Using Multimodal Sensing and Neural Networks to Predict and Detect Confusion

Tristan Chavez
Regis University
tchavez003@regis.edu

Sarah Smith
University of Rochester
ssm180@u.rochester.edu

Shweta Wahane
Rochester Institute of Technology
sw9910@rit.edu

Reynold Bailey and **Cecilia O. Alm** and **Alex Ororbia**
Rochester Institute of Technology
{rjbvcs, coagla, agovcs}@rit.edu

Abstract

Confusion is a complex emotional affective state that can impact anyone in everyday life, but is especially common now with the rise of online learning and its lack of individualized attention. Thus, it is a worthwhile and important effort to create computer devices that can detect confusion in order to provide effective interventions. Our study expands on previous research on the subject by devising an extensively multimodal Recurrent Neural Network (RNN) model that can accurately predict and detect confusion during a complex collaborative task. Several datastreams are collected and synchronized: screen-based eye tracking and eye-tracking glasses, motion capture, facial action units, galvanic skin response, and audio transcripts. A self-reported review task applies confused/not confused labels to the data. An RNN is then trained to recognize confusion patterns from these modalities.

Introduction

Machine learning can be used to develop a computational model which predicts confusion based off of the input of multiple modalities. Such information is valuable, especially in a world with a sizeable reliance on remote and asynchronous learning. To be able to adjust and assess the impacts of teaching techniques, having technology that detects confusion is an important development. Many past studies have focused on confusion in relation to a single modality, such as facial analysis or language features, but our study aims to combine many more modalities to produce a stronger and more human-like confusion detection model. Included in our study is a galvanic skin response (GSR) sensor, facial tracking software, two types of eye tracking, body tracking, and speech analysis creating the data library to develop a neural network model.

Confusion can be produced as a result of cognitive disequilibrium, which is to say, the result of some form of contradiction or unexpected anomaly (Lehman et al. 2011). In the interest of inducing confusion in such a way, the setup of our study was as follows: Two subjects in separate labs attempt to cooperate on building a structure out of large building blocks. One subject builds while the other relays instructions, but half of the Builders had been given instructions to be deliberately uncooperative before commencement of the study. This would then cause the necessary cognitive disequilibrium to induce confusion on the Instructor. The complexity of the build, on the other hand, would cause confusion in the Builder.

The data collected from each sensor must be synchronized so all data properly lines up sequentially. This prepares it to be input into a neural network which can carry out the confusion detection by using multiple modalities.

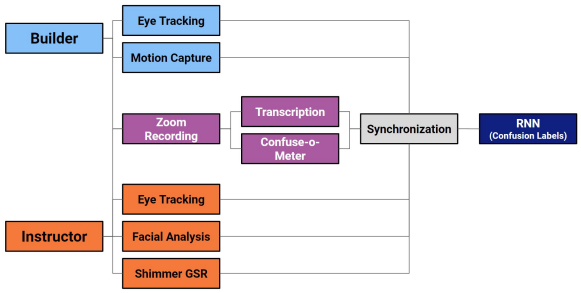

Figure 1: A chart showing the flow of all input streams from both participants and processing course of the data.

This project was designed to answer the following research questions:

- RQ1: Which features from each type of data are most relevant to detecting confusion?
- RQ2: Does a multimodal dataset perform better than a unimodal one in detecting confusion? How many and which modalities are ideal?

Some examples of data which have been examined in successful uni- or bi-modal machine learning models are facial expressions and eye tracking. This guides our research on features to examine for confusion detection. In relation to facial expression, actions which have been closely associated with confusion are inner and outer brow raising, brow lowering, squinting and the appearance of dimples (Grafsgaard et al. 2013). To evaluate a multimodal vs unimodal approach to confusion detection, the most helpful measures will come from the accuracy scores of the neural networks

on the test data. To reach the aforementioned goals, the modalities would be analyzed individually and in conjunction with one another in the neural network to evaluate the individual modalities' impacts on the accuracy in classification of confusion.

