Multimodal Behavioral Disengagement Detection for Collaborative Game-Based Learning

Fahmid Morshed Fahid[1][0000-0002-4802-3979], Halim Acosta[1], Seung Lee[1], Dan Carpenter[1][0000-0003-3253-8369], Bradford Mott[1][0000-0003-3303-4699], Haesol Bae[2], Asmalina Saleh[2], Thomas Brush[2], Krista Glazewski[2], Cindy E. Hmelo-Silver[2], and James Lester[1]

[1]North Carolina State University, Raleigh, NC 27695, USA
{ffahid, hacosta, sylee, dcarpen2, bwmott, lester}@ncsu.edu
[2]Indiana University, Bloomington, IN 47405
{haebae, asmsaleh, tbrush, glaze, chmelosi}@indiana.edu

Abstract. Collaborative game-based learning environments offer significant promise for creating effective and engaging group learning experiences. These environments enable small groups of students to work together toward a common goal by sharing information, asking questions, and constructing explanations. However, students periodically disengage from the learning process, which negatively affects their learning, and the impacts are more severe in collaborative learning environments as disengagement can propagate, affecting participation across the group. Here, we introduce a multimodal behavioral disengagement detection framework that uses facial expression analysis in conjunction with natural language analyses of group chat. We evaluate the framework with students interacting with a collaborative game-based learning environment for middle school science education. The multimodal behavioral disengagement detection framework integrating both facial expression and group chat modalities achieves higher levels of predictive accuracy than those of baseline unimodal models.

Keywords: Multimodal Learning, Collaborative Game-Based Learning, Behavioral Disengagement.

1 Introduction

Game-based learning environments are designed to enhance positive cognitive and affective outcomes among learners by embedding curricular content in gameplay [4]. Collaborative game-based learning environments integrate collaborative elements such as group chat so that students can interact with each other, which can potentially increase engagement [3, 4]. However, disengagement may appear throughout the learning process in game-based learning environments and the impact is even higher in a collaborative learning space as students' behavioral disengagement can distract other students, impeding the learning process and engendering negative attitudes [6]. Prior work has characterized different types of disengagement behaviors and their impacts [5], but

identifying disengagement behaviors is challenging, as what constitutes disengagement is often dependent on the context and modality of the learning environment [2].

This paper introduces a multimodal behavioral disengagement detection framework leveraging facial video recordings and group chat logs from middle school students and investigates the effectiveness of different modalities when identifying disengagement behaviors in a collaborative game-based learning environment, CRYSTAL ISLAND: ECOJOURNEYS. We labeled the two modalities for disengagement behaviors and investigated the impacts of features from both modalities for predicting disengagement. Results show that multimodal models can achieve higher predictive accuracy when automatically detecting disengagement behaviors compared to unimodal baselines.

2 Multimodal Disengagement Modeling

A classroom study was conducted with 26 middle school students (8 females; 18 males) as they interacted in a collaborative game-based learning environment for ecosystem science, CRYSTAL ISLAND: ECOJOURNEYS. In the game, the students were divided into groups of three or four (total 7 groups) to determine the cause of the sudden sickness of fish on a remote island using in-game chat and a collaborative whiteboard system. A total of 2,560 chat messages and 44 hours of facial video recordings were collected.

We first annotated both modalities (group chat and video recording) as *engaged* or *disengaged*. Two raters marked all the chat messages that are either content or task related as *engaged* (35.35%) and the rest as *disengaged* (κ=0.925). Next, for tagging the video recordings, two raters used the HELP coding framework [1] with a window size of 10 seconds [7] and tagged each window as *disengaged* whenever a student was disengaged for more than 4 seconds (κ=0.763). The final labels contain 4,863 disengaged segments (30.85%) and 10,898 engaged segments (69.15%). To combine the chat-based labels with video-based labels, we considered the "window of a chat" to be a window of 10 seconds after a chat message was sent. We combined the labels using three different heuristics: *chat_first* prefers *chat-based* labels whenever there is a chat message; *engaged_first* prefers engaged behavior whenever there is disagreement between *chat-based* and *video-based* labels; and, *disengaged_first* prioritizes disengaged behavior whenever there is disagreement.

For chat-based features, we transformed the tokenized in-game chat messages into distributed vector representations using ELMo. For each chat message, we averaged a 1024-dimension vector over all tokens to generate a single word embedding. We refer to these as *chat_only* features. For video-based features, we utilized the action unit features, pose features, and gaze features (total 49 features) from the OpenFace[1] behavior analysis toolkit. We refer to these as *video_only* features. We also created a multimodal feature set called *all* features that combines *chat_only* and *video_only* features.

