2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 979-8-3503-2745-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACIIW59127.2023.10388144

Early vs. Late Multimodal Fusion for Recognizing Confusion in Collaborative Tasks

Anisha Ashwath[1], Michael Peechatt[1], Cecilia Alm[1,2], and Reynold Bailey[1]
[1]*Golisano College of Computing and Information Sciences,* [2]*College of Liberal Arts*
Rochester Institute of Technology
Rochester, USA
{aa6106, mp6510, coagla, rjbvcs}@rit.edu

***Abstract*—There has been a rapid transformation in the medium of learning and communication due to the pandemic. Multitudes have adopted online video platforms to learn and work from any corner of the world. Emotion detection is vital for understanding how well instructions are communicated through online interactions and for building cognitive systems that can identify human behavior. Confusion is a key emotion that can impact online learning and can be used to verify whether students using an online platform understand the material being taught. Our research expands on previous work regarding confusion detection, focusing on data fusion techniques. We explore the impact of early fusion (feature-level) vs late fusion (decision-level) on modeling confusion identification during a collaborative block building task. Experimenting with different classifiers, our results show that late fusion performs better with larger time windows. This fusion approach can aid in model interpretability.**

***Index Terms*—data fusion, multimodal data, affective computing, early fusion, late fusion**

I. Introduction

Humans can experience various emotions during communication. There may be rapid changes in the types and intensity of emotions experienced in a collaborative context. The pandemic has brought a substantial change in the medium of teaching, with online learning gaining popularity in many educational institutions. Affective computing specializes in developing systems that can identify and reproduce human emotions. Among the variety of emotions users may experience during online interactions, we focus on confusion. For the context of this study, we define confusion as a state of mind where an individual is uncertain about the information communicated to them. Compared to happiness, sadness, or anger, confusion is a more subtle affective state and poses a challenge [1]. This study seeks to identify this ambiguous emotion using multimodal data.

Data fusion techniques handle the combination of disparate data types for inference tasks. We perform our experiments on the MULTICOLLAB corpus [2], a heavily multimodal dataset involving pairs of participants working on collaborative block building tasks over Zoom. Subjects then provide timestamped, self-annotated instances of confusion from the video recording of the call. Sensor data surrounding the labeled timestamp are sampled based on two parameters: $time\_window$ and $time\_dimension$. The former determines how much sensor data to select, and the latter defines the number of averaged time steps to divide from the selection. This paper poses the following research questions: **RQ1**: Which data fusion technique, early or late fusion, performs better in detecting confusion? **RQ2**: Does $time\_window$ or $time\_dimension$ play a more important role in influencing the performance of model accuracy?

II. Related Work

Hori et al. [3] studied the confusion a car driver can experience. The corpus was collected with various sensors, including video cameras, car state sensors, and a GPS navigation system. They inferred that LSTM performed better for its use of context, indicating that time-series signals can be used to detect driver status. Our work focuses on confusion experienced in a task-driven setting. We make use of biophysical sensing data including facial expression, eye movement, and galvanic skin response. Kaushik et al. [4] also explored confusion in dialogue tasks. They annotated the confusion labels in intervals, while our work uses continuous annotation for the confusion labels. Shi et al. [5] examined confusion in an academic context, with coursework offered through Massive Open Online Course (MOOC) platforms. Facial movements were recorded across several videos to detect confusion. The experiment involved audio, video, text, and electrodermal activity. For the data collected and used in our study, participants communicated to each other directly through a Zoom call. Gunes and Piccardi [6] examined data fusion techniques and explored which ones led to better recognition of facial features, including whether fusion at the feature-level (early fusion) or fusion at the decision-level (late fusion) performed better. In our work, we focus on finding which data fusion technique performs better on a heavily multimodal corpus.