Relevant Prior Work

Work on deep learning and using neural networks in multimodal practical problems has become popular both academically and in industry in recent years. Ramachandram and Taylor (2017) give a detailed survey of recent advances in the field, where the importance of multimodal regularization methods and ways to find optimal deep multimodal architectures are stressed. Main takeaways of their work include the knowledge that using multiple modalities in machine learning nearly always results in better performance, deep learning methods are powerful in their allowance for flexible fusion approaches, and a cost function that enforces inter- and intra-modality is key, all of which are aspects we will consider in this project. A further consideration regards what to do when a modality is severely missing from multimodal data. This problem was covered by Ma et al. (2021) who worked with the SMIL method, which uses Bayesian meta-learning that allows both the training and testing models to be flexible with the amount of available modalities. A reconstruction network approximates data in the event of a missing modality, after which features are fed into a regularization network and used in jointly optimizing all of the networks. This method demonstrated consistent efficiency across different datasets with different ratios of missing modalities in a step forward to eliminating the need for complete sets of modalities being necessary to train and test in a neural network. This method is something that could be applicable to our data in further work, but is not employed in the current study.

Work on confusion has strongly leaned towards a learning environment context. With the rise of Massive Open Online Courses (MOOCs) in recent years, it has become important to be able to sense and resolve student confusion without individual human attention. Several studies have looked at the recognizable appearance of confusion in MOOC discussion forum posts, focusing on linguistic features as well as clickstream data to be able to classify confused posts. The practical goal of such work is to develop tools to automatically detect confusion and apply interventions ("just-in-time support") that allow just enough confusion to have a beneficial learning effect but a quick resolution of that confusion to keep students from becoming frustrated and disengaging. Atapattu et al. (2019) found—through a MANOVA test with linguistic features as independent variables and post confusion category as independent variables—the significantly different variables between confused and non-confused posts to be number of pronouns, question stem, opinion, confusion expressions, and incomplete expressions. They tested several classifiers and found Random Forest to be the best performing model, measured by 10-fold cross-validation. Atapattu et al. (2020) found similar results looking at language and discourse features and analyzing the students' reactions from various articles rated on a 1-7 scale of confusion in an MOOC setting, getting 79%-92% accuracy from a Naïve Bayes Random Forest to predict confusion. Yang et al.'s (2015) study attempting to quantify the effect of confusion on dropout from MOOCs looked at linguistic features (weighted by classifier) and click data (which shows context patterns). Logistic Regression was used for models and performance was evaluated with 10-fold cross validation, and several model versions considering different combinations of the input were considered, with a reduced all inputs + unigram feature set performing the best. Further, survival analysis was performed using parametric regression and a Weibull distribution, which showed that "(1) the more students express their confusion and are exposed to confusion in the MOOC forums, the less likely students are to remain active in the learning community; (2) helping resolving or providing responses to student confusion reduces their dropout in the courses; (3) the extent to which different types of confusion affect dropout is determined by specific courses."

Shi et al. (2019) also looked at academic confusion, but their study focused on facial expressions. There was a self-report aspect where participants chose between binary confused and non-confused options for their experience. A CNN + SVM model was used to classify facial features for confusion. They only collected 2D facial images to be fed into the feature extractors which allows room for growth in the study of confusion detection, which we hope to contribute to.

Most importantly, this project is a direct descendant of two previous studies looking at modeling confusion during collaborative tasks, Kaushik et al. (2021) and Mince et al. (2022). Kaushik et al.'s (2021) study investigates bimodal vs. unimodal machine learning for detecting confusion, with facial and linguistic features being the two modalities under investigation. Participants completed two tasks (an impossible scheduling task and a logical reasoning problem) together over Zoom and then independently rated their confusion levels on a 5-point scale for 30-second spans of time during a replay of their previous work. A Random Forest algorithm was utilized for the machine learning aspect, with three tests—training on each individual modality, then the bimodal data, with 100 trees and an 8:2 test-train split, and Gini impurity to partition the data for feature importance analysis. It was found that the facial action units AU4 (brow furrow) and AU7 (lid tighten) were the most strongly correlated to confusion, corroborating previous research on the topic. For linguistic features, token-based features, questions, and silent pauses were most important. Additionally, the model which considered both modalities performed better than individual ones.

