2022 International Conference on Advanced Learning Technologies (ICALT) | 978-1-6654-9519-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICALT55010.2022.00033

Tackling the design and evaluation of a theater-based intelligent system to monitor audience experience in virtual public speaking settings

Eduardo Rivero-Rodríguez
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
eduriver@ucm.es

Pablo Villalobos-Sánchez
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
pavill01@ucm.es

Meriem El-Yamri
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
melyamri@ucm.es

Alejandro Romero-Hernández
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
alerom02@ucm.es

Borja Manero
Dept. of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Madrid, Spain
bmanero@ucm.es

***Abstract*— COVID-19 has brought about a sharp increase in the use of videoconferencing tools. In education, this complicates the monitoring of student experience, which is essential to perform adequate classroom management. Researchers have designed tools to aid teachers within on-site settings, but they focus only on student engagement and are not suitable for virtual environments.**

In this paper, we present our system's architecture and evaluation. First, we adapted a theater-based framework to measuring audience experience beyond engagement in online settings. Secondly, we designed a proof-of-concept computer vision system and a companion video conferencing tool to automatically measure audience experience in the classroom and present near real-time feedback. We also describe the experiment we conducted to obtain a dataset to test our system and present the results. Although the predictive accuracy of our proof-of-concept system is limited, it opens several directions for future research.

Keywords*— *computer vision, machine learning, sentiment analysis

I. INTRODUCTION

The relationship between students' classroom behavior and their academic outcomes is not new [1]. A recent meta-analysis [2] analyzed 69 independent studies on the topic and established a positive correlation between students' engagement in the classroom and their academic achievement and learning. The recent COVID-19 pandemic and the move to virtual settings have made this task even harder. It is much easier for a teacher to identify when an in-person audience is getting bored than when a virtual one is.

Every technical solution we found in the literature measures only one specific dimension of audience experience: engagement. Engagement is a construct that includes observable behaviors, internal cognition, and emotions. However, audience experience goes beyond engagement and, therefore, it seems necessary to incorporate other factors to achieve a more accurate model.

Theater is a discipline that has something to say about audiences. A report commissioned by different theater associations [15] identified five major components of audience experience in the theater: engagement, learning, energy, shared atmosphere, and emotional connection.

Taking that into account, we propose a system capable of providing real-time feedback by monitoring and analyzing the audience of online meetings. Our system has 3 differentiating factors: it extends engagement-based frameworks with a theater-inspired approach, it is designed to operate within online learning environments, and it provides near real-time feedback. In addition, we show an initial proof of concept, which we were able to test in a small-scale experiment with mixed results.

This paper is structured as follows. First, we introduce the conditions that make virtual settings different and explain our framework for quantifying audience experience. Then, we give a general overview of the architecture and a more detailed technical explanation. After that, we describe the experiment performed and the results obtained. Finally, we highlight the conclusions and limitations of our system.

This project has been funded by the Ministry of Science, Innovation and Universities of Spain (Didascalias, RTI2018-096401-A-I00).

978-1-6654-9519-6/22/$31.00 ©2022 IEEE
DOI 10.1109/ICALT55010.2022.00033

II. THE SYSTEM

A. The context of virtual settings

Meetings in virtual settings differ significantly from in-person meetings: only the head and shoulders are visible, there may be variations in video quality (resolution, lighting, etc.) and each video stream may fluctuate or disconnect over time. These differences prompted us to adapt the framework proposed by the report on audience experience in the theater [15] by removing the shared atmosphere dimension, using affective response [4] to measure energy, and incorporating the research of Curtis, Jones, and Campbell on engagement [3] and comprehension [5]. Thus, our adapted framework is composed of affective response (Af), engagement (En), emotional connection (Ec), and learning (Le).

One challenge we faced when setting up a data pipeline for this framework is that most widespread applications did not have a simple way of treating the video feeds in a meeting room as separate input data streams. To overcome said challenge, we developed an ad hoc video conferencing application based on the WebRTC API [6] (see Fig. 1).

Fig. 1. Respectively, participant view (left) and host view (right) side by side.

B. Architectural overview

The general architecture is composed of a central server, a host for the video conference (the presenter's device), and multiple clients (the participants' devices). The server runs the machine learning (ML) pipeline, which outputs a metric value for each participant and each dimension of audience experience.

