Tracking Individuals in Classroom Videos via Post-processing OpenPose Data

PAUL HUR, University of Illinois Urbana–Champaign, USA
NIGEL BOSCH, University of Illinois Urbana–Champaign, USA

Analyzing classroom video data provides valuable insights about the interactions between students and teachers, albeit often through time-consuming qualitative coding or the use of bespoke sensors to record individual movement information. We explore measuring classroom posture and movement in secondary classroom video data through computer vision methods (especially OpenPose), and introduce a simple but effective approach to automatically track movement via post-processing of OpenPose output data. Analysis of 67 videos of mathematics classes from middle school and high school levels highlighted the challenges associated with analyzing movement in typical classroom videos: occlusion from low camera angles, difficulty detecting lower body movement due to sitting, and the close proximity of students to one another and their teachers. Despite these challenges, our approach tracked person IDs across classroom videos for 93.0% of detected individuals. The tracking results were manually verified through randomly sampling 240 instances, which revealed notable OpenPose tracking inconsistencies. Finally, we discuss the implications for supporting more scalability of video data classroom movement analysis, and future potential explorations.

CCS Concepts: • **Applied computing** → **Education**; • **Computing methodologies** → **Tracking**; *Object detection.*

Additional Key Words and Phrases: classroom video, video analysis, posture, movement

ACM Reference Format:
Paul Hur and Nigel Bosch. 2022. Tracking Individuals in Classroom Videos via Post-processing OpenPose Data. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3506860.3506888

1 INTRODUCTION

Video recordings of traditional classrooms capture detailed interactions of students and the instructor. The value of video recordings as data has been established in educational research, where they have long been used for qualitative research to analyze pedagogy and to obtain rich classroom insights through teacher reflection [28], classroom comparison [22], and stimulated recall [24]. Furthermore, qualitative coding of video data has been used to identify patterns in classroom dialogue [15]. Gestures are also readily recognizable in video; thus, teacher and student gestures have been a central interest of embodied cognition research, which has examined roles of gestures in mathematics education such as communicating abstract representations [3] and scaffolding [2] concepts. When combined with machine learning methods, qualitative video coding (and other types of manual coding) can serve as ground-truth labels to help automatically detect attentional states [6], analyze patterns in student group collaboration [16], and other constructs (e.g., [9, 33, 34]).

Although recorded classroom videos are a valuable source of information-rich data for education researchers and teachers, they can be difficult to translate into insights because of their heterogeneity. Videos can capture a variety of learning environments, and researchers may collect video data to answer various project-specific research questions

Manuscript submitted to ACM

about diverse modalities. Using existing analysis methods designed for an entirely different set of research questions can lead to issues. These incompatibility issues have led to researchers needing to develop project-specific video coding schemes, which are a set of definitions that the researcher determines to be representative of particular behaviors, such as whether students are on task or verbally interacting with their peers [30]. Developing video coding schemes can be a time-consuming process that relies on researchers' domain expertise and idiosyncratic observations from the video. Once created, video coding schemes are usually most applicable to the researchers' own dataset. Manual coding of video data introduces further complexities as project members need to be trained on the coding process and the coders need to obtain a reasonable level of inter-coder reliability.

More recently, as a result of the progress made with applying machine learning approaches to educational data, there is a body of work in developing automated methods for analyzing video data [8, 12, 13, 19, 29]. These methods have provided alternate means of obtaining person movement, position, and posture information that enable larger-scale qualitative and quantitative analyses. Constructs like movement alone may be too low-level for automatically coding variables of interest like student engagement or affect, however. Consequently, researchers have used multimodal approaches to study movement in learning to account for the unpredictability of movement, such as developing affect detectors using facial expressions as the primary channel and body movements as the secondary channel [7], and physiological sensors (camera, pressure mouse, pressure-sensitive chair, and conductance bracelet) along with students' self-reports [4]. Movement and related variables that can be extracted from videos serve to complement these detailed data for quantitative analyses (e.g., machine learning), and may serve qualitative research by guiding researchers toward interesting points in videos for in-depth analysis.