Mince et al. (2022) furthered this research by creating a multimodal model which differed from past studies by also considering audio information (prosody, pauses, etc.). There were three confusion-eliciting tasks that were assigned to the participants and their text, speech audio, and video-based facial expressions were recorded. A playback video was then shown to each participant and they were given the task to self label using an in-house designed web app called the Confuse-o-Meter at what points they were confused on a sliding scale. This tool provided significant improvement

over the review method for previous studies by using continuous time rather than discrete time segments. Our project borrows and utilizes this same Confuse-o-Meter tool. This information was stored in a model that implements (1):

$$p(y_t|x_0^m, x_1^m, ..., x_t^m; \Theta_m) = f^m(x_t^m; \Theta_m) \quad (1)$$

where y_t is the time vector where the participant was confused at time t and m is the modality with {vis, aud, txt}. Θ_m contains all learnable weight parameters. This Recurrent Neural Network model was trained with 5 epochs to disallow overfitting and validated using one random male and female. The model found that facial expressions such as an inner brow raise was a good indicator of confusion but audio information was a less effective one. Ultimately, this project designed a neural network model that was capable of accurately predicting confusion and measures how well text, audio and video were accurate predictors of this cognitive state. In our work, the multimodal aspect of Mince et al.'s (2022) project will be extended to look at several more modalities and a physical task rather than cognitive ones, but will use a similar machine learning framework.

Data Collection

Subjects

32 participants were recruited for the study, coming in as 16 groups of two, with 13 female, 16 male, and three non-binary participants aged 18-58. 28 participants spoke US English as their first language and four participants did not. Participants were equally split into two roles: Builders and Instructors, which paired together to form groups. The Builders were further divided equally into Cooperative and Uncooperative assignments, this aspect being kept secret from the Instructors.

The participants were not initially informed about which emotion the study was focused on beforehand, and the participants' consent forms informed them that they were not being told everything about the study in order to cooperate with IRB standards. All participants were debriefed about the purpose and the additional instruction given for uncooperative Builders after completion of the the demographic survey which came after the second task.

Tasks

Each group was instructed to complete two building block tasks, where the Builder and the Instructor were in different rooms, the Builder having the blocks and the Instructor having a picture of the intended structure. The Instructor had to give instructions on how to build the structures. The first task was a simple tower and time capped at two minutes, intended just to get participants used to the procedure. The second task was a complicated truck, time capped at five minutes, the complexity of which and the lack of direction on how to put together the wheels were intended to elicit confusion on the Builder's part. During the second task, half of the Builders were instructed to be uncooperative with their Instructor for three 30-second segments, queued by one of the experimenters holding up a brightly colored block in their line of view but behind the camera. This deliberate uncooperative behavior (picking up the wrong colored block, taking apart the structure, etc.) was meant to elicit confusion on the Instructor's side. We found that several Instructors from uncooperative groups reported feeling frustration more than confusion, which prompted the addition of a frustration measure to the study. This measure was not examined in the current analysis, but the data collected could be used in future work.

Upon completion of the building tasks, sensor recording was turned off and participants individually filled out a general Google Forms survey covering demographics, the effectiveness of their group's communication, and the extent to which various emotions were felt during the previous tasks. This survey was not utilized in data analysis.

Participants were then debriefed before continuing to the fourth task: individually using the Confuse-o-Meter tool developed by Mince et al. (2022) to continuously rate their confusion level while watching the Zoom recording of their group completing the first two tasks. It was important that each subject complete this on their own, as confusion levels during a collaborative task can vary greatly between interlocutors (Kaushik et al. 2021). Following this, participants watched the recording again, this time rating frustration with our modified Frustrate-o-meter. The base-85 output of these tools was collected and stored.