These metrics are then pooled and can be displayed to the host as feedback individually (Af, En, Le, Ec), or aggregated as a global compound score (S). The metrics for each dimension correspond to the average across all participants, and the compound score (S) is the average of the 4-dimensional scores. Depending on the context, it might make sense to add weights for each dimension to calculate the compound score.

The ML pipeline first performs face detection and tracking in each frame, associating a persistent identity with each participant in the batch and keeping track in future frames (see Fig. 2). This is necessary to incorporate information from previous batches.

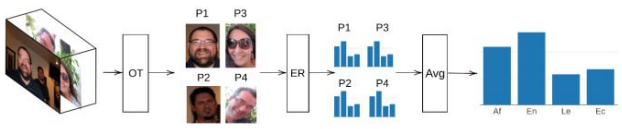

Fig. 2. Overview of the system architecture (OT – Object Tracking, ER – Emotion Recognition, Avg - Average).

Then, the streams associated with each identity enter the emotion recognition module, which takes a batch of frames and outputs the value of the 4 metrics for each frame in the batch. These metrics represent the scoring determined by the emotion recognition module for each dimension of audience experience. Finally, a pooling step returns the metrics for the whole batch.

C. Architectural implementation

The object tracking module works in 2 steps. First, we apply standard YOLOv4 [7] to detect participants in the frame. We created a custom algorithm to track specific participants. This algorithm keeps track of the positions of the object bounding box centers and box sizes across frames and uses the variation in the last 2 frames to predict the bounding box centers and sizes in the next frame.

With the predicted values for bounding box center (c_{n+1}) and size (bb_{n+1}) and with each bounding box bb being represented as a pair of height and width (h, w) in pixels, we match existing identities to the observed participants with the most similar bounding box characteristics, minimizing the metric

$$m(c, bb) = \left|\left|c - c_{n+1}\right|\right| + f(h, h_{n+1}) + f(w, w_{n+1}) \quad (1)$$

where $f(x, y) = log\ (1 + |x - y|)$.

To obtain the matching efficiently, other systems [8] use the Hungarian algorithm, which runs in time complexity $O(n^3)$ [9]. We have found an alternative approach that remains largely unexplored in the literature and is only implemented by a few select systems [10]. If we formulate the problem above as a stable matching problem by converting the distance matrix determined by metric $m(\cdot,\cdot)$ into preference lists for the previous detections and the detected object, we can use the Gale-Shapley algorithm [11] to perform the matching in $O(n^2)$.

The emotion recognition module also works in 2 steps. First, we use MobileNetV3 [12] to obtain a low-dimensional embedding of the input frames. This embedding is then passed onto 4 fully connected layers with linear activation, each trained to predict a single dimension-specific score (Af, En, Le, Ec). Finally, the pooling module extracts global scores by averaging across the number of agents. We can calculate the summary metric by averaging the dimensional scores.

III. EXPERIMENT, DATASET AND RESULTS

A. Experimental design and participants

While there are some audiovisual datasets on audience experience and affective response [3], [13], they are not available to the general public and do not contain data in online settings. For this reason, we ran a data collection experiment and created a dataset of online audience response.

A total of 8 last year Spanish college students (6 males and 2 females, all around 20 years old) volunteered to take part in the study. The experiment was conducted as follows: the volunteers connected to our custom video conferencing tool with their personal computers and webcams. We streamed three

TABLE I. RESULTS ON THE TEST DATASET

	MSE	MAE	MAPE	R^2
Aff. Resp.	0,0861	0,2557	1,2240	-0,1142
Engagement	0,1556	0,3093	14,0547	-0,3560
Em. Con.	0,0757	0,2324	0,7135	-0,4725
Learning	0,1775	0,2906	6.3201	-0,4725
Combined	0,1099	0,2642	1,5040	-0,2330

short (10min) video performances while we recorded participants with their webcams. After watching each performance, the participants filled out a short questionnaire.

After data collection, the recordings were cleaned up and score labels were generated for each frame by linear interpolation from questionnaire results for each video.

