JESK J Ergon Soc Korea 2023; 42(5): 417-433
https://dx.doi.org/10.5143/JESK.2023.42.5.417
https://jesk.or.kr eISSN:2093-8462

Real-time Multimodal Emotion Recognition Based on Multithreaded Weighted Average Fusion

Miracle Udurume[1], Erick C. Valverde[1], Angela Caliwag[1], Sangho Kim[2], Wansu Lim[1]

[1]Kumoh National Institute of Technology, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Gumi-si, 391777
[2]Kumoh National Institute of Technology, School of Industrial Engineering, Gumi-si, 39177

Corresponding Author
Wansu Lim
Kumoh National Institute of Technology, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Gumi-si, 39177
Email : wansu.lim@kumoh.ac.kr

Received : January 27, 2023
Revised : September 18, 2023
Accepted : October 11, 2023

Copyright@2023 by Ergonomics Society of Korea. All right reserved.

Objective: The previous study explored the use of multimodality for accurate emotion predictions. However, limited research has addressed real-time implementation due to the challenges of simultaneous emotion recognition. To tackle this issue, we propose a real-time multimodal emotion recognition system based on multithreaded weighted average fusion.

Background: Emotion recognition stands as a crucial component in human-machine interaction. Challenges arise in emotion recognition due to the diverse expressions of emotions across various forms such as visual cues, auditory signals, text, and physiological responses. Recent advances in the field highlight that combining multimodal inputs, such as voice, speech, and EEG signals, yields superior results compared to unimodal approaches.

Method: We have constructed a multithreaded system to facilitate the simultaneous utilization of diverse modalities, ensuring continuous synchronization. Building upon previous work, we have enhanced our approach by incorporating weighted average fusion alongside the multithreaded system. This enhancement allows us to predict emotions based on the highest probability score.

Results: Our implementation demonstrated the ability of the proposed model to recognize and predict user emotions in real-time, resulting in improved accuracy in emotion recognition.

Conclusion: This technology has the potential to enrich user experiences and applications by enabling real-time understanding and response to human emotions.

Application: The proposed real-time multimodal emotion recognition system holds promising applications in various domains, including human-computer interaction, healthcare, and entertainment.

Keywords: Emotion recognition, Multimodality, Multithreading, Real-time implementation, Weighted average fusion

1. Introduction

Emotion Recognition (ER) has become an essential part of the research field due to the rapid development of intelligent technology and the influences of emotions strongly perceived in our daily activities. Putting together diverse sorts of information that express human emotions is one of the most crucial jobs in the recognition of

emotions. Human emotion is inherently multimodal in its expression. Aside from using words, gestures, and facial expressions, there are many other ways to express emotions (Nguyen et al., 2022). Emotion is part of our everyday life and influences human decision-making, behavior, and communication. Emotion recognition has found its application in various fields such as human-computer interaction (e.g. social robots, monitoring systems for car drivers' conditions), affective computing (which deals with the development of systems that can recognize, process, and simulate human effects), and social signal processing (deals with the analysis and non-verbal information extracted during social interactions) (Tzirakis et al., 2017). Such applications would frequently require a real-time interpretation due to the dynamic and erratic character of emotions. Unfortunately, the majority of the software programs available today for recognizing emotions call on offline post-practice analysis of recorded data, which do not yield results in real time. Rare real-time techniques frequently rely just on one modality (speech, facial expression, skin resistance, posture, etc.), which significantly reduces their accuracy in identifying emotions (Vogt, 2011). Recent studies have demonstrated that the incorporation of multimodal discriminant information (such as facial and audio features) in emotion recognition systems enhances their robustness. This is due to the increased interest in developing real-world scenarios datasets as well as the increased computer processing capabilities (Ristea et al., 2019; Bhattacharya et al., 2021; Eskimez et al., 2021; Radoi et al., 2021). While several multimodal fusion approaches for emotion recognition exist, most of the current approaches consider emotion recognition in an offline manner and are not suitable for real-time implementation. Furthermore, a complicated problem is the difficulty in combining multimodalities synchronously.

To resolve the mentioned challenges, real-time multimodal emotion recognition is proposed in this study which takes three modalities of face, speech, and EEG signals. Though each signal is acquired and processed individually, all the information from all three signals is used to recognize the user's emotion. To this end, we propose a weighted average fusion algorithm for accurate predictions between the three modalities. The contributions of this study are as follows:

- We proposed a real-time multimodal emotion recognition with three different modalities. A multithreaded system was incorporated to enable synchronization of all modalities. Multithreading can be used to enable multiple threads execute at the same time.