Technical Setup

Each group worked together from two separate rooms. The Builder room's setup involved eye tracking and body motion capture. Eye tracking was done using SMI eye-tracking glasses and the iView and BeGaze softwares, and motion capture was done using the program FreeMocap with three webcams directed at the Builder at different angles. FreeMocap splices these angles to create a 3D skeleton of the subject's movements. Additionally, there was a fourth webcam and a monitor dedicated to the Zoom call. Builders wore a microphone and stood in front of the cameras behind the table that held the blocks.

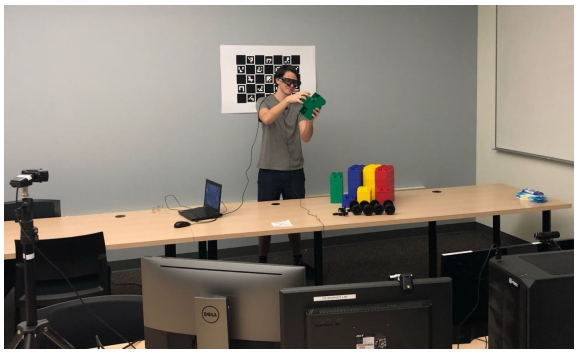

Figure 2: Image of Builder's side of the setup, shown from behind the mocap cameras.

The Instructor's part was carried out in a WhisperRoom sound booth, and utilized screen-based eye tracking, facial

analysis, and a Shimmer galvanic skin response sensor. iMotions was used to collect and synchronize data from these modalities. Instructors looked at a monitor with the screen split, half showing the Builder on Zoom and half showing a picture of the task's intended structure. Additionally, one of the experimenters in the Builder's room acted as the Zoom Facilitator, with the job of recording the Zoom and keeping track of when tasks started and were completed for later synchronization.

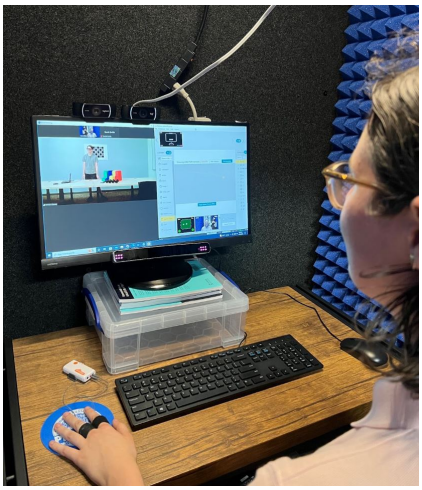

Figure 3: Image of Instructor's side of the setup, showing their split screen view of the Builder and instructions.

Methods

Data Preprocessing & Synchronization

The output of the Confuse-o-Meter is a base-85 string that, when decoded, provides tuples with timestamps and corresponding confusion levels that range from 0 (not at all confused) to 3 (extremely confused) as self-reported by the participants. The Confuse-o-Meter's start and end times are synchronized with the input streams from the iMotions data, FreeMocap output, SMI eye tracking software, and dialogue transcript (which took the Zoom-recorded audio and processed it through IBM Watson Speech-to-Text and the deep-disfluency Python package (Hough and Schlangen 2017)) in preparation for data analysis. The particular features selected for each modality are shown in Table 1. These were selected either because they have been associated with confusion or strong emotion in prior research, because they help us know where a participant is looking at and therefore where they are focusing attention at a given point, or because they gave promising or changing values in naive examination. We note that the transcript data was unable to be implemented in the dataset due to how long it took at process and the project's time constraints.

