Multimodal Features for Group Dynamic-Aware Agents

Iliana Castillon*[1], Videep Venkatesha*[1], Hannah VanderHoeven[1], Mariah Bradford[1], Nikhil Krishnaswamy[1], and Nathaniel Blanchard[1]

Colorado State University, Colorado 80521, USA
Iliana.Castillon@rams.colostate.edu

Abstract. Collaborative problem solving (CPS) is a necessary skill in education and the workforce. Here, we provide a brief tutorial for educational stakeholders on how acoustic and video data of group collaboration can be distilled into interpretable features. These features can then inform an AI agent about a group and their progress through the task, ultimately enabling the agent to make appropriate decisions about when and how to interact with the group. We hypothesize that there may be features, specific to education, that could further inform the agent, and discuss pending problems with how data can be utilized to debug/improve an agent, versus inevitable privacy concerns.

Keywords: Collaborative Problem Solving · Multimodal features · Agents

1 Introduction

For an AI agent to facilitate a Collaborative Problem Solving (CPS) task, the agent needs to both follow the team's progress through a task and interpret the team's multimodal communication strategies [3], i.e., group interactions. Interpreting group interactions requires access to features that are essential to communication (speech, gesture, pose, and expression) and the collaborative task that students are engaging with. Pinpointing students' progress through a task is dependent on what steps students have taken, and what future steps they might take. Unless a task is integrated into a program, understanding progress requires the agent to recognize and interpret the physical world, such as the position of key objects and how object position may correspond with evaluation. Ideally, all of this information will be collectively utilized by an AI agent to make a decision about how to engage with the group.

Here, we showcase a toolbox of solutions we have compiled to provide agents with this information. For example, a participant's body language corresponds to their engagement with the task and their interactions with others. The agent is provided information about participants' body language with the OpenPose [4] library, which includes a pre-trained machine learning model, that can be utilized to track basic poses, or expanded upon to track poses specific to CPS tasks.

* These authors contributed equally.

Our ultimate aim is to showcase and justify to educational stakeholders what, and how, information is provided to the AI agent. Ultimately, we plan to extend our agent to K-12 classrooms, and have the agent facilitate CPS. We hope to understand what additional information educational stakeholders believe such an agent should be aware of, and understand how educational stakeholders believe information the agent ingests should be handled in post-processing (after the agent makes a decision). Finally, we conclude with a discussion on privacy and improvement tradeoffs, centered around what data, if any, should be stored, and who should have access.

2 A Toolbox for Providing Information to AI Agents

For an AI agent to decide how to guide an interaction, it needs to consider multiple relevant information channels, such as video and audio [8]. In turn, the raw data from these channels is often distilled into features — for example audio data with speech can be automatically transcribed. For our purposes, we emphasize interpretable features, such as a gesture label. Additionally, we note that many of these interpretable features are predicted by deep learning models, which process raw data using non-interpretable features.

Here, we outline several categories of features as well as current technologies that can be used to automatically provide said features. In Section 2.1, we discuss language and auditory features, including automatic speech recognition and audio-prosodic features. Section 2.2 focuses on visual features, such as pose estimation and object tracking.

2.1 Language and Auditory Features

Linguistic features derived from conversational transcripts are key for modeling group work [9]. Typically, raw audio is automatically segmented into individual utterances and each utterance is transcribed with automatic speech recognition. Utterance segmentation can also be refined with speech diarization, which ensures only one speaker is included in each utterance. For example, Google's open source Automated Speech Recognition (ASR) can be used to segment utterances, and has a speech diarization option.

Utterances provide descriptive features about participation and speaking time, such as speaking length, total turns speaking, etc. Further, prosodic speech features (such as energy and pitch features) can be extracted using tools like Wavesurf [7] and OPENsmile [5]. These features and data have the potential to provide a deeper insight into the inner workings of language and acoustic features during group work.

Linguistic features can be extracted from the automatic transcripts of the utterances [5] using the SpaCy Library. Typically, multimodal systems (prosodic and linguistic features are considered separate modalities) yield better performance than just using a single modality [5].

2.2 Visual Features

Visual features grant the agent access to the physical world, which includes interactions among participants, interactions with key CPS task objects, and potentailly even CPS task progress.

For example, 6D Object pose estimation [10], a method to determine an object's location and the rotation, is relevant for group work because task-specific objects can be identified and tracked, something that is essential for understanding a group's progress as they work through the task. This can be applied in the Fibonacci Weights Task, which prompts groups to determine the weights of various blocks using a scale. For a fuller description of the task, see [2]. 6D pose could be used to track the blocks users are interacting with, and evaluate student solutions based on block placement.

Another key visual facet is gestures, which are a key component of multimodal communication employed in everyday interactions [6, 11, 12]. Gestures can be combined with other information — for example, an identified 'pointing' gesture can be further processed to identify what a participant was pointing at. This, in turn, could be passed to an agent (participant one was pointing at the green block). Google's MediaPipe library [13] allows hands to be tracked using 21 landmarks that consist of x, y, and depth coordinates. The vector of the finger used to 'point' can be projected through the scene, and the 6D pose of key objects can be used to determine if the participant is pointing at a key object. In a similar vein, the pose of participants, extracted using libraries like OpenPose [4], can be used to track and analyze positions of participants' bodies.

Finally, visual features are essential for interpreting actions and attention of participants. There are many face detection models which can be used to identify a participant's face — this is an important first step for enumerable downstream tasks like face recognition, which can be used to keep track of which participant is which, even if positions change. Detected faces can also be utilized to identify where participants' attention is focused, such as capitalizing on head pose and eye gaze [1]. Accessing these actions and features paves way for the educational community to provide insight on further analysis.