1.1 Novelty and Contribution

Little is currently known about the specific challenges that arise when applying automatic video analysis methods for secondary analysis of classroom data. Classroom video data exist widely for a few reasons: they may originally be collected for different purposes and research questions [27], they may be publicly available on YouTube, or teachers may record and store their classroom sessions for their own instructional development and reflection [31]. Modern computer vision tools such as OpenPose [10], Deepcut [25], RMPE [14], and VIBE (Video Inference for human Body pose and shape Estimation) [17] are easily applied to videos to explore new research questions. In particular, the scalability of video data analysis may potentially be increased via investigating the application of computer vision tools on existing data. However, whether or not such tools are already well suited to the kinds of video that arise from typical middle school and high school classroom recordings is another matter.

In this paper, we examine one such computer vision method, OpenPose, to analyze classroom videos that were not originally recorded for such analysis. We examine the extent to which OpenPose can be useful for detecting the position, posture, and movement of students and teachers in the context of middle school and high school classrooms, and propose a simple yet effective approach for extending OpenPose's capabilities via post-processing of output data to address challenges encountered during secondary analyses of classroom video data. We aim to explore the particular characteristics of classrooms videos when analyzing movement data.

We first discuss methods that researchers have previously used to track and study movement in learning, and describe the secondary video data used for our analysis. Then, we describe the OpenPose configurations we used and introduce our post-processing methods to track individuals throughout the videos. We then discuss the results, our manual verification process to better support the validity the results, and the implications of this type of analysis for potentially increasing the scalability of analyzing existing classroom videos.

We organize our analysis in this paper around the following two research questions:

RQ1: What challenges arise when automatically measuring movement in classroom video data via OpenPose?
RQ2: How well can the issues be resolved via post-processing OpenPose output?

2 RELATED WORK

There is a growing body of work using diverse methodological approaches to examine movement data in learning. Research in the area has been partly motivated by the idea of spatial pedagogy as coined by Lim et al. [18], who describe spatial pedagogy as the teacher's physical positioning and movement through the learning environment with respect to the students and learning materials, and these spatial factors' meaning in relation to creating effective pedagogy.

In order to track and analyze positioning and movements of individuals in classrooms, some researchers have utilized low-cost wearable badge sensors to track individuals' patterns of movement, which have been represented and analyzed as visual heatmaps in higher education contexts in design courses [20] and lab sections [21]. Researchers have also developed systems using a variety of different custom sensors and features [26]. Perhaps the most developed work published on multimodal approaches for automatically analyzing classroom interactions is EduSense [1]. EduSense is a system for instructor-facing dashboards in higher education instructional feedback. It integrates various visual and audio features, such as detecting sitting and standing, hand-raising, and speech data patterns. EduSense researchers utilized a single wide-lens camera mounted onto electric boxes near the ceiling to give a bird's-eye view that is able to capture a comprehensive perspective of classroom activity. They then processed video data through EduSense, which consists of applying custom-tuned OpenPose processing to reduce false-positive body detection.

Video recording is often less invasive than other sensors (e.g., sensors worn on the body), which may improve the ecological validity of the work as it may be less likely to impact subjects' behaviors. While there is work which has analyzed existing classroom video data (collected through videos publicly available on YouTube) to automatically analyze types of movements in video such as eye gaze following [5], there is no such work on automatically tracking persons across time in existing videos from real-world middle school and high school classroom contexts. Thus, it may be valuable to develop these methods for positioning and movement on existing video in order to help improve the empirical understanding of spatial pedagogy.

3 DATA

In line with our research motivation, the data used for analysis were not originally collected for the purposes of this research. The video data were collected in 2014 and 2015 for a different research project which closely examined the processes teachers engage in when teaching using point-of-view cameras; specifically, the project aims were to better understand mathematics teachers' responsive teaching practices. To that end, cameras were positioned in various middle school and high school mathematics classrooms located in the United States to capture the interactions between students and instructors.

We utilized our tracking approach with diverse video data to evaluate its generalizability. Our study analyzed a sample of 67 classroom videos of around 90 minutes in length in class periods of one hour long—videos started earlier and later than classes in order to capture footage of the full class period. The collection of videos represent videos from 6 different teachers, across two years mathematics classes with different students. The videos were recorded with either a Sony HDR-MV1 camera or a Zoom Q4 Handy Video Recorder in 1080p (1920×1080 pixels) resolution at 30 frames per

Fig. 1. An example video frame from a classroom video

Fig. 2. Example video frames showing the variety of video data

second. The Sony and Zoom cameras had 120-degree and 130-degree fields of view, respectively. The cameras used to record the videos in this paper had been positioned in locations of classrooms that were able to capture interactions from the entire class, such as stage left or stage right. Figures 1 and 2 show the wide variety of videos, with each video having differences in camera placement, captured perspective, and classroom lighting.

