PG23

Daejeon, Korea

Avatar Emotion Recognition using Non-verbal Communication

Jalal Safari Bazargani , Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi*

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

* Corresponding Author

PROBLEM

Traditionally, it was believed that working with computers efficiently required discarding emotions. However, in today's world, emotions play a significant role in almost every computer-related activity. Among the sources of information about emotions, body movements, recognized as "kinesics" in non-verbal communication, have received limited attention, although they offer valuable information from a distance.

This research gap suggests the need to investigate suitable body movement-based approaches for making communication in virtual environments more realistic.

RELATED WORK

More than 50% of human communication is conveyed through non-verbal components, which is a testimony to the large amount of research on these components [1]. With regard to body movements, it has been demonstrated that a person's gait or body expressions can influence how others perceive their feelings. In addition, body movements can be considered a better approach when emotion recognition is required from a distance. That said, this area of research has not been explored extensively [2].

OVERVIEW

This study proposes an automated emotion recognition approach suitable for use in virtual environments, wherein avatars and digital characters can communicate.

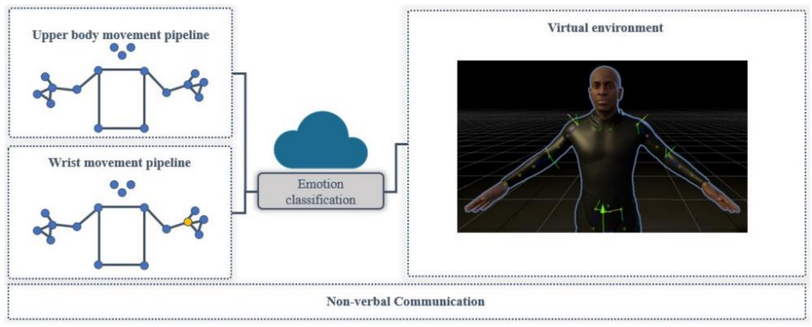
This paper consists of two pipelines for emotion recognition. For the first pipeline, i.e., upper-body keypoint-based recognition, the HEROES video dataset was employed to train a bidirectional long short-term memory model using upper-body keypoints capable of predicting four discrete emotions: boredom, disgust, happiness, and interest, achieving an accuracy of 84%. The virtual environment was then created for an avatar, and 17 keypoints were extracted with the help of the meta-movement software development kit as the input features of the trained model, which was stored on a server offering a web socket connection. For the second pipeline, i.e., wrist-movement-based recognition, a random forest model was trained based on 17 features computed from acceleration data of wrist movements along each axis. The model achieved an accuracy of 63% in distinguishing three discrete emotions: sadness, neutrality, and happiness. The findings suggest that the proposed approach is a noticeable step toward automated emotion recognition, which is appropriate for current devices and networks, without using any additional sensors other than the head mounted display.

ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC), Supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2023-RS-2022-00156354; and in part by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative Research and Development Program under Project P0016038.

세 종 대 학 교
초실감 XR 연구센터
EXTENDED REALITY RESERCH CENTER

METHODOLOGY



This methodology can be divided into two main pipelines: upper-body and wrist-movement-based recognition. For the upper-body keypoints, the HEROES video dataset was chosen for the experiment. The dataset presented four emotions: boredom, disgust, happiness, and interest. As body language includes different indicators, 17 upper-body keypoints were selected for each frame. The left eye, right eye, and nose were selected for eye movement and head orientation. Head orientation provides valuable information about emotional states. For example, tilting the head may be a sign of interest [3]. After the face, arms are believed to be a great source of body language information. In this regard, five keypoints including the elbow, wrist thumb, index finger, and little finger were chosen for each arm. The remaining four keypoints are indicators of the shoulders and hips, i.e., the torso.

The dataset provided by [4] was used for the wrist-movement pipeline. This dataset was created from participants walking 250 m while wearing a smartwatch and heart rate gadget.

Three discrete emotions: happy, sad, and neutral, were offered in this dataset. For this study, only the accelerometer data were employed because after the initial tests, retrieving the gyroscope data from the controllers seemed unreliable. 17 features were extracted from the accelerometer data along each axis, resulting in 51 features per sample.

The processed data was then regarded as input features for training a random forest model. The remaining procedure is similar to that used in upper-body pipeline, wherein the LSTM model is connected to the virtual scene with the help of a web socket.

Inside the virtual scene, the acceleration was calculated based on the velocity of the left controller, assuming that the data came from a smartwatch. This approach eliminates the need for a smartwatch and an interface to connect the data obtained from the smartwatch to the engine. The data is then preprocessed to extract17 features. Finally, it is passed to the model via a web socket for prediction.