- We proposed a weighted average fusion technique. By assigning weights for each modality, we can select the best prediction result obtained from the training process.

2. Related Works

To further increase the accuracy of emotion recognition, a multi-modal fusion approach is now being proposed. A Convolutional Neural Network that uses Autoencoder Units for emotion detection in human faces is being tested in (Prieto and Oplatkova, 2018) using a multimodal fusion model that may combine various physiological inputs to produce results for emotion recognition. A tensor fusion network that fuses modalities dynamically in a hierarchical manner (Liang et al., 2018), a capsule GCN taking into account information redundancy and complementarity (Liu et al., 2021), and late or early fusion networks (Liang et al., 2018), Cimtay et al. (2020) and Lu et al. (2015), which emphasize a relative place of network, have all been proposed and have outperformed single modality emotion recognition systems in terms of performance. Guo et al. (2019) created a hybrid classifier to categorize emotional states using compressed sensing representation by fusing fuzzy cognitive map and support vector machine (SVM). Angadi and Reddy (2019) proposed a technique for recognizing facial expressions while speaking. The technique performs multi-modal fusion of EEG signals and facial emotions using a hybrid deep network architecture. According to experimental findings, the multi-modal fusion model recognizes emotions more frequently than any individual modal model. Tsai et al. (2019) developed a cross-modal transformer. They proposed a multimodal transformer that offers latent cross-modal adaptability and unifies multimodal data by paying close attention to low-level properties in other modalities. For tasks involving emotion recognition, a self-supervised training strategy that makes use of a pre-training procedure using a sizable unlabeled

dataset is employed (Rahman et al., 2020; Zadeh et al., 2018; Khare et al., 2021). In all of the aforementioned investigations, multimodal fusion systems were found to be superior to unimodal input sources for emotion recognition. However, none of these provided a method for multimodal emotion identification that worked in real-time. Our multithreaded weighted average fusion presented in the next section, offers a model for a real-time, continuous, reliable, and accurate multimodal emotion recognition system that incorporates three modalities (face, voice, and EEG).

3. Proposed Method

Our model focuses on multimodal emotion recognition implementation using audio, face, and EEG. The overview of the system model is shown in Figure 1. As shown in the Figure, the signals from the users are extracted using the input devices. These signals are then sent to the local PC for feature extraction and model training. To enable the processing of multiple signals simultaneously, multithreading is used. Multithreading involves the use of shared memory and task queue. Then, the main process of emotion recognition per single modality is performed. Using these results, a weighted average is applied to determine the final emotion. The details are discussed in the following subsections.