To facilitate analysis, important packages such as TensorFlow, Keras, Pandas, and NumPy were implemented to process the data into tensors. Tensors are functional-based multidimensional arrays which aid in the training of the model, which will assign labels based on these input streams with

Role	B or I
Mocap	Raw Data (flattened Frame x Position x Axis)
Eye-Tracking Glasses	Event Category, Index Binocular, Pupil Diameter, Area of Interest Name, Gaze Vector (x, y, z)
Dialogue Transcript	*Disfluency Features*, Current Sentence Length, Speech Rate, Content (multi-hot encoded)
Facial Action Units	Attention, *Brow Furrow*, Dimpler, Jaw Drop, *Lid Tightener*, Mouth Open, Nose Wrinkle
Screen Tracking	Coordinates, Fixations (x, y)
GSR	Conductance

Table 1: Utilized features for each modality. Features in italics have been associated with confusion in prior research.

the goal of detecting confusion in the test data. First, after organizing the data into .csv files based on subject, the data from each modality was cropped so that the start of their individual data stream began at the start of the first task. From there, each individual modality's dataset was made into a separate tensor with the features on the x-axis, timestamp (in seconds) on the y-axis, and subject enumerated on the z-axis. The data was then normalized with a min-max range function in the tensor in preparation for the creation of a final tensor which had every modality from every subject. The role of the participant was either given to the computer as a 0 or 1 at the beginning of the stream of data so that the program would adjust accordingly to which modalities would be significant during the machine learning process. After this datapoint came all the features extracted from the other modalities which were appended to the tensor, thus creating one large tensor which was trimmed according to the task times in conjunction with the tensor containing the confusion labels. Before being fed into the RNN model, the subjects were split up for three-fold cross validation with a test set of 2 subjects, validation set of 3 and with the remaining 26 subjects encompassing the training set. Each subject was then split up into sets of varying time lengths called "windows" for batching and faster performance. The window size was a variable that changed depending on the experiment being run.

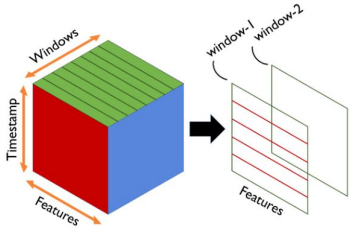

Figure 4: Diagram of the tensor structure for the data input.

Recurrent Neural Network

A Recurrent Neural Network (RNN) model was then trained to recognize the patterns from these modalities that correspond with the levels of confusion (Figure 1). RNNs are designed to handle sequential data so an RNN model is best suited for this project as every modality provides sequential data which can be processed as a tensor array. The RNN was trained with the formulas:

$$h_t = \emptyset(W * x_t + V * h_{t-1} + b)$$

$$\hat{y} = softmax(U * h_t + c)$$

These formulas corresponds with the diagram in Figure 5, with W, V and U representing the weights initialized with the Gaussian random number distribution and t adjusted according to the loss calculated by categorical cross entropy recorded with the gradient tape function. The constants in the function are a and b, representing the bias of the function and ŷ is the prediction of the confusion on a categorical vector of length 4 based on the input of x_t. Adam optimizer was chosen with a learning rate of 0.001 to apply the gradients to the function.

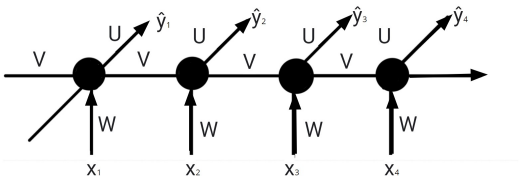

Figure 5: Diagram showing the RNN model unrolled with weights applied.

Results

Hyperparameter Experimentation

Nine experiments were run on the data set to determine the best performing hyperparameters. We used a learning rate of .001 for all experiments and altered the window size with the set 100, 150, 200 and hidden nodes from the set 32, 64, 128, trying all unique combinations to find the lowest loss on the validation set. 250 iterations of the RNN were allowed to run for each experiment. Once the lowest loss was produced by the RNN, the instance of its class with the weights would be pickled, along with a text file which specified the parameters and epoch of the lowest loss. The accuracies in Table 2 were produced from the RNN's predictions on the validation when the loss function produced its lowest number. The graphs of accuracy and loss are shown below, notice the machine properly learning the training data but having a more difficult time with the validation set which did improve over time after a steep initial decline.

