International Journal of **STEAM**
vol. 15, No. 1, (2024), pp.80~83

Emotion Recognition Device for the Visually Impaired

Seoyun Baek, Jungwoo Brian Kim and Jihoon Oh*
Korean Minjok Leadership Academy, Hoengseong-gun, Gangwon-do, Republic of Korea

bsy2430735@gmail.com, brianjwkim25@gmail.com, ojhoon0621@gmail.com

[Abstract] Communication is a fundamental element in human social interaction. However, approximately 2.2 billion visually impaired individuals worldwide encounter significant obstacles in this regard, as they are unable to perceive the faces of others during conversation. In this study, we developed a device that recognizes the other person's facial expressions and voices through a camera and built-in microphone, extracts emotions through deep learning-based emotion recognition algorithms and operates a vibration sensor according to emotions to convey the other person's emotions to user. Speech emotion recognition algorithm (SER) and facial expression recognition algorithm (FER) showed remarkable prediction accuracy at 98.55% and 97.25%, respectively. For wireless data transmission, an ESP32 board was used, allowing the vibration sensor to operate immediately according to the emotion prediction results of the emotion recognition algorithm. The device is designed in the form of glasses, allowing users to operate it without the use of their hands. Furthermore, the device is easily transportable as it can be wirelessly connected to a smartphone and controlled through a mobile application.

[Keywords] Speech emotion recognition algorithm (SER), Facial emotion recognition algorithm (FER), ESP32, wireless data transmission, vibration sensor, hands-free usage, mobile application

I. Introduction

With the recent development of computer vision technology, including convolutional neural networks, interest in recognition technology is increasing. Among them, Facial Emotion Recognition (FER), which recognizes emotions through the other person's facial expressions, and Speech Emotion Recognition (SER), which recognizes emotions through voice tone changes, are highly utilized in various fields, and accordingly, pre-trained models based on various quality datasets and augmentation methods are being developed.

The recognition technology is ideal for developing devices that help visually impaired people communicate better. Individuals who are visually impaired have difficulty understanding the other person's emotions because they are not able to check the other person's expression during conversation. Emotion recognition devices that utilize recognition technology can instead recognize emotions and display them to the visually impaired, allowing them to communicate better with the other person.

Various attempts have been made to create a device that recognizes emotions and transmits them to the visually impaired. Sungkyunkwan University conducted a study on a Thermal Display-Based Emotional Communication System for the visually impaired and developed a device that classifies emotions into favorable, unfavorable, and normal through a pre-trained model and operates a temperature sensor depending on the classified emotion [1]. However, it is challenging to confirm that accurate emotion recognition was achieved in this study, since the user's subjectivity was a significant factor in determining temperature, and the classification of emotions was too simplistic in comparison to the full spectrum of emotions that humans can experience.

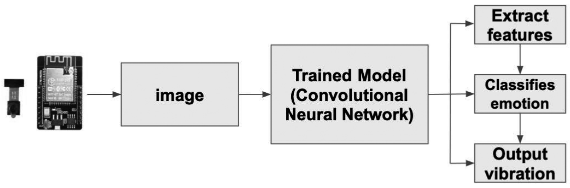
Huawei in China has developed an application that conveys emotions recognizable through facial expressions through sound [2]. However, the sound produced by the device is disruptive to the conversation itself, as it operates too much. Consequently, this device is not considered to be an appropriate means of enhancing the quality of conversation.

This research implemented a model trained on a high-quality dataset with a data augmentation method and deep learning technology for emotion recognition. The model was developed using a vibration sensor to minimize user subjectivity. This resulted in the development of a glasses-shaped wireless device and an application to control the device, allowing users to easily access the device without difficulties.

2. Methodology

2.1 Project design

As illustrated in Figure 1, the operation of the device is divided into two distinct parts: a software component that recognizes facial expressions and voices through a camera installed inside the device and a microphone installed in the user's cell phone, and a hardware component that transmits the extracted emotions to the ESP32 board to operate the vibration sensor according to the type of emotion. The software component utilizes an artificial intelligence model trained based on a convolutional neural network to extract emotions, while the hardware component utilizes the extracted emotions to operate the vibration sensor.