3 Discussion

It is important to note that the above-mentioned features do not encompass all the research that is done in the field of studying group work, and is far from an exhaustive survey. However, most of the tools and models cited are open source solutions that are relatively easy to integrate into an agent processing pipeline. Ideally, these tools will be easy for anyone in the Artificial Intelligence in Education (AIEd) or Educational Data Mining (EDM) community to utilize.

We have explicitly illustrated how we, as machine learning scientists, have attempted to distill CPS for an AI agent. We expect the education community will have ample insights into additional features that we should be accounting for, or additional tools that can be added to the toolbox.

4 Conclusion

In this paper, we have identified a number of tools that can be integrated in situations that involve monitoring group work. Some of the tools highlighted provide agents the ability to interpret key components of group work by those who are not experts in Artificial Intelligence or related fields. For instance, out-of-the-box hand tracking models are sufficient for identifying hand joints. However, there are portions of using these tools that require experts to implement custom solutions. For example, there are a number of open gesture recognition solutions, but these solutions may not include gestures that are relevant for group work. Further, if the pointing gesture is identified, a custom solution could identify what or who the gesture is directed at. The specifics of which gestures are important for group work is a key area that other specialists can help machine learning specialists with, for example by identifying which gestures should be included in a model.

5 Collaborating with Educational Stakeholders

The educational community can provide perspectives and insight that may not have been addressed from a solely technical perspective. In order for an agent to be able to facilitate group work, it must receive information that is crucial in identifying successful collaboration. Given that our ultimate goal is to provide a virtual agent in a live classroom, we invite all inputs from educational stakeholders to weigh in on the importance of the listed features, as well as classroom specific features. Considering that communication during group work is multimodal, we want to collaborate with the educational community to investigate other features that would give the agent relevant information. Further, we are interested in exploring intervention strategies for the agent to employ which enhance the CPS.

In addition to this, we are interested in discussing how educational stakeholders feel about the trade-off of privacy in the classroom (all information is discarded) and understanding/debugging agent behavior (information is stored but only utilized for upgrading the agent). In the former case, information the agent utilized to make a decision could still be stored, but with the caveat that any machine learning models that provide information will be imperfect, and the model prediction may not accurately reflect reality. In the latter case, we are interested in whom educational stakeholders (e.g., teachers, students, parents, administrators, etc.) believe should have access to such data. Finally, we seek feedback from education and human computer interaction communities about key group work features that we may be overlooking, and that an agent should have access to.

References

1. Aung, A.M., Ramakrishnan, A., Whitehill, J.R.: Who Are They Looking At? Automatic Eye Gaze Following for Classroom Observation Video Analysis. International Educational Data Mining Society (2018), publisher: ERIC
2. Bradford, M., Hansen, P., Beveridge, J.R., Krishnaswamy, N., Blanchard, N.: A deep dive into microphone hardware for recording collaborative group work. In: Proceedings of the International Conference on Educational Data Mining (2022)
3. Bradford, M., Hansen, P., Lai, K., Brutti, R., Dickle, R., Hirshfield, L.M., Pustejovsky, J., Blanchard, N., Krishnaswamy, N.: Challenges and opportunities in annotating a multimodal collaborative problem solving task. In: Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIEd. International AIEd Society (2022)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (Jan 2021). https://doi.org/10.1109/TPAMI.2019.2929257, https://ieeexplore.ieee.org/document/8765346/
5. Murray, G., Oertel, C.: Predicting Group Performance in Task-Based Interaction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 14–20. ACM, Boulder CO USA (Oct 2018). https://doi.org/10.1145/3242969.3243027, https://dl.acm.org/doi/10.1145/3242969.3243027
6. Quek, F., Mcneill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., Mccullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Trans. Comput.-Hum. Interact. **9**, 171–193 (Jan 2002)
7. Sanchez-Cortes, D., Aran, O., Mast, M., Gatica-Perez, D.: A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. IEEE Transactions on Multimedia **14**, 816–832 (Jun 2012). https://doi.org/10.1109/TMM.2011.2181941
8. Stewart, A.E.B., Keirn, Z., D'Mello, S.K.: Multimodal modeling of collaborative problem-solving facets in triads. User Modeling and User-Adapted Interaction **31**(4), 713–751 (Sep 2021). https://doi.org/10.1007/s11257-021-09290-y, https://doi.org/10.1007/s11257-021-09290-y
9. Sun, C., Shute, V.J., Stewart, A.E.B., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N., D'Mello, S.K.: The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. Computers in Human Behavior **128**, 107120 (Mar 2022). https://doi.org/10.1016/j.chb.2021.107120, https://www.sciencedirect.com/science/article/pii/S074756322100443X
10. Trabelsi, A., Chaabane, M., Blanchard, N., Beveridge, R.: A Pose Proposal and Refinement Network for Better 6D Object Pose Estimation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2381–2390. IEEE, Waikoloa, HI, USA (Jan 2021). https://doi.org/10.1109/WACV48630.2021.00243, https://ieeexplore.ieee.org/document/9423353/
11. Turk, M.: Multimodal interaction: A review. Pattern Recognition Letters **36**, 189–195 (Jan 2014). https://doi.org/10.1016/j.patrec.2013.07.003, https://www.sciencedirect.com/science/article/pii/S0167865513002584
12. Wang, I., Fraj, M.B., Narayana, P., Patil, D., Mulay, G., Bangar, R., Beveridge, J.R., Draper, B.A., Ruiz, J.: EGGNOG: A Continuous, Multi-modal Data Set of Naturally Occurring Gestures with Ground Truth Labels. In: 2017 12th IEEE

International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 414–421 (May 2017). https://doi.org/10.1109/FG.2017.145

13. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: MediaPipe Hands: On-device Real-time Hand Tracking. CoRR (2020), https://arxiv.org/abs/2006.10214