4 METHODS

We randomly sampled one-minute video clips for analysis from each of the 67 videos in the collection and extracted keypoints from each clip using OpenPose. Then, we performed post-processing on the resulting data in order to implement inter-frame tracking of individuals. Finally, we performed manual verification of our tracking method to assess its accuracy. In this section, we describe these processes in more detail.

4.1 Video Sampling and OpenPose Processing

We sought to explore the generalizability of our tracking method. We thus focused on analyzing short clips from many classroom videos to capture data diversity, rather than focusing in depth on specific videos. Based on our preliminary observations from the videos, the 30 to 60 minute time span represented portions of the video when the classroom was fully settled, and the most representative of typical classroom activity for the respective classes. This time span included instances of students moving around, engaging with the course material and instructor, and interacting with each other in the classroom. Thus, we randomly sampled one-minute video clips from the 67 videos starting from from the 30th to the 59th minute of class. The resulting 67 one-minute videos were subsequently processed via OpenPose in order to extract body keypoints (x and y coordinates of various points on each visible body).

Like other pose estimation tools such as Deepcut [25], and RMPE [14], OpenPose is a computer vision tool for identifying individuals in video data by jointly detecting human body, hand, facial, and foot keypoints on a single image [10]. For our analysis, we used the 25-keypoint body and foot keypoint configuration. We expected that the addition of the other facial features and hands keypoints provided by other configurations would not substantially improve our planned tracking implementation of persons in post-processing, while complicating output interpretation and greatly increasing the processing time. OpenPose output consists of JavaScript Object Notation (JSON) files, each of which holds an array of objects representing identified persons with body part locations as coordinates and detection confidence. Since OpenPose processes videos per frame, the total number of output files per video was $\approx$ RECORDED FRAMES-PER-SECOND $\times$ SECONDS OF VIDEO.

4.2 Post-processing OpenPose Output Data

OpenPose does not have native support for inter-frame person tracking. Each frame of video is newly analyzed and persons in the current frame are detected without information about the previously processed frames. This makes it difficult to use OpenPose for analyses like examining teacher interaction patterns or peer movement interactions over a class period, since the analyses require tracking individual people over time. Here, we outline an approach to track students and teachers between frames through post-processing. We applied our approach to the output data of each one minute video (at 30 frames per second), or around 1,800 frames per video.

First, we concatenated the JSON output files containing the keypoint coordinates and confidence values of each detected person per frame, forming a single output file per video for easier access to values during calculation. We considered low-confidence keypoint detections (≤ 0.3) as non-detections and filtered them out prior to calculations, based on empirical observations that the coordinates of these low-confidence keypoints varied widely. Then, for each frame in each video output, we calculated individual Euclidean distances from the 25 keypoints of each person in the current frame to the coordinates of all corresponding keypoints in the previous frames. The calculations continued with progressively earlier frames in the video until finding a distance of less than 10 pixels or until reaching 5 frames previous. We recorded the two smallest non-zero distance values along with the keypoint index of the smallest distance. This allowed us to determine if the best match across frames is close (i.e., closest distance is small) and unambiguous (i.e., second-smallest distance is large).

We then assigned person IDs based on these inter-frame matches by iterating through frames in reverse order and linking IDs based on close matches. If a detected person had no clear matches, we assigned a new person ID. Alternatively, if a person had multiple matches in the same frame, the ambiguity was resolved via voting for the most keypoints matched. Furthermore, our tracking also accounted for instances when OpenPose briefly failed to detect a person or where one person was detected as more than one, which is described in section 5.1 in further detail. In cases when one person's keypoints were fragmented into two or more sets of keypoints and incorrectly detected by OpenPose to belong to different people, we automatically merged them into one person with one person ID if the keypoints were complementary and there was evidence from adjacent frames that they corresponded to a single person. On the other hand, in instances when OpenPose had inconsistent person detection between frames, we estimated missing keypoints by interpolating the missing information from earlier and later frames.