[Figure 1] Overall process of operating emotion recognition device

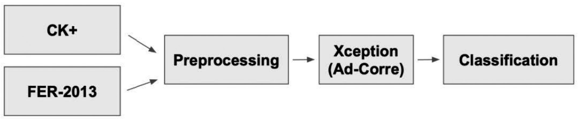
2.2 Emotion Recognition Algorithm

The emotion recognition algorithm combines two distinct sources of emotional data: facial expression recognition and speech emotion recognition. The former identifies and extracts emotions from facial expressions, while the latter employs voice recognition to extract emotional information. These two sources of emotional data are then integrated through an algorithmic process, and the resulting emotional value is transmitted to the ESP32 board via Bluetooth Low Energy (BLE) communication.

2.2.1 Facial Expression Recognition (FER)

In the context of facial expression recognition, the facial expression of the other person is detected through the ESP32 Camera Module. Subsequently, the facial expressions are classified according to emotions through an artificial intelligence model trained using the Ad-core loss technique applied to the Xception model [3] and trained using the pre-trained model [4], following the application of data augmentation to the CK+ dataset and FER-2013.

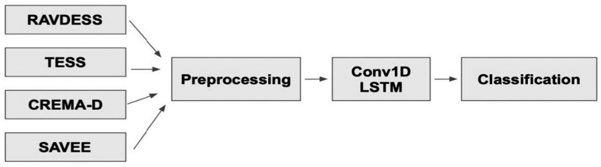
[Figure 2] Esp32 Camera Module



[Figure 3] Process of Facial Expression Recognition

2.2.2 Speech Emotion Recognition (SER)

In the case of Speech Emotion Recognition, the voice is recognized by the microphone built into the user's mobile phone, and the emotions are extracted by a model trained by Conv1D and LSTM based on data augmentation applied to RAVDESS, TESS, CREMA-D, and SAVEE.



[Figure 4] Process of Speech Emotion Recognition

2.2.3 Integration

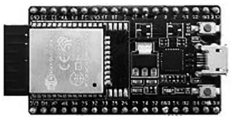
First, the seven extractable emotions (angry, disgust, fear, happy, neutral, sad, and surprise) are divided into three categories. The first category, Positive, includes the positive emotions of Happy and Surprise, the second category, Negative, includes Angry, Disgust, Fear, Sad, and the last category, Neutral.

The emotion recognition algorithm transmits the predicted emotion extraction value from facial expression recognition to the ESP32 board only when the category of the emotion extraction value from facial expression recognition and the category of the emotion extraction value from voice recognition are the same. To simplify, the emotion extraction value by voice recognition is a process of correcting the emotion extraction value by facial expression recognition.

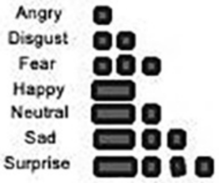
2.3 Hardware Architecture

2.3.1 Vibration sensor Activation with ESP32

After the ESP32 board receives the emotion extraction value through BLE communication, the vibration sensor is operated by varying the number of vibrations according to the emotion. The number of vibrations for each emotion is as follows, and long vibrations were added to happy, neutral, sad, and surprise instead of increasing the number of vibrations so that the number of vibrations does not exceed 4 times.



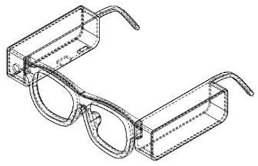
[Figure 5] ESP32 Dev kit for data receiving and transmission



[Figure 6] Vibration signal for each emotion

2.3.2 Hardware Modeling

The hardware was designed using 3D modeling, considering the ESP32 board, the battery for power supply, and the size of the vibration sensor, and it was designed in the form of glasses, considering the user's wearability.