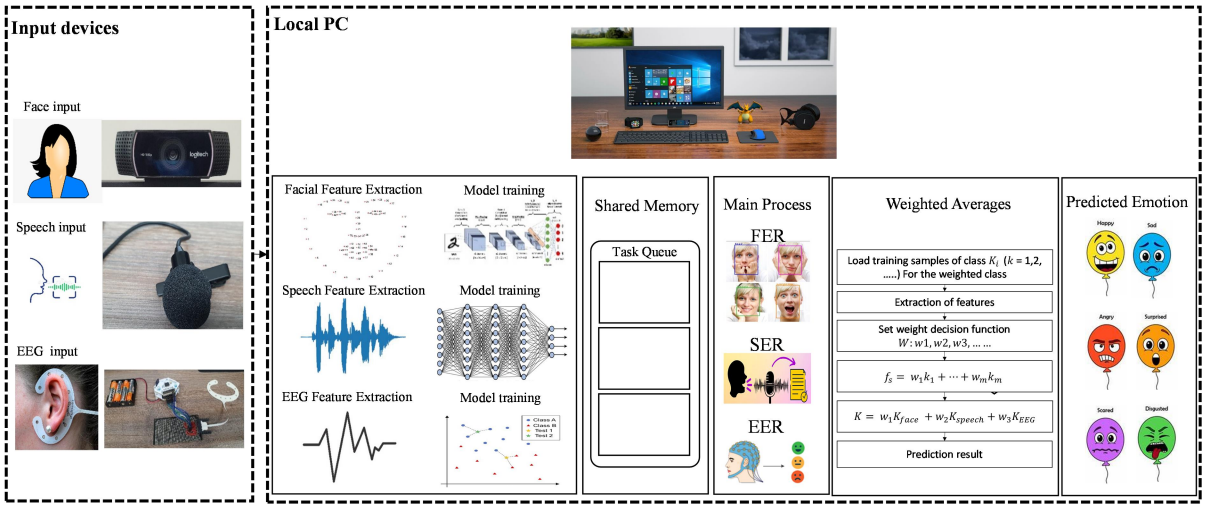


Figure 1. System overview

3.1 Input device

Each emotion recognition modality was performed on a local PC and set up using the Pycharm environment. The input devices used included the hardware device used for real-time implementation which includes specifically an ACPI X64 PC, webcam (c922 pro stream), A USB microphone, and the daisy module OpenBCI Cyton plus cEEGrid device. These devices offer tools for face, voice, and EEG emotion recognition.

3.2 Datasets

In this paper, emotion recognition consists of two phases: individual emotion recognition, and the weighted average fusion-based emotion recognition. Each model used in the first phase of emotion recognition is developed considering different datasets: FER2013 (Goodfellow et al., 2013), RAVDESS (Livingstone and Russo, 2018), TESS datasets (Pichora-Fuller and Dupuis, 2020), and

DEAP (Koelstra et al., 2012). Specifically, the models used in this study are trained using those datasets. That is, the facial emotion recognition model is trained using FER2013 dataset, the speech emotion recognition model is trained using the RAVDESS and TESS datasets, and the EEG emotion recognition model is trained using the DEAP dataset. Once the emotion recognition models are obtained, the model for the second phase is then developed. Specifically, a weighted average fusion method is applied to recognize the emotion considering the output from three different individual emotion recognition models from the first phase. This process of model development is performed offline. After obtaining the models for the first and second phases of emotion recognition, it is also tested in real-time deployment.

3.2.1 FER2013 database

The FER2013 dataset was produced by Pierre Luc Carrier and Aaron Courville. The dataset was developed by searching for photographs of faces that correspond to a list of 184 emotion-related phrases, such as "Anger", "Blissful", etc., using the Google image search API. The phrases associated with gender, age, or ethnicity were then added to these keywords to create approximately 600 strings that are utilized as facial picture search queries. A total of 28709 training photos, 3589 public test images, and 3589 private test images make up the FER2013 dataset. The numbers 0 through 6 in this dataset correspond to seven phrases. Words like "angry", "disgusted", "frightened", "sad", "glad", "surprised", and "neutral".

3.2.2 RAVDESS dataset

A publicly available dataset for speech emotion recognition, the Ryerson Audio-Visual database of Emotional Speech and Song (RAVDESS) dataset include 24 professional actors 12 men and 12 women pronouncing lexically compatible sentences in a North American neutral dialect. There are 1440 files in the dataset, each representing one of eight emotions: "calm", "sad", "disgust", "angry", "fearful", "surprise", "glad", and "neutral". Each emotion has a total of 192 samples available, with the exception of neutral, which has 96 samples. Each expression in the RAVDESS dataset, which includes audio and video files, is produced at two different intensities, normal and strong, with a third neutral expression.

3.2.3 TESS dataset

The Toronto Emotional Speech Set (TESS), which was developed by Kate Dupuis and M. Kathleen Pichora-Fuller, is an open-access dataset that is available for research purposes. Two actresses (26 and 24 years old) from the Toronto region are represented by 2800 audio files in the datasets. 200 target words are vocalized in a set in the carrier phrase. Seven different emotional states are represented in the datasets, including "disgust", "anger", "happiness", "sadness", "neutral", "fear", and "pleasant surprise", The dataset consists of 14 folders, each containing 200 audio files. We integrated the RAVDESS and TESS datasets for this study. To add more variety to the final dataset, the two datasets have been combined.

3.2.4 DEAP dataset

The Database for Emotion Analysis using Physiological Signals (DEAP) comprises EEG recordings from 32 subjects acquired during one-minute viewing sessions of each of the 40 music videos. Each participant was asked to rate each film on a scale of one to nine for its level of arousal, valence, like/dislike, dominance, and familiarity. The EEG has 32 channels. From 16 male and 16 female subjects, the signals were gathered.

3.3 Feature extraction

To improve the accuracy of the model in this work, essential features for each modality were chosen in order to get the best

possible emotion detection prediction.