4.3 Manual Verification of Inter-frame Tracking

We performed a manual verification step in order to support the validity of our tracking approach. Each of the 240 randomly selected samples represented the 25 keypoints' coordinates (x, y) of one person in one frame of one video. These data were then compared to the keypoint data of the person which was determined by our tracking approach to belong to the same person (same person ID). The two corresponding video frames of comparison were opened in a photo editing tool which displayed photo pixel coordinate values. The identity of the person was checked by carefully examining keypoint coordinates via the photo editing tool. We then recorded whether the person ID belonged to the same person in both frames.

We conducted a binomial distribution power analysis to determine how many samples (i.e., pairs of consecutive frames) were needed to detect an error rate of 20% or higher with 80% power, using the `pwr` package in *R* [11, 32]. Power analysis showed that 197 samples would be needed. We selected slightly more samples to account for incomplete

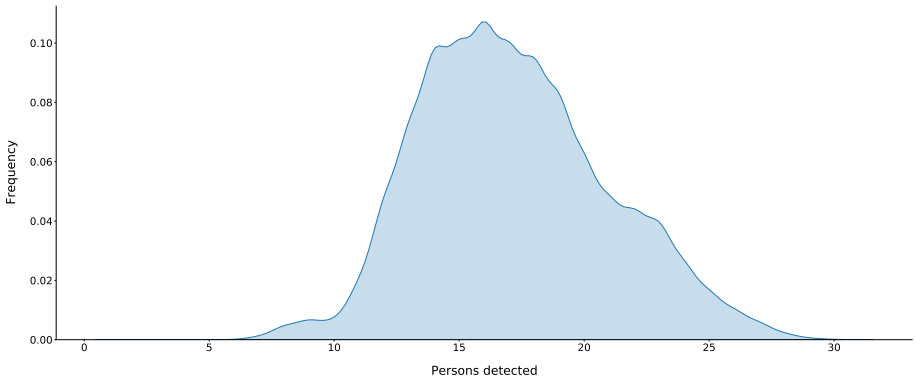

Fig. 3. Density plot of the frequency of persons detected by OpenPose across all 67 videos. The true number of individuals in each video ranged from 16 to 31 (avg. 22.4)

or empty data points and to allow for even sampling across the 6 teachers; specifically, we randomly selected 40 per teacher for a total of 240 samples.

5 FINDINGS

In this section, we describe our observations from examining OpenPose output data in this classroom context and findings from our post-processing tracking method. While OpenPose output revealed consistency issues with keypoint and person detection, inter-frame tracking was highly accurate for the individuals detected. Furthermore, we describe the instances when tracking was unsuccessful during the manual verification process.

5.1 Inconsistencies in Person Detection

An exploration of OpenPose's detection process revealed consistency issues with person detection. During observations of OpenPose's detection process (viewing detected keypoints in the video while the software was extracting keypoints), we observed numerous instances of intermittent detection failures. Identified keypoints, and sometimes entire individuals, would alternate unpredictably between being detected and not being detected. This resulted in inconsistent numbers of persons detected across the video as shown in the density plot of the numbers of detected individuals in Figure 3. Across all videos, the average percentage of individuals detected by OpenPose out of the manually-counted true number of classroom individuals was 77.2%.

Intermittent detection manifested in two ways. OpenPose sometimes split one individual's keypoints into multiple different individuals, increasing the apparent number of individuals detected. The split-up keypoints coordinates were so close to the keypoints of multiple individuals that all keypoints could be interpreted as correctly detected, which complicated tracking. Alternatively, OpenPose can merge multiple individuals' keypoints into one single detected individual, leading to a smaller number of detected individuals.

5.2 Tracking Performance

Our tracking approach was largely successful. There were an average of 30,868 instances of OpenPose person detections per one-minute video clip. Persons were tracked across 93.0% of these detections. We calculated this success rate as the percentage of detections for which person IDs could be matched (versus creating a new ID).

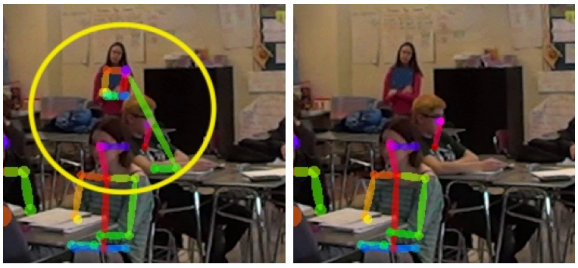

Fig. 4. An example of when the tracking method failed to recognize the individual as the same person due to inconsistency in person detection. Relevant person is circled in yellow in Frame A.