	Hidden Layer Nodes		
Window Size	32	64	128
100	0.67	0.68	0.69
150	0.68	0.67	0.70
200	0.59	0.66	0.68

Table 2: Table of validation accuracies for each combination of window size and number of hidden layers.

Test Set

After determining that the window size of 150 and hidden node size of 128 performed the best on the validation set based off the accuracies in the table above, the test set would be validated using the instance of the RNN that provided that result. This resulted in the test set prediction providing an accuracy of .68.

Discussion

The accuracies in Table 2 demonstrate little variability based on the hyperparameters set before each trial. It should, however, be noted that the specific instance when the accuracies are recorded is at the iteration during which the RNN produces the lowest result from the loss function. This means that the accuracy could still increase while the loss continues to increase. Even though the two are seen as inversely related, this is not always true as verified by our

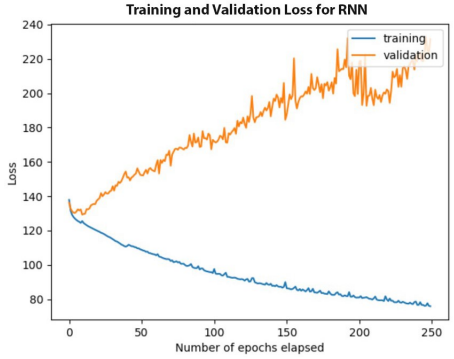

Figure 6: Outputs of the loss function up to 250 iterations.

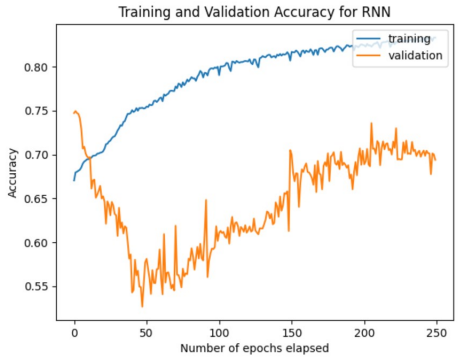

Figure 7: Accuracy of the validation and training sets.

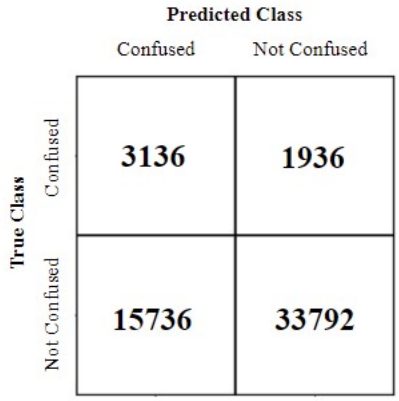

Figure 8: Confusion matrix of the test set.

RNN. Figures 6 and 7 demonstrate this principle as accuracy and loss can operate somewhat independent of one another. The confusion matrix provided in Figure 8 for the test set demonstrated a large number of false positives, which means the model developed is biased towards predicting that a subject is confused at a specific instance of time. A quantitative analysis of this bias gives us a low precision value of .17 but a better recall of .62. The best case scenario for the algorithm was predicting when a subject was not confused (a true negative). To improve on this, an experiment setup that induces more confusion would be beneficial since as demonstrated by the test set the experiment largely did not elicit a balanced amount of confused and not confused data points as the test set was composed of 17% confused data points which may not be enough for this RNN to function off of which was further imbalanced in other subject trials.

Limitations and Future Work

The major limitation of this work thus far comes from the project's time limit. Due to this constraint, we were unable to get to answering our second research question and the opportunity window for tuning the RNN was bounded. In further work, we hope to perform more testing and experiments on hyperparameters and possibly in altering the RNN implementation itself with the aim of improving its accuracy. It is possible that a smaller window size may lead to better results, but machine processing limitations barred us from trying anything under 100 and, furthermore, impacted the number of experiments that could be run.