3.3.1 Face features

This is the process of extracting facial features from the face for face emotion classification. First, we perform face detection which is a process of detecting a person's face in real-time (Figure 2). Face detection is performed using Dlib an open-source library that is a landmark facial detector with trained models (Castellano et al., 2008). Dlib is used to estimate the location of coordinates (x, y) that map the facial points on a person's face to enable the incorporation of features. Once the face is detected, feature points are extracted from the user, these points include face tracking and landmark detection algorithms which are used to track the face of the user in real time. Face landmark detection enables the computer to detect and localize regions of the face such as eyes, eyebrows, nose, and mouth extracted from human faces. These landmarks can be associated with the expressions of certain emotions such as "happy", "sad", "angry", "surprise", "calm", "neutral", "fearful", etc. Facial landmarks can be used as a base task to perform other computer vision tasks such as head pose estimation, detecting facial gestures, facial emotion recognition, etc. facial landmark recognition can help in the identification of human emotions (Figure 3).

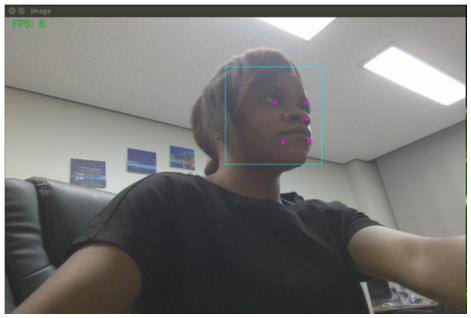


Figure 2. Example of Face detection performed by user in real-time

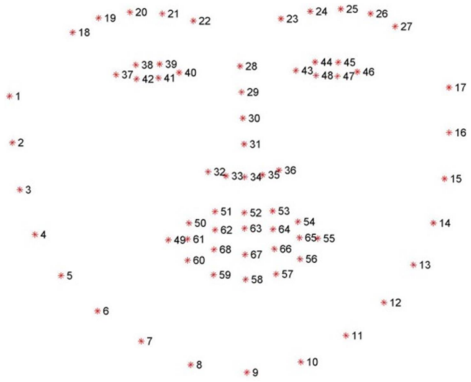


Figure 3. 68 facial landmarks extracted from a frontal face view

3.3.2 Speech features

This process involves the extraction of features from voice intonations. Similar to face feature extraction, this work detects the voice of the user before features are extracted. Speech features are extracted using the Librosa library package (McFee et al., 2015) which is a Python library used for audio and feature extraction. Features such as Mel- frequency cepstrum coefficients (MFCC), and zero crossing (ZCR). MFCC is one of the effective methods used for feature extraction in speech emotion recognition. It is a

coefficient that represents audio based on perception with their frequency bands logarithmically positioned and mimics the human vocal response. MFCC represents the real cepstral of a windowed short-time signal. The MFCC is obtained using the Fast Fourier transform (FFT) of a windowed fragment of a signal while the spectrum's power can be derived with the use of the Mel scale (see Figure 4). For the extraction, the logs of the Mel frequencies' power are used as well as the discrete cosine transform which is stored for further analysis (Nasim et al., 2021). Another means by which speech features can be extracted is by the use of Mel spectrograms. A spectrogram is a signal analysis technique that shows the evolution of the frequency spectrum at the time of a signal. Spectrograms are obtained when the FFTs are computer on overlapping windowed segments of the audio signals. When we take into account the whole frequency spectrum for each window by decomposing the signal magnitude into its components corresponding to the Mel scale's frequencies, it produces spectrograms. The features were extracted with frame_length = 2048, and hop length = 512 (similar to Chunk which is a batch of sequential samples to process at once). The main focus of this work was on the MFCC which is the most popularly extracted feature for speech emotion recognition. Every sample (sequences of 0.2 secs) is analyzed and translated to 4 sequential features (2048/512 = 4). MFCC is grouped into different stages: re-emphasis, windowing, spectral analysis, filter bank processing, log energy computation, and Mel frequency cepstral computation. It must be noted that the communication or speech is performed using the English language.