Affective computing and its application towards building human-like systems have been explored in prior work. Pantic and Rothkrantz [7] advocated for the incorporation of emotion recognition in building cognitive systems. By using multimodal affective signals, human-computer interaction (HCI) design decisions can be informed by nonverbal feedback cues. Barnum et al. [8] have argued that multimodal processing happens almost immediately, i.e., at the feature-level, for humans. In the case of video analysis with multimodal data, Snoek et al. [9] provided evidence that late fusion performs better than early fusion. In our work, we explore a multitude of configurations for both early and late fusion with varying time windows and time dimensions.

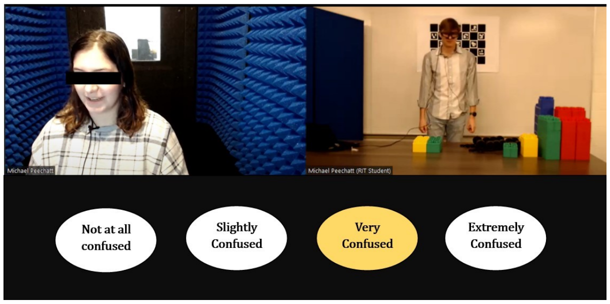

Fig. 1. Participants annotate their instances of confusion using the Confuse-O-Meter, an annotation tool which synchronizes ratings with timestamps.

III. Dataset

The MULTICOLLAB corpus [2] captured time-series sensor data in a collaborative scenario. The dataset is comprised of 48 participants (24 groups), 27 male and 20 female and 1 undisclosed. Pairs of individuals worked together to complete two block-building tasks. Each pair consisted of an instructor and a builder. Among the builders, they were further divided into cooperative and non-cooperative groups. The cooperative builders followed all the instructions given by the instructor, while the non-cooperative builders were privately told to disobey the instructor's directions to induce confusion. All groups were assigned two tasks. The first task was a simple six-block structure, used to familiarize subjects with the experimental setup. The second task was a much more complicated thirteen-block structure which utilized wheel components.

TABLE I
The modality features used in early and late fusion experiments.

Voice Features	
Intensity (dB)	F0 (Hz)
Facial Features	
Brow Furrow	Chin Raise
Lid Tighten	Lip Corner Depressor
Eye Gaze Features	
Saccade Duration	Saccade Peak Velocity
Fixation Dispersion	Fixation Duration
Gaze Velocity	
Biophysical Features	
GSR Conductance	

Features were extracted from various sensors including screen-based eye tracking, TASCAM microphones, webcam recordings, and Shimmer Galvanic Skin Response (GSR). These were further processed using z-score normalization to allow for comparison across participants. Table I shows the features used in our experiments, grouped by modality. Instructors were then asked to watch their own Zoom recording and annotate timestamps where they were confused. Confusion ratings were self-reported by the instructors on a scale of *Not At All Confused*, *Slightly Confused*, *Very Confused*, and *Extremely Confused*. Figure 1 shows the Confuse-O-Meter annotation tool used by the participants.

A. Rating Distribution

The counts of instructor confusion ratings separated by group type is shown in Figure 2. This distribution indicates an imbalance, with *Slightly Confused* being the most frequent label and *Extremely Confused* being the least frequent. This label imbalance increased the classification challenge. Additionally, the rating counts seemed unaffected by the cooperation of the builder, with there being more instances of *Slightly Confused* and *Very Confused* in cooperative groups. This implies the difficulty of the tasks induced confusion more effectively than the disobedience of the builder.

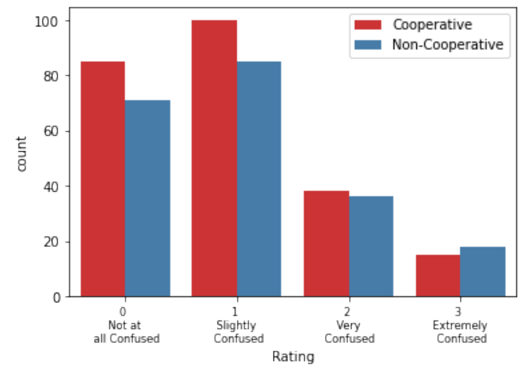

Fig. 2. Label distribution of Confuse-O-Meter instructor ratings for both cooperative and non-cooperative groups. Whether or not builders obeyed instructions seemed to have little impact on rating.