[1] https://openface-api.readthedocs.io

3 Results and Discussion

To compare our feature groups across different labels based on different modalities, we utilized three off-the-shelf classifiers, namely, Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). All classifiers were trained to predict engagement (1) and disengagement (0) behaviors among students on individual 10-second time segments. We repeated all experiments with three random seeds, each with five-fold cross-validation, and only report the mean scores.

Table 1. Precision, Recall, and F1 scores (in percent) for unimodal and multimodal features across multiple labels using different classifiers. Highest F1 scores are bolded.

		***chat_only* unimodal features**			***video_only* unimodal features**			***all* multimodal features**		
Labels	**Classifier**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
chat-based	LR	76	69	73	62	33	43	77	71	**74**
	DT	60	62	**61**	45	42	43	58	61	59
	RF	78	63	**70**	64	34	44	79	61	69
video-based	LR	16	1	1	74	44	55	72	47	**57**
	DT	20	2	3	60	56	58	61	57	**59**
	RF	34	1	1	79	60	**68**	80	56	66
chat_first (combined)	LR	77	12	21	72	43	54	75	52	**61**
	DT	61	11	18	58	54	56	60	56	**58**
	RF	77	10	18	78	58	66	79	59	**68**
engaged_first (combined)	LR	25	1	1	74	43	54	73	47	**57**
	DT	20	1	2	59	54	57	60	56	**58**
	RF	39	0	1	79	59	**68**	79	57	66
disengaged_first (combined)	LR	74	13	22	73	45	55	74	53	**62**
	DT	59	11	19	59	55	57	61	57	**59**
	RF	75	13	21	78	60	68	79	60	**69**

For unimodal features, we can see that *video_only* unimodal features outperform *chat_only* unimodal features in all cases except for *chat-based* labels. Only for *chat-based* labels, *chat_only* unimodal features are significantly better at predicting disengagement behaviors than *video_only* unimodal features. This shows that *video_only* features are dominant in predicting disengagement behaviors when disengagements are defined either using *video-based* annotations or using a combination of *video-based* annotation and *chat-based* annotations. But *video_only* features are ineffective in predicting *chat-based* disengagement behaviors. When *all* features are used, most of the classifiers perform better than the unimodal features, indicating that combined labels achieve better predictive scores when both modalities of features are present.

A potential explanation for *chat_only* features performing worse in most cases maybe the low number of segments that contain chat messages (2,560) when compared

against the total number of segments (15,761) that contain *video_only* features. As for other limitations, we defined disengagement behaviors using 10-second segments for *video-based* annotations. Previous work also used different sizes of segmented windows to define student behavior [7], but in reality, such behaviors are continuous.

4 Conclusion

Collaborative game-based learning environments offer engaging learning opportunities for groups of students. However, students may become disengaged from the learning process, and these disengagement behaviors can be detrimental to the learning process of individuals as well as their groups. Automatically detecting disengagement behaviors is challenging as they are often difficult to identify with a single modality. We have introduced a multimodal behavioral disengagement detection framework that leverages in-game chat messages in conjunction with facial video recordings of individual learners to detect disengagement behaviors among students. The results show that the multimodal behavior disengagement detection framework outperforms unimodal models for detecting student disengagement. These findings suggest that the multimodal behavior disengagement detection framework can inform the design of adaptive scaffolding for collaborative game-based learning environments.

Acknowledgements. This work is supported by the National Science Foundation through grants IIS-1839966, and SES-1840120. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Aslan, S., Mete, S.E., Okur, E., Oktay, E., Alyuz, N., Genc, U.E., et al.: Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. Educational Technology. 53–59 (2017).
2. Giannakos, M.N., Sharma, K., Pappas, I.O., Kostakos, V., and Velloso, E.: Multimodal data as a means to understand the learning experience. International Journal of Information Management. 48, 108–119 (2019).
3. Jeong, H., Hmelo-Silver, C.E., and Jo, K.: Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. Educational research review. 28, 100284 (2019).
4. de Jesus, Â.M., and Silveira, I.F.: A collaborative game-based learning framework to improve computational thinking skills. In: 2019 International Conference on Virtual Reality and Visualization (ICVRV). pp. 161–166 IEEE (2019).
5. Langer-Osuna, J.M.: Productive disruptions: Rethinking the role of off-task interactions in collaborative mathematics learning. Education Sciences. 8, 2, 87 (2018).
6. Park, K., Sohn, H., Mott, B., Min, W., Saleh, A., Glazewski, K., et al.: Detecting disruptive talk in student chat-based discussion within collaborative game-based learning environments. In: Proceedings of the 11th International Learning Analytics and Knowledge Conference. pp. 405–415 (2021).

7. Thomas, C., and Jayagopi, D.B.: Predicting student engagement in classrooms using facial behavioral cues. In: Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education. pp. 33–40 (2017).