Next steps include the addition of the transcript data into our model, a more systematic selection of features to provide a tailored answer to RQ1, and experimentation with various combinations of modalities in pursuit of answering RQ2. systematic addition and removal of features would be useful as some modalities could be providing more noise to the machine rather than helping in the learning process. FreeMocap for instance is responsible for over 1,600 features fed into the machine and that only composes one modality for the builder whereas the next closest modality is a tie between the SMI eyeglass data (which were also for the builder) and iMotions facial analysis which have 7 features each. This imbalance of data points between builder and instructor could heavily impact the performance of the RNN especially since it is trained on 26 subjects which can be thought of as 13 pairs.It would be beneficial to have at least 20 sets of participants to allow the algorithm access to more varied data which would improve its predictions.

Conclusion

This project did produce a recurrent neural network that tries to detect confusion in response to multimodal data, however its accuracy and especially precision could be improved substantially with further review and refinement. Currently the model does over perform the most common class classifier but it likely can improve further in future iterations. Several of the features used in our model are possibly not actually correlated with confusion and may bring too much noise to the data or be inversely related. A closer look into what parts of the data collected are actually statistically relevant to confusion would help resolve that flaw. The current model does the best performance-wise in predicting the true negative of the subject being not confused which did make up the majority of the self-reported confusion labels. This RNN had a low precision score but it is something that may be improved on with an adjustment of optimizer and learning rate in future experiments.

The secondary product of this study is an extensively multimodal, synchronized, and high-quality set of data which is a feat in of itself. Although the number of participants is still on the smaller side and a few participants had a single modality fail, a modality such as the audio features can be useful to other researchers in future projects revolving around confusion detection as this is a modality which has data for 32 subjects and their corresponding confusion labels.

Ethics Statement

This research was conducted with approval from an institutional review board. All subjects gave informed consent before participation and were told that they had the authority to stop the study if they became uncomfortable at any point. Participant data remained anonymous and demographics were collected to ensure a broad subject pool for the data in order to properly train the models without biases. All equipment was sanitized with antibacterial wipes after every session of data collection for the safety of subjects. Instructors were limited to 10 minutes behind a closed door in the booth due to the potential for the room to induce claustrophobia.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591. 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. We also acknowledge and thank Rajesh Titung for his generous help and support on this project.

References

Atapattu, T.; Falkner, K.; Thilakaratne, M.; Sivaneasharajah, L.; and Jayashanka, R. 2019. An identification of learners' confusion through language and discourse analysis. *CoRR* abs/1903.03286.

Atapattu, T.; Falkner, K.; Thilakaratne, M.; Sivaneasharajah, L.; and Jayashanka, R. 2020. What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs. *IEEE Transactions on Learning Technologies* 13(4):878–888.

Grafsgaard, J. F.; Wiggins, J. B.; Boyer, K. E.; Wiebe, E. N.; and Lester, J. C. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. In *EDM*.

Hough, J., and Schlangen, D. 2017. Joint incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 326–336". Association for Computational Linguistics.

Kaushik, N.; Bailey, R.; Ororbia, A.; and Alm, C. O. 2021. Eliciting confusion in online conversational tasks. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 1–5.

Lehman, B.; D'Mello, S. K.; Strain, A. C.; Gross, M.; Dobbins, A.; Wallace, P.; Millis, K.; and Graesser, A. C. 2011. Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In Biswas, G.; Bull, S.; Kay, J.; and Mitrovic, A., eds., *Artificial Intelligence in Education*, 171–178. Springer Berlin Heidelberg.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: multimodal learning with severely missing modality. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2302, 2310.

Mince, C.; Rhomberg, S.; Alm, C.; Bailey, R.; and Ororbia, A. 2022. Multimodal modeling of task-mediated confusion. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 188–194. Association for Computational Linguistics.

Ramachandram, D., and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34(6):96–108.

Shi, Z.; Zhang, Y.; Bian, C.; and Lu, W. 2019. Automatic academic confusion recognition in online learning based on facial expressions. In *2019 14th International Conference on Computer Science & Education (ICCSE)*, 528–532.

Yang, D.; Wen, M.; Howley, I.; Kraut, R.; and Rose, C. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, 121–130. Association for Computing Machinery.