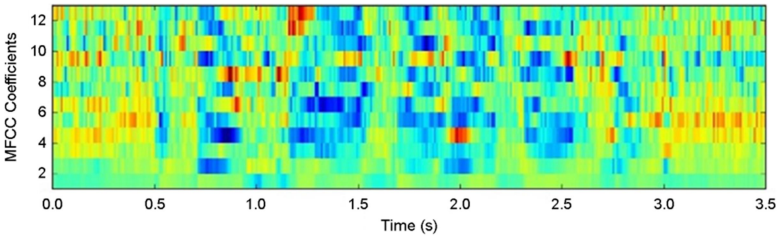


Figure 4. MFCC of a sampled audio file

3.3.3 EEG features

The v3 Daisy module Open-BCI Cyton Plus cEEGrid device was used to measure the EEG data and extract the EEG features for this study. The Open-BCI data format allows for the reading out of raw data for post-processing. We created a signal processing tool for unprocessed Open-BCI data using the Matlab programming language. Signal values with a 24-bit resolution make up EEG data. For signal acquisition, a total of 8 channels are used. Additionally, the X, Y, and Z axes of the accelerometer data are stored as 16-bit signed integers. The Open-default BCI's data sampling rate is 256Hz, however, it can be adjusted using the software the project has made available. EEG data is often studied in the frequency domain, whereas the Open-raw BCI's output is in the time domain. There are eight primary brain waves that make up the component frequencies of an EEG: Theta (4~7Hz), Alpha low (8~9Hz), Alpha high (10~12Hz), Beta low (13~17Hz), Beta high (18~30Hz), Gamma low (31~40Hz), and Delta (1~3Hz) (41~50Hz). These frequencies stand for particular brain states, such as being very attentive, falling asleep deeply, meditating, feeling anxious, and so forth. In order to conduct an efficient analysis, the raw data from the Open-BCI board is converted from the time domain to the frequency domain. The general state of emotions in EEG is described by two different types of models: the discrete emotion model, which covers fundamental emotions like "sadness", "anger", "fear", "surprise", "disgust", "pleasure", etc. The other is a multi-dimensional emotional model of valence and arousal, where valence stands for an individual's level of delight and ranges from negative to positive, and arousal stands for an emotional state's level of arousal, which ranges from calm to excited.

3.4 Multithreading

This section discusses multithreading which allows the execution of several threads concurrently by the provision of multiple hardware contexts and hardware scheduler mechanisms to support fast context switching and thread management within the processor. We introduce the multithreading algorithm used for parallel implementation of each modality synchronously. We define classifier K_i^n for category i data and assume that there are M expressions per category in the training sample set

$$K_i(F_i) = \frac{1}{N}\sum_{n=1}^{N} k_i^n(F_i^n) \quad \#(1)$$

where N defines the total number of training samples; k represents the defined model classifier assigned to modality i; and F represents the features extracted from individual modality. At the same time, we test every model $k_i^n(F_k)$ in conjunction with the preceding $n-1$ boosting rounds. In other words, we tested $K_i^{n-1}(F_i) + k_i^n(F_i)\ for\ k_i^n(F_i)$ on all training samples:

$$J\left(K_i^{n-1}(F_i) + k_i^{\mathrm{n}}(F_i)\right) \quad \#(2)$$

The algorithm in Table 1 summarizes the multithreaded technique used in this paper. We implement a weighted fusion on the multithreading algorithm to enable the appropriate selection of models with the best performance and the capability to remove redundant weak classifiers. In so doing, it is not only able to reduce the number of weak classifiers in each stage but also improve the generalization capability of the strong classifiers. We structure our multithreading model to fit three modalities: face, speech, and EEG.

Table 1. Proposed multithreading algorithm

	Input : Extracted features from modalities F, model classifiers k.
	Output: Prediction results of modalities on training samples N.
1	Initialize N number of samples;
2	k_i^n
3	**while** $F_k > F$ do
4	$n = n - 1$
5	Train classifiers by samples k
6	Evaluate the models on the whole training
7	Update training samples;
8	Determine threads
9	**for** $i \leftarrow 1$ **to** n **do**
10	$K_i(F_i) = \frac{1}{N}\sum_{n=1}^{N} k_i^n(F_i^n)$
11	Update the features for each modality
12	**end**
13	$J\left(K_i^{n-1}(F_i) + k_i^n(F_i)\right)$;
14	Update samples;
15	Evaluate accuracy;
16	**End**

3.5 Emotion recognition

In this paper, emotion recognition consists of two phases: individual emotion recognition, and the weighted average fusion-based emotion recognition. In the first phase, three different models are used to recognize the emotion based on three different input signals (voice, face, and EEG). Hence, each model performs unimodal emotion recognition. In the second phase, the final emotion is recognized based on the emotions obtained in the first phase. That is, the emotion recognized is based on three different modalities. Hence, multimodal emotion recognition is performed. The emotions are derived from the datasets discussed in Section III.B. Since the different emotions are considered by different dataset sources, we limit the emotions to the following: happy, angry, sad, fear, disgust, surprise, neutral, and average. That is, among the data samples provided by different datasets, only those with labels of the emotions mentioned are considered.