B. Instructor Utterances

Word tokens were transcribed from instructor microphone recordings. Figure 3 visualizes the words captured around annotated instances of confusion. The top shows the frequency of words for *Not At All Confused* instances, and the bottom shows the frequency of words for both *Very Confused* and *Extremely Confused* combined instances. The words in each word cloud can provide a better understanding of the kind of words used during an online call, as well as which ones can be leveraged for identifying confusion. Disfluencies or filled pauses like *ah, uh, um, uhm* could potentially indicate that the instructor was confused at that point in the recording, though further analysis seems needed. While lexicon from the utterances were considered in this visual inspection, they were not used in this study for developing classifiers.

IV. Method and Results

There are three ways to combine data using fusion: early fusion, late fusion, and intermediate fusion. Our experiments focus on early and late fusion. Data-level fusion (early fusion) involves combining modality features before the prediction task. Early fusion captures the interactions between modalities through time. In contrast, decision-level fusion (late fusion) is a technique where each modality is separated and analyzed independently, then later combined to obtain predictions. This method loses interactions between the modalities but has

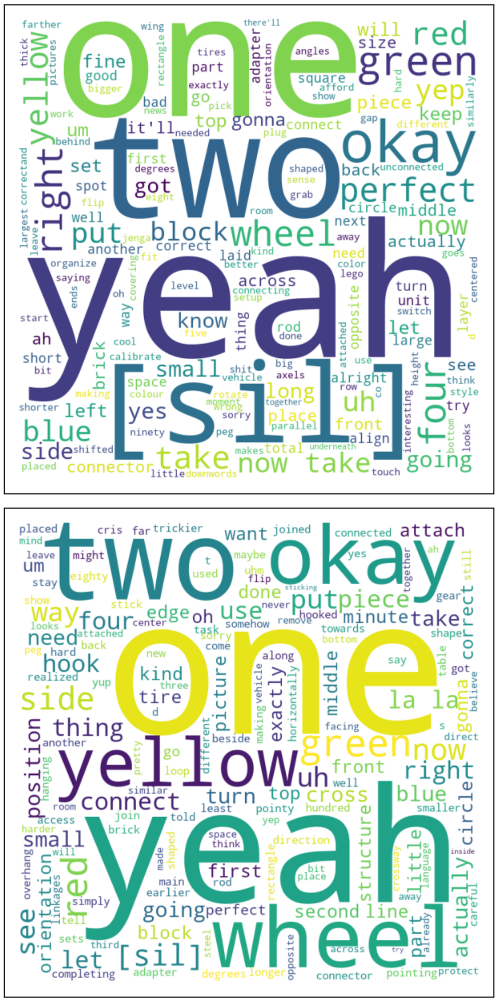

Fig. 3. Word clouds for *Not At All Confused* (top) and the combination of *Very Confused* and *Extremely Confused* (bottom). The [sil] token (representing silence) is larger in the former class, indicating instructors may have been more inclined to speak when they felt confused.

improved interpretability when compared to early fusion. Because we can see an independent prediction for each modality, this may provide improved interpretability regarding how the final prediction was made.

A. Early Fusion

For our dataset, early fusion is performed by concatenating the modality features through time. We use two parameters to represent modality sensor values: $time\_window$ represents the amount of sensing data captured before and after the timestamped rating, and $time\_dimension$ represents the number of splits for averaged time steps. Various configurations of both parameters were used to compare the influence on classification performance. The following $time\_window$ values were used: 1500 ms, 2000 ms, 2500 ms, 3000 ms, 3500 ms, 4000 ms, 4500 ms, and 5000 ms. The following $time\_dimension$ values were used: 10, 15, and 20. We performed our experiments on the Cartesian product of all $time\_window$ and $time\_dimension$ configurations.