B. Experimental results

For the final training, we used an Adam optimizer [14] with a learning rate of 0.0005, L1 loss, and batch size of 256. The model was trained for 2 epochs using 72% of the data, while the remaining data, recordings belonging to 2 participants, was used for validation (14%) and testing (14%). To assess the performance of the model we have tracked Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R^2 score (See Table 1).

As we can see, MSE is relatively low, while MAPE is very large. However, since the error for each frame is between 0 and 1, this is to be expected and is mostly an artifact of working with small numbers. The R^2 is negative, which implies that our model performs worse than a constant prediction that matches the mean of the test dataset. However, we must interpret this result considering the data generation process: since the scores were interpolated from three points, the variance in the dataset is very low, which might explain such a low R^2 score.

All in all, it seems that MAE best reflects the actual performance of the system. This metric indicates that our system is making average errors of around 0.3. This is not very precise but may allow the speaker to rule out extreme situations (e.g., very low or very high engagement).

IV. Conclusions and Future Work

In this paper we proposed and evaluated a general framework that can be used to design intelligent systems to automatically evaluate audience experience in virtual settings. The framework is based on how the theater world evaluates their audiences. It goes beyond the one-dimensional (engagement-based) current trend and specifies four dimensions: affective response (Af), engagement (En), emotional connection (Ec), and learning (Le). Besides, we specified a particular implementation to predict audience experience scores. We also described the experiment we carried out to obtain a dataset and test our system and presented the final results.

The proposed 4-dimensional framework captures aspects of the audience experience that were not considered in the one-dimensional measurements. An audience may be highly engaged but fall short in terms of learning. In the same manner, an audience might not be fully engaged but still have a high degree of affective response. The ability to capture these subtleties makes the proposed framework a better choice than engagement-based ones to see the full picture when it comes to measuring audience experience. At the same time, while existing systems were designed for in-person use, the proposed system is designed specifically for virtual settings.

Acknowledgment

We would like to thank all the volunteers who helped us build our dataset and, therefore, enabled us to work on the predictive system.

References

[1] J. D. McKinney, J. Mason, K. Perkerson, and M. Clifford, "Relationship between classroom behavior and academic achievement," *J. Educ. Psychol.*, vol. 67, no. 2, pp. 198–203, Apr. 1975, doi: 10.1037/h0077012.

[2] H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Soc. Behav. Pers.*, vol. 46, no. 3, pp. 517–528, 2018, doi: 10.2224/sbp.7054.

[3] K. Curtis, G. J. F. Jones, and N. Campbell, "Effects of good speaking techniques on audience engagement," in *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, Nov. 2015, pp. 35–42, doi: 10.1145/2818346.2820766.

[4] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994, doi: 10.1016/0005-7916(94)90063-9.

[5] K. Curtis, G. J. F. Jones, and N. Campbell, "Speaker impact on audience comprehension for academic presentations," in *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Oct. 2016, pp. 129–136, doi: 10.1145/2993148.2993194.

[6] "WebRTC." https://webrtc.org/ (accessed Feb. 16, 2021).

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv*, Apr. 2020, Accessed: Feb. 16, 2021. [Online]. Available: http://arxiv.org/abs/2004.10934.

[8] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T. K. Kim, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, 2021, doi: 10.1016/j.artint.2020.103448.

[9] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist. Q.*, vol. 2, no. 1–2, pp. 83–97, 1955, doi: 10.1002/nav.3800020109.

[10] A. B. Godbehere and K. Goldberg, "Algorithms for visual tracking of visitors under variable-lighting conditions for a responsive audio art installation," *Control. Art Inq. Intersect. Subj. Object.*, pp. 181–204, 2014, doi: 10.1007/978-3-319-03904-6_8.

[11] D. Gale and L. S. Shapley, "College Admissions and the Stability of Marriage," *Am. Math. Mon.*, vol. 69, no. 1, p. 9, 1962, doi: 10.2307/2312726.

[12] A. Howard *et al.*, "Searching for MobileNetV3," *arXiv*, 2019.

[13] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, 2012, doi: 10.1109/T-AFFC.2011.25.

[14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, 2015.

[15] Review of *Capturing the Audience Experience: A Handbook for the Theatre*. n.d. New Economic Foundation. https://itc-arts-s3.studiocoucou.com/uploads/helpsheet_attachment/file/23/Theatre_handbook.pdf.