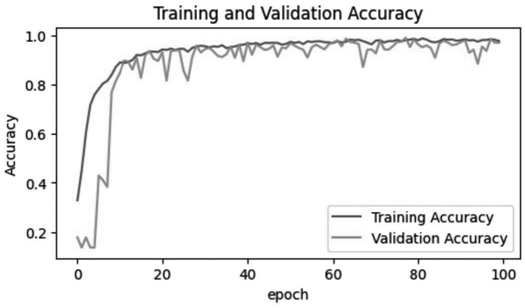
[Figure 7] Device modelled by Fusion360

3. Results

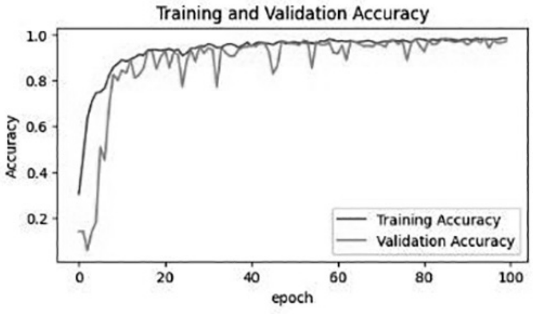
3.1 Emotion Recognition Algorithm

As illustrated in Figure 8, Facial Expression Recognition, trained by the Xception model and applied Ad-Corre loss, achieved an accuracy of 98.55%.

Speech Emotion Recognition, trained by a one-dimensional Convolutional Neural Network and Long Short-Term Memory, achieved an accuracy of 97.25%, as illustrated in Figure 9.



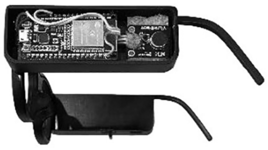
[Figure 8] Accuracy of Facial Expression Recognition (FER)



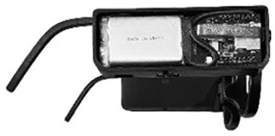
[Figure 9] Accuracy of Speech Emotion Recognition (SER)

3.2 Hardware Development

The device modeled as Fusion360 was printed via the ANICCUBIC Photon M3. The left part of the device, shown in Figure 10, consists of a vibration sensor and ESP32 dev kit that receives emotions from an emotion recognition algorithm. The right part consists of an ESP camera module for facial expression recognition, a lithium-ion battery to power the device, and a switch and charger module. More details are shown in Figure 11.



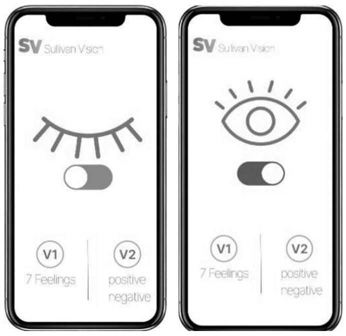
[Figure 10] Structure of the device (left side)



[Figure 11] Structure of the device (right side)

3.3 Mobile Application

It was developed using the Flutter framework to enable BLE communication between mobile phones and devices. It supports conversion to simple mode, which only receives signals with three emotions (positive, negative, and neutral).



[Figure 12] User interface of mobile application

4. Conclusion

To minimize the inconvenience of communication experienced by the visually impaired, this study has developed a device that costs less than $35 and transmits the other person's emotions to the user through wireless communication using an emotion recognition algorithm with a delay of 0.002 seconds and 98% accuracy. In the future, research will focus on reducing the weight of the device and increasing the battery life through battery research.

Acknowledgements

This work was supported by Korea Science Service (KSS, ksef@ksef.or.kr)

References

[1] Noh, J. (2013). Development of a thermal display for material representation. Journal of Korean Information Processing Society, 20(5), 123-130. http://cg.skku.edu/pub/korean/papers/2013-noh-kips-thermal-display.pdf

[2] designboom, sofia lekka angelopoulou I. (2018, December 28). Huawei releases AI-powered app that translates emotions into sounds for the blind and visually impaired. https://www.designboom.com/technology/huawei-app-facing-emotions-for-the-blind-12-30-2018/

[3] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.195

[4] Fard, A. P., & Mahoor, M. H. (2022). Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access, 10, 26756–26768. https://doi.org/10.1109/access.2022.3156598

Seo-yun Baek

- 2022.2~ Korean Minjok Leadership Academy

Jungwoo-Brian Kim

- 2021.2~ Korean Minjok Leadership Academy

Ji-hoon Oh

- 2022.2~ Korean Minjok Leadership Academy