B. Late Fusion

Late fusion was performed by considering each feature as a different dataset. The features extracted from each of the modalities had its own trained model, which was then fused with other model predictions at the decision level. An ensemble classifier was built using each model output, as each modality prediction was used to make a final inference. Both hard voting and soft voting were used in building this ensemble classifier. Hard voting selected the class with the highest number of votes, while soft voting averaged the probabilities of each prediction from each model and selected the class with the highest total probability.

TABLE II

ACCURACY COMPARISON ACROSS DIFFERENT CLASSIFIERS FOR $time\_window$: 4500 MS AND $time\_dimension$: 10, WHERE E_F IS EARLY FUSION, L_FHV IS LATE FUSION WITH HARD VOTING, AND L_FSV IS LATE FUSION WITH SOFT VOTING.

Model	E_F Acc	L_FHV Acc	L_FSV Acc
Random Forest	56.8	**75.0**	**72.7**
XGBoost	**65.9**	65.9	65.9
Decision Tree	27.3	56.8	56.8
Logistic Regression	38.6	38.6	38.6
SVM	43.2	43.2	34.1

C. Computational Modeling Methods

Models using Scikit-learn implementations were used to evaluate the performance of both early and late fusion techniques. Logistic Regression, SVM, XGBoost, Decision Tree, and Random Forest were used to build the models and compare their performance. Accuracy was used as a metric to measure the performance. Our experiments were performed using the configurations outlined in Section IV-A to compare the accuracy of the classifiers. Table II shows the results of different classifiers for the configuration $time\_window$: 4500 ms and $time\_dimension$: 10. Experiments were performed using a development data split of 80%-20% for training and validation. A test set of three groups had been set aside to evaluate accuracy.

Early fusion performed better than late fusion using the configuration $time\_window$: 4000 ms and $time\_dimension$: 10. Using the Random Forest classifier, the early fusion technique was able to achieve an accuracy of 68%. Figure 4 shows the results for late fusion performing better than early fusion using the configuration $time\_window$: 4500 ms and $time\_dimension$: 10. Using the Random Forest classifier for individual models and a hard voting classifier on their output, this late fusion technique was able to achieve an accuracy of 75%, giving the best results overall.

V. DISCUSSION

Results show that late fusion provided better results in most cases with an increasing $time\_window$. The best results were obtained for the smallest $time\_dimension$: 10. Late fusion can provide better prediction interpretability, even though it loses the interactions between modalities through time. The results obtained show that late fusion performs better in

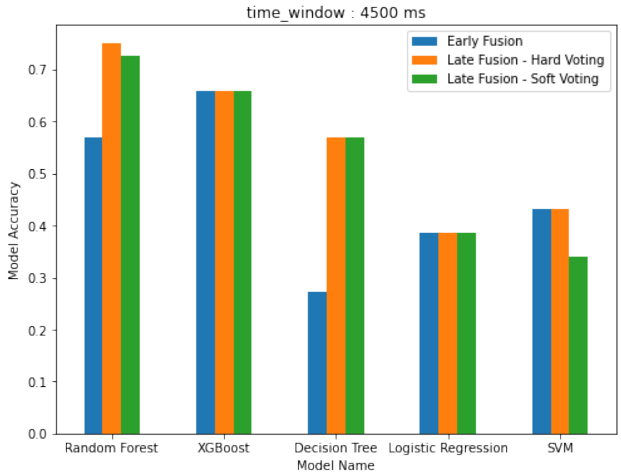

Fig. 4. Results of early fusion and late fusion for $time\_window$: 4500 ms and $time\_dimension$: 10. The late fusion model of Random Forest performs the best overall with an accuracy of 75%

