PG23

Daejeon, Korea

Emotion-based Interaction Technique Using User's Voice and Facial Expressions in Virtual and Augmented Reality

Beom-Seok Ko[1] ,Ho-San Kang[1], Kyuhong Lee[1], Manuel Braunschewiler[2] , Fabio Zünd[2] , Rober Sumner[2] and Soo-Mi Choi[1]

1. Department of Computer Science and Engineering and Convergence Engineering for Intelligent Drone, XR Research Center, Sejong University, Korea

2. Game Technology Center, ETH Zürich, Switzerland

PROBLEM

The metaverse is a virtual space for immersive experiences and communication, with AR and VR technologies gaining importance. They find applications in various fields like healthcare, education, and industry. AI characters and NPCs are crucial for enhancing immersion, and incorporating AI enables natural dialogues and emotion recognition through facial expressions in VR. Recent advancements in AI, like ChatGPT, offer potential for lifelike interactions in AR/VR, enhancing user experiences based on emotions and conversation recognition.

RELATED WORK

An increasing number of efforts are underway to utilize both AR and VR technologies within immersive environments. Instead of employing each technology on separate devices, researchers are exploring ways to enhance user experiences by integrating them into a single device [1]. Furthermore, research on the use of Non-Player Characters (NPCs) in AR/VR is on the rise. Studies are investigating interactions with NPCs, such as generating interactions through simple non-contact hand gestures in AR/VR, aiming to improve user trust and immersion [2]. Additionally, research on communication methods involving NPCs and facial expressions is also gaining traction. [3] demonstrates the potential for users to communicate with virtual humans by recognizing facial expressions in VR environments.

OVERVIEW

This study proposes an interaction method in AR/VR environments, where non-player characters (NPCs) immerse themselves in each environment based on the user's emotions. This paper distinguishes between AR and VR environments, employing different approaches to gauge user emotions. In the AR environment, a natural language processing model, RoBERTa, is used during user-NPC conversations to identify emotion-related keywords in user dialogues. These keywords are then used to ascertain emotions, subsequently altering the style or attributes of specific objects for interaction. In the VR environment, user emotions are inferred by monitoring facial expressions using an HMD's eye tracker and an additional facial tracker for capturing eye and lower facial movements, vectorizing these values. The vectorized data serves as input for EmotionNet, determining the user's emotions and subsequently modifying the VR environment accordingly.

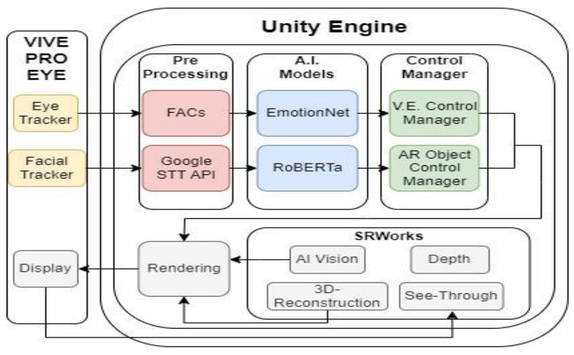
ACKNOWLEDGEMENT

This research was supported in part by the MOTIE and KIAT through the International Cooperative R\&D program (No. P0016038) and in part by the MSIT and IITP (IITP-2023-RS-2022-00156354).

초실감 XR 연구센터
eXtended Reality Research Center

Game Technology Center

METHODOLOGY



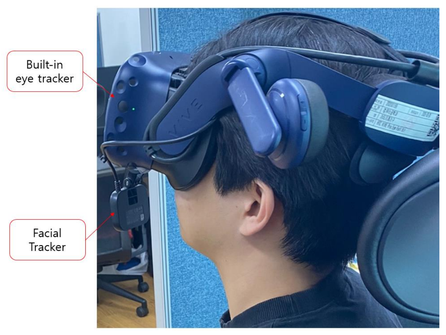
This system is an interactive system that detects emotions from the user's voice and facial expressions for interaction. The interactions of this system can be broadly categorized into AR environment interactions and VR environment interactions. To implement interactions in both AR and VR, we conducted research using VIVE Pro Eye and Facial Tracker.