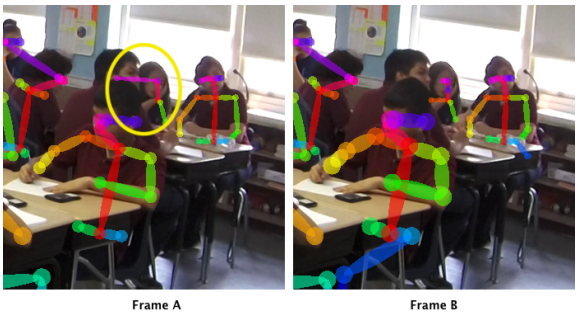

Fig. 5. An example of when the tracking method failed to recognize the individual as the same person due to a small number of detected keypoints which varied between frames. Relevant person is circled in yellow in Frame A.

5.3 Observations from Manual Verification of Tracking Approach

Out of 240 manually checked samples, there were 15 samples when our tracking method did not identify the same person ID. These instances of unsuccessful tracking manifested due to either OpenPose's person detection inconsistencies as outlined in section 5.1 above, or our tracking method was unable to accommodate different keypoints being detected in cases when few keypoints were detected in both frames. Out of the 15 instances of tracking failure, 13 instances were due to OpenPose detection inconsistencies, and 2 were due to our tracking method failing to track persons with a small number of keypoints. Figure 4 shows an example of when OpenPose detected a person in one frame (Frame A) but did not in the next frame (Frame B), despite very little movement differences between the two frames. On the other hand, Figure 5 shows the tracking method failing due to a small number of detected keypoints in both frames: 8 keypoints in Frame A, and 10 keypoints in Frame B. The small number of total keypoints for the person, combined with different keypoints identified in the two frames, led to the inability of our tracking approach to track effectively.

6 DISCUSSION

Compared to collecting original video data for movement analysis, we had no control over optimizing the data collection process for our analysis. Occlusions may have been a major contributor to our tracking performance not reaching greater tracking accuracy. In ideal scenarios, cameras near the ceiling avoid issues with occlusion [1]; however, in secondary data analyses like ours, comparatively low angles are common since this typically makes camera setup more

straightforward. This created many instances of occlusion, as students positioned very close together appear even closer in the 2D plane. Furthermore, the lower half of students' bodies were often obscured since students were seated at or around desks, so OpenPose was not able to consistently detect many lower-body keypoints. Thus, the keypoints that were more rarely detected were those located in the lower half of the body. As shown in the box plots in Figure 6, some keypoints (1, 2, 5, 0) were detected with reasonable consistency, while some keypoints (13, 10, 11, 14, and 19 through 24) were rarely detected across the 67 videos—the bottom 10 least often detected keypoints were found in the legs.

Further analysis into keypoint differences revealed movement magnitude difference patterns based on location. The five most frequently detected keypoints (1, 2, 5, 0, 8) were found in the torso, lower, or head and neck, and had average inter-frame distances (in pixels) of 1.67, 2.03, 2.07, 1.93, and 3.54, respectively. The five least frequently detected keypoints (22, 19, 20, 23, 21) were all found in the legs, and had average inter-frame distances (in pixels) of 9.80, 10.01, 9.66, 9.72, and 9.57, respectively. Despite these larger distances for some keypoints, the tracking performance was still quite good across the videos since the difference between closest and second-closest distance typically provided strong evidence of matches even when keypoint detection was imprecise. Our manual verification process revealed that our approach successfully tracked 225 out of 240 instances, with 15 instances of tracking failures. A majority of these instances (13) were due to persons having variable detection between the two consecutive frames. This shows that there is room for improving our tracking approach through accounting for instances when keypoint information is more sparse.