most cases, but there are configurations where early fusion performed better. A limitation of our research is the modest size of the data, making it difficult to ascertain generalizable results. Confusion is also a subtle and ambiguous emotion to predict and may add complexity compared to other emotions such as happiness or anger. For future work, word embeddings derived from spoken dialogue can be used to provide a richer representation when combined with other sensing data. Incorporating transcribed words can also provide insight as to which lexical items or utterances contribute to distinguishing certain emotions. Expanding the dataset with more participants could be beneficial for learning more about data fusion and its role in detecting complex emotions, such as confusion, using multimodal data. Finally, to address **RQ1**, the late fusion technique performed the best in most cases giving an overall best accuracy of 75% using the Random Forest classifier. For **RQ2**, varying $time\_window$ added more context in detecting confusion and three different dimensions were also considered, of which $time\_dimension$: 10 yielded the best results for both early and late fusion.

VI. Conclusion

This paper explored the concepts of data fusion with multimodal data. Two different forms of data fusion techniques were used: early fusion and late fusion. In exploring which fusion technique yielded better performance, late fusion performed better in most cases with larger time windows. However, there were cases where the early fusion performed better with somewhat smaller time windows. Working with different subsets of the data gave insight into how fusion prediction accuracy was influenced by the hyperparameters. Varying $time\_window$ and $time\_dimension$ values also helped in determining which fusion technique yielded better results. The experiments discussed can be further expanded by developing neural network models for the early and late fusion approaches and increasing the volume of the dataset with more labeled instances.

Ethical Impact Statement

The dataset was collected in an IRB-approved study that used informed consent prior to subjects' participation. They were given the option to end the study at any time. During the setup, the equipment used for collecting each modality was explained to the participant before data collection began. After providing annotations, both builders and instructors were debriefed on the nature of the study, data use, and whom to contact about the study. Each participant was compensated $25 USD for participating in the study.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. 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] O. AlZoubi, *Automatic Affect Detection From Physiological Signals: Practical Issues*. PhD thesis, 09 2012.
[2] M. Peechatt, C. Alm, and R. Bailey, "MULTICOLLAB: A multimodal corpus of dialogues for analyzing collaboration and frustration in language," tech. rep., Rochester Institute of Technology, 2023.
[3] C. Hori, S. Watanabe, T. Hori, B. A. Harsham, J. Hershey, Y. Koji, Y. Fujii, and Y. Furumoto, "Driver confusion status detection using recurrent neural networks," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016.
[4] N. Kaushik, R. Bailey, A. Ororbia, and C. O. Alm, "Eliciting confusion in online conversational tasks," in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–5, 2021.
[5] Z. Shi, Y. Zhang, C. Bian, and W. Lu, "Automatic academic confusion recognition in online learning based on facial expressions," in *2019 14th International Conference on Computer Science & Education (ICCSE)*, pp. 528–532, 2019.
[6] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3437–3443 Vol. 4, 2005.
[7] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.
[8] G. Barnum, S. J. Talukder, and Y. Yue, "On the benefits of early fusion in multimodal representation learning," in *NeurIPS 2020 Workshop SVRHM*, 2020.
[9] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, (New York, NY, USA), p. 399–402, Association for Computing Machinery, 2005.
[10] S. W. McQuiggan, S. Lee, and J. C. Lester, "Early prediction of student frustration," in *Affective Computing and Intelligent Interaction* (A. C. R. Paiva, R. Prada, and R. W. Picard, eds.), (Berlin, Heidelberg), pp. 698–709, Springer Berlin Heidelberg, 2007.
[11] D. Guo, W. Zhou, H. Li, and M. Wang, "Online early-late fusion based on adaptive hmm for sign language recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, dec 2017.
[12] C. Mince, S. Rhomberg, C. Alm, R. Bailey, and A. Ororbia, "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*, (Hybrid: Seattle, Washington + Online), pp. 188–194, Association for Computational Linguistics, July 2022.