A weighted sum of probabilities is used to conduct decision-level fusion in emotion recognition. For the result of emotion recognition to be fused at the decision level across all data modalities, each modality must be segmented into equal emotion units before fusion so that the results of each modality are aligned with each other. Emotion units are difficult to identify given the differences in the rate of emotion recognition across the different modalities. For this work, a weighted average fusion process is implemented to illustrate a method of determining multimodal emotion recognition for face, voice, and EEG as shown in Figure 5. We present a weighted average fusion for each respective emotion. The formula is given below with w_1 being the weight given to facial expression emotion probabilities K_{face}, w_2 being the weight given to speech emotion probabilities K_{speech}, and w_3 being the weight given to EEG emotion probabilities K_{EEG}. The fused weighted score is calculated using the equation

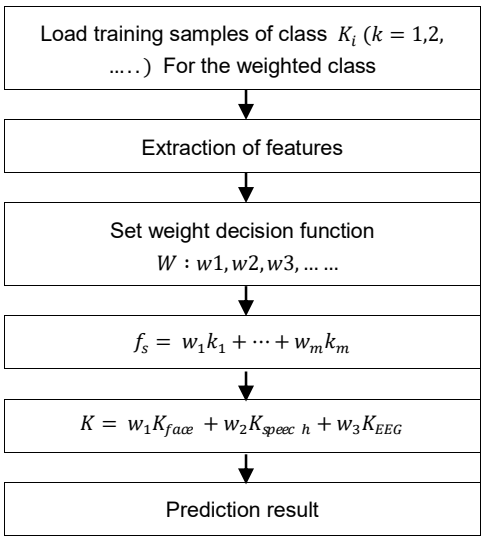


Figure 5. Block diagram of weighted average approach

$$f_s = w_1k_1 + \cdots + w_mk_m \quad \#(3)$$

From equation (3), the weights are varied from the range 0 to 1, such that the limitations are $w_1k_1 + \cdots + w_m k_m$ is satisfied. Hence, equation (4) based on three modalities is given by:

$$K = w_1K_{face} + w_2K_{speech} + w_3K_{EEG} \quad \#(4)$$

Weighted averages are assigned to each unimodality independently to check the emotion recognition performance of each

modality before fusion. Weights are assigned based on features extracted from each modality, and a weight decision function is assigned to check the prediction probability performance for each model. We propose a decision-level weight fusion network to combine each independent classifier. Weighted fusion is based on a weight matrix. Calculating the weight is the key element that gives weight to various features. First, recognition results are obtained from each modality using separate classifiers. Each emotion recognition rate of each classifier is $P_i = \left(p_{i1,\ldots,} p_{i\,m}\right)^T (1 \le i \le n)$ treated as a weight matrix W_i $(1 \le i \le n)$ as follows:

$$W_i = \begin{pmatrix} P_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_{ij} \end{pmatrix} \quad \#(5)$$

In particular, from each classifier, we obtained a weight matrix. Were W_i is a weight matrix assigned for each classifier from the classifiers, P_i is the class for each modality classifier, and n is the number of training samples. Secondly, let $C_i = \left(c_{i1,\ldots,}c_{im}\right)^T (1 \le i \le n)$ as the classifier probabilities resulting from each modality signal are n-dimensional vectors, where $|C_i| = 1$ and $C_{ij} \in \{0,1\}(1 \le i \le n, 1 \le j \le m)$ and the classifier C is obtained according to the linear data fusion principle as follows:

$$C = \sum_{i=1}^{n} C_i \; W_i = \sum_{i=1}^{n} \begin{bmatrix} C_{i1} \\ \vdots \\ C_{im} \end{bmatrix} \begin{bmatrix} P_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_{ij} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} C_{i1} \; p_{i1} \\ \vdots \\ \sum_{i=1}^{n} C_{im} \; p_{im} \end{bmatrix} \quad \#(6)$$