Notably, because the approaches employed in this study do not offer the same discrete emotion recognition, the authors decided to use a condition and switch to the appropriate pipeline. In this condition, the movement of the hips is monitored to determine whether the user is walking or in an almost stationary position. If the walking state is confirmed, the engine switches to using the wrist movement for emotion recognition. Otherwise, the engine continues to use the upper-body keypoints to recognize the emotions of the user.

RESULTS

For the upper-body pipeline, after several experiments, a sequence of five frames yielded the highest accuracy. The trained Bi-LSTM model was evaluated on the same dataset; 10% of the data were selected for testing, which achieved an accuracy of 83.63%.

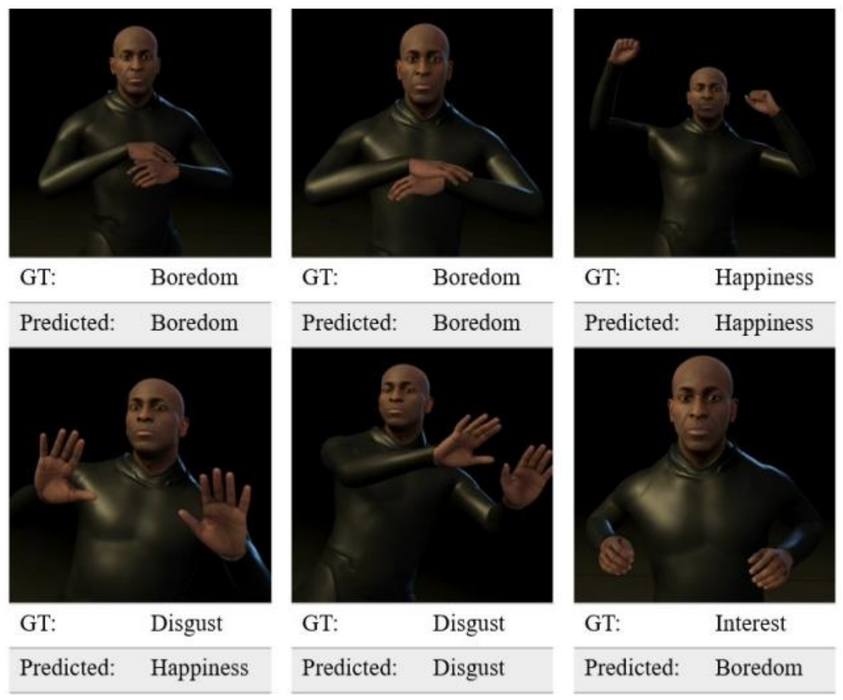
Real-time tests were conducted without any noticeable delay, primarily because the feature extraction part was omitted, and SDK was utilized to retrieve keypoints. This approach also prevents challenges arising from the distance between the point responsible for visual feature extraction and avatar location.

In terms of emotions, for boredom, the associated body language normally involves having the hands parallel to the torso or palms on top of each other. These movements may not be tracked when relying only on the HMD because hand tracking depends on the line of sight, and the head orientation does not allow a line of sight to the hands, in some configurations. In this case, the hands turn into standby mode, and the body configures a boredom posture, resulting in an incorrect prediction by the trained model. This is also somewhat true for happiness, wherein users choose to open and move their arms. In addition to hands covering the head or neck area, there is disgust, backing, and moving to one side. These movements can be similar to happiness in some cases.

Additionally, head orientation plays a significant role in emotional interest. This indication of interest can be misunderstood because of the inability to track hands when the head is oriented forward. This was the case for the incorrect prediction for the emotion of interest.

For the wrist-movement pipeline, an individual model was trained for each participant, and the average accuracy was 62.43% for predicting the three discrete emotions for each individual dataset.

Real-time tests revealed that the feature extraction part was more consistent compared to the upper-body pipeline. In other words, losing the line of sight between the headset and hands in the upper-body pipeline is a major problem that prevents feature extraction. Having said that, feature extraction in the wrist-movement pipeline was more reliable owing to the presence of controllers for retrieving acceleration data.



REFERENCES

[1] Chaudhary Muhammad Aqdus Ilyas, Rita Nunes, Kamal Nasrollahi, Matthias Rehm, and Thomas B. Moeslund. 2021. Deep Emotion Recognition through Upper Body Movements and Facial Expression. VISIGRAPP 2021 (2021), 669–679.

[2] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. 2019. Emotion Recognition from Skeletal Movements. Entropy 2019, Vol. 21, Page 646 21 (6 2019), 646. Issue 7.

[3] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kaminska, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on Emotional Body Gesture Recognition. IEEE Transactions on Affective Computing 12 (1 2018), 505–523. Issue 2.

[4] Juan Carlos Quiroz, Elena Geangu, and Min Hooi Yong. 2018. Emotion Recognition Using Smart Watch Sensor Data: Mixed-Design Study. JMIR Ment Health 2018;5(3): Issue 3.