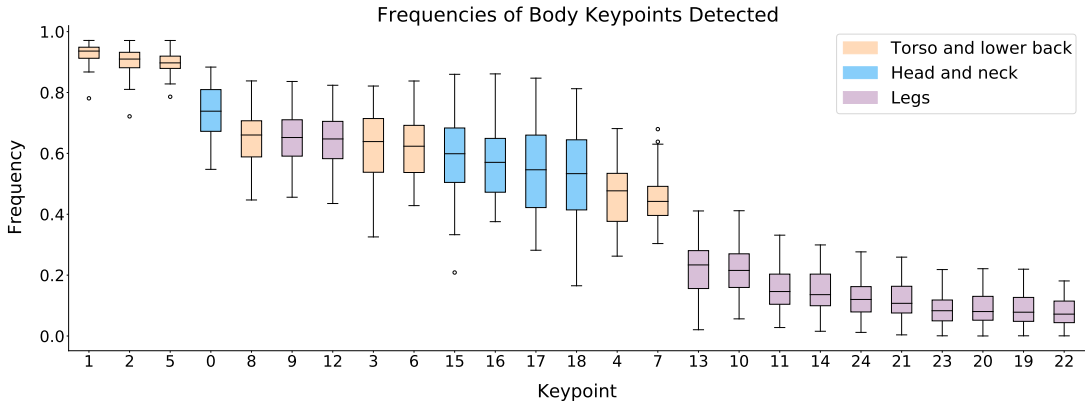

Fig. 6. Box plots of keypoint detection frequencies of all 67 videos

7 FUTURE WORK

Future work could improve the process by integrating the keypoint detection frequency analysis which showed detection rate variations based on keypoint locations on the body. The keypoints could be assigned different weights denoting their importance in the tracking process, in which more commonly detected keypoints (e.g., in the upper body) might receive higher importance weights. This will help to reduce noise from less commonly detected keypoints (e.g., in the lower body) with large inter-frame distances, which could contribute to fewer tracking failures. Furthermore, we can perhaps account for the inconsistent contrast across videos caused by uneven classroom lighting or camera hardware limitations. Recent work has shown that image pre-processing operations such as increasing the contrast and sharpness of the target could increase OpenPose's detection accuracy by up to 38.37% [23]. Our tracking approach could also be

applied to videos of a wider range of classroom environments, as the collection of videos analyzed in this study had some similarities in classroom layout. There are opportunities for further improving the robustness of our inter-frame tracking approach, such as scanning over longer periods of time to re-identify persons when tracking fails, using motion to anticipate where persons are likely to be in subsequent frames, and other approaches. Such improvements will allow for the investigation of trends in movement over entire class periods, and further develop spatial pedagogy research. Finally, while many current pose estimation software packages like OpenPose support person detection but not person *recognition*, there are still concerns in maintaining the privacy of individuals' data. Future work should consider ways to depersonalize such potentially sensitive data through blurring of faces after pose estimation but before data analysis, or separating available student characteristic data (names, grades, etc.) from video data.

8 CONCLUSION

In this paper, we described a method for automatically measuring the positions and movements of teachers and students in classroom videos. We were motivated by the potential scalability of these methods when compared to more manual qualitative methods, and its less intrusive nature compared to using custom sensor systems. Tracking persons in classroom settings highlighted several challenges, like large amounts of occlusion and intermittent detection failures, along with less-than-ideal video angles, all of which are expected in real-world classroom settings. Much work remains to be done to fully address these challenges. However, our post-processing solution for overcoming these challenges while tracking students and the instructor shows promise, which was validated through a manual verification process. Our code, including ongoing improvements, is documented and publicly available (https://github.com/tca2/videodata-processing).

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (DRL-1920796). 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.

REFERENCES

[1] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical classroom sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[2] Martha W Alibali and Mitchell J Nathan. 2007. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In *Video Research in the Learning Sciences*. Routledge, New York, NY.

[3] Martha W Alibali, Mitchell J Nathan, and Yuka Fujimori. 2013. Gestures in the mathematics classroom: What's the point? In *Developmental Cognitive Science Goes to School*. Routledge, New York, NY, 233–248.

[4] Ivon Arroyo, David G Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. 2009. Emotion sensors go to school.. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, Vol. 200. Citeseer, IOS Press, 17–24.

[5] Arkar Min Aung, Anand Ramakrishnan, and Jacob R Whitehill. 2018. Who Are They Looking At? Automatic Eye Gaze Following for Classroom Observation Video Analysis.. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, 252–258.

[6] Nigel Bosch and Sidney D'Mello. 2019. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing* 12 (2019), 974–988. Issue 4.

[7] Nigel Bosch, Sidney K D'Mello, Jaclyn Ocumpaugh, Ryan S Baker, and Valerie Shute. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 2 (2016), 1–26.