In the case of AR environment interaction, Non-Player Characters (NPCs) discern user emotions through voice during conversations. After identifying the user's emotions, specific object styles or attributes are modified to engage with the user. To facilitate AR environment interaction, we first scanned the user's surroundings using the HMD's camera and Unity's SRWorks to locate objects for interaction.

The emotions that NPCs can detect through conversation are limited to anger, fear, joy, love, sadness, and surprise, and the research proceeds with these six emotions. Voice can provide important cues for emotion detection.

For example, users can provide direct emotional clues, such as "I'm feeling down right now" or "I love you." To classify emotions, the system uses the RoBERTa natural language processing model. To utilize the user's voice as input for this model, the user's speech is converted to text using Google's Speech-to-Text (STT) API. The RoBERTa model classifies emotions based on the converted voice text, and these emotions are used to modify the style or attributes of specific objects. Another means of emotion detection is facial expressions. In the VR environment, facial expressions are used to infer user emotions.

However, since users wear an HMD, only eye movements, mouth, and lower face movements are tracked to recognize facial expressions. The eye tracker embedded in the HMD and the additional Facial Tracker device are used to weight eye and lower facial movements.



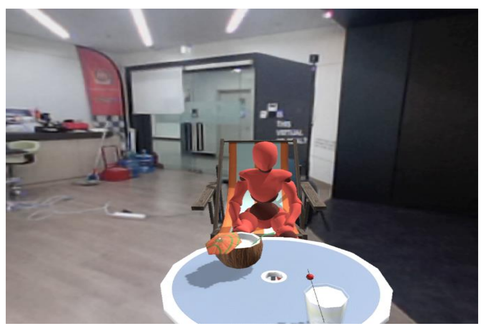
These weights are vectorized as feature vectors for the eyes and lower face, serving as inputs for a CNN-based classification model to perform emotion classification.

RESULTS

In this study, we have proposed an interaction method based on the user's emotions in both AR and VR environments. To measure the accuracy of interactions, we conducted experiments using pre-prepared scripts with NPCs instead of free-form conversations. In the AR environment, we accurately scanned the user's surroundings using the HMD's built-in camera and the SRWorks library, achieving precise object recognition, including items such as chairs and tables. User voice recognition, facilitated by Google's API, enabled real-time conversion of speech into text, with an observed error rate of approximately 4.8%.

For emotion classification using text, RoBERTa was employed. While the pre-trained model exhibited an average accuracy of over 80% on well-curated text datasets, the actual text converted from speech using the Google STT API resulted in an average accuracy of approximately 61%, likely due to API-related errors and other factors.

In the VR environment, we tracked eye and lower facial movements using the HMD's eye tracker and facial tracker. Tracking lower facial movements presented challenges, such as difficulties capturing changes like tongue movement or excessive mouth opening. Additionally, distinguishing between lips and the tongue when the mouth was widely open led to misinterpretation. Furthermore, predicting facial expressions solely based on eye and lower facial movements, rather than the entire face, demanded higher accuracy. The primary challenge was the intermittent tracking of lower facial movements during the process of vectorizing eye and lower facial movements. The facial expression recognition rate in the VR environment yielded an accuracy of approximately 52%.



In summary, this study has explored the potential for NPC and user interaction in AR/VR environments. As part of our future research, we plan to integrate generative AI technology to create more advanced forms of emotional communication between intelligent characters and users in AR/VR settings. This integration holds the promise of enriching user experiences and expanding the possibilities for emotional interactions within mixed reality environments

REFERENCES

[1] Kang, H., Yang, J., Ko, B. S., Kim, B. S., Song, O. Y., & Choi, S. M. (2023). Integrated augmented and virtual reality technologies for realistic fire drill training. IEEE computer graphics and applications.

[2] Carroll, D. (2022). Attention and Communication in Virtual Worlds: Interacting with Non-Player Characters in Virtual Reality.

[3] Vicente-Querol, M. A., Fernández-Caballero, A., Molina, J. P., González, P., González-Gualda, L. M., Fernández-Sotos, P., & García, A. S. (2022, May). Influence of the Level of Immersion in Emotion Recognition Using Virtual Humans. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 464-474). Cham: Springer International Publishing.