[8] Nigel Bosch, Caitlin Mills, Jeffrey D Wammes, and Daniel Smilek. 2018. Quantifying classroom instructor dynamics with computer vision. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018)*. Springer, Cham, CH, 30–42.

[9] Fiona J Buckingham, Keeley A Crockett, Zuhair A Bandar, and James D O'Shea. 2014. FATHOM: A neural network-based non-verbal human comprehension detection system for learning environments. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE,

Piscataway, NJ, 403–409.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[11] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, and Helios De Rosario. 2020. pwr: Basic Functions for Power Analysis. https://CRAN.R-project.org/package=pwr

[12] Yuxuan Chen, Nigel Bosch, and Sidney D'Mello. 2015. Video-based affect detection in noninteractive learning environments. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, 440–443.

[13] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1–20.

[14] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*. Computer Vision Foundation, 2334–2343.

[15] Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction* 9 (2016), 16–44.

[16] Paul Hur, Nigel Bosch, Luc Paquette, and Emma Mercier. 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*. International Educational Data Mining Society, 104–114.

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, 5253–5263.

[18] Fei Victor Lim, Kay L O'Halloran, and Alexey Podlasov. 2012. Spatial pedagogy: Mapping meanings in the use of classroom space. *Cambridge journal of education* 42, 2 (2012), 235–251.

[19] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 325–347.

[20] Roberto Martinez-Maldonado. 2019. "I Spent More Time with that Team": Making Spatial Pedagogy Visible Using Positioning Sensors. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, New York, NY, 21–25.

[21] Roberto Martinez-Maldonado, Vanessa Echeverria, Jurgen Schulte, Antonette Shibani, Katerina Mangaroska, and Simon Buckingham Shum. 2020. Moodoo: Indoor positioning analytics for characterising classroom teaching. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Springer, Cham, CH, 360–373.

[22] Ashleigh Mayo, Manjula D Sharma, and Derek A Muller. 2009. Qualitative differences between learning environments using videos in small groups and whole class discussions: A preliminary study in physics. *Research in Science Education* 39, 4 (2009), 477–493.

[23] Jannik Christian Lærkegård Pedersen, Mattias Foltmar Sander, Niklas Fruerlund Jensen, Jonas Lasham Lakhrissi, Mikkel Gede Hansen, Patrick Staalbo, and Andreas Wulff-Abramsson. 2019. Improving the Accuracy of Intelligent Pose Estimation Systems Through Low Level Image Processing Operations. In *International Conference on Digital Image & Signal Processing (DISP'19)*. 4 pages.

[24] Susan EB Pirie. 1996. Classroom Video-Recording: When, Why and How Does It Offer a Valuable Data Source for Qualitative Research?. (1996), 17 pages.

[25] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Piscataway, NJ, 4929–4937.

[26] Luis Pablo Prieto, Kshitij Sharma, Łukasz Kidzinski, María Jesús Rodríguez-Triana, and Pierre Dillenbourg. 2018. Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of computer assisted learning* 34, 2 (2018), 193–203.

[27] Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. 2015. Translating head motion into attention-towards processing of student's body-language. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, 320–326.

[28] Rossella Santagata and Giulia Angelici. 2010. Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of teacher education* 61, 4 (2010), 339–349.

[29] Arjun Sharma, Arijit Biswas, Ankit Gandhi, Sonal Patil, and Om Deshmukh. 2016. LIVELINET: A Multimodal Deep Recurrent Neural Network to Predict Liveliness in Educational Videos.. In *Proceedings of the 9th International Conference on Educational Data Mining*. International Educational Data Mining Society, 215–222.

[30] Saadeddine Shehab and Emma Mercier. 2019. Visualizing Representations of Interaction States during CSCL. In *Proceedings of the 13th International Conference on Computer Supported Collaborative Learning*. International Society of the Learning Sciences (ISLS), Lyon, France, 871–872.

[31] Miriam Sherin and Elizabeth van Es. 2005. Using video to support teachers' ability to notice classroom interactions. *Journal of technology and teacher education* 13, 3 (2005), 475–491.

[32] R Core Team. 2013. R: A language and environment for statistical computing. (2013).

[33] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[34] Andrea Stevenson Won, Jeremy N Bailenson, and Joris H Janssen. 2014. Automatic detection of nonverbal behavior predicts learning in dyadic interactions. *IEEE Transactions on Affective Computing* 5, 2 (2014), 112–125.