The multimodal emotion recognition result is obtained using a max-min win strategy as follows:

$$\max_{j=1} m \; \left\{\sum_{i=1}^{n} C_{ij} \; p_{ij} \right\} = \sum_{i=1}^{n} C_{ik} \; p_{ik} \quad \#(7)$$

And the most likely category label is k. A property of the weighting fusion approach is that the system that obtained the highest emotion recognition is multiplied by the smallest weights and vice versa. A simple method employed to combine the decisions from the different classifiers is the use of a voting scheme. If each classifier has one vote, it votes for the class with the lowest representation error and all classes receive no vote. A voting classifier is a kind of cooperative learning which engages multiple individual classifiers and combines their predictions to attain better performance than a single classifier. This work also employed the use of a voting classifier. Voting can be grouped into two types: hard and soft voting. Hard voting involves the use of multiple individual models to make predictions. First, each model makes its prediction which is counted as one vote in a running tally after which an outcome based on the majority of votes is recorded for each model. Hard voting gives us the chance to foresee the class name in place of the last class mark which has been anticipated often through models of characterization. Soft voting relies on probabilistic outcome values generated by classification algorithms. The soft voting averages each component model's predicted probability of "1" class membership for record and then assigns a classification outcome prediction based on that average.

4. Results and Discussion

In this section, we describe the hardware implementation and requirements for the proposed real-time multimodal emotion recognition based on multithreaded weighted average fusion.

4.1 Environmental setup and test scenario

Stage 1. Preparing before Initiating Interaction:

- Software Setup: We utilized the BrainProduct software for EEG signal acquisition and synchronization. Timestamp recording was implemented to ensure precise data alignment.

- Electrode Placement: To capture EEG signals related to emotional responses, we carefully prepared the participant by applying conductive gel and attaching electrodes to specific scalp locations. Impedance levels for each electrode point were meticulously checked to maintain signal quality.
- Camera Calibration: We adjusted the camera angle to focus on the participant's frontal face during the interaction to capture facial expressions effectively.

Stage 2. Initiating the Interaction:

- Participant Queries VUI: The interaction began with the participant asking a question to the VUI system. During this phase, the participant's voice was recorded.
- Listening to VUI Response: As the participant received responses from the VUI, we recorded the entire interaction. Importantly, for emotion extraction, we had prepared a database of audio responses encompassing various emotional tones.
- Post-Interaction Steady State: To capture the participant's emotional state after the VUI response, we introduced a 3-second steady state.
- Data Collection: During steps 2 and 3, we simultaneously collected EEG signals and frontal face images.

4.2 Participant information and list of questions

Table 2. Summary of participant information

Subject	Age	Gender	Innovativeness	Personality	Subject	Age	Gender	Innovativeness	Personality
1	22	Male	Majority	Extraversion	6	27	Female	Majority	Extraversion
2	25	Male	Majority	Extraversion	7	24	Female	Majority	Extraversion
3	26	Male	Majority	Introversion	8	20	Female	Majority	Introversion
4	25	Male	Majority	Introversion	9	22	Female	Early adopters	Extraversion
5	22	Male	Early adopters	Extraversion	10	23	Female	Early adopters	Introversion

Table 3. List of Questions

	Question		Question
1	Is the **Dead Sea** a lake or a sea?	11	Recommend me a good food for **diabetes**.
2	What is the angle of the leaning **Tower of Pisa**?	12	What is the **distance** from the earth to the sun?
3	Which animal has the longest **lifespan**?	13	How many **calories** are in tomatoes?
4	Tell me about the principal of spin **free kick**.	14	Who was the first to coin the term **quantum mechanics**?
5	What is the **PCR** test?	15	When is **Joseon** National day?
6	Tell me about the benefits of the **red ginseng**.	16	How much does the **Netflix** premium plan cost?
7	How many **Ikea** stores are in Korea?	17	Where is the **Grand Canyon** located?
8	What is the smartest **breed** of dog?	18	What is the **flower language** of red roses?
9	What is the number of **public holidays** in Korea this year?	19	What is the proper **indoor** temperature in winter?
10	What is the official language of the **United Nations**?	20	What ingredients are needed to make a **mulled wine**?

Table 2 provides a summary of the demographic information for the participants in the experiment. A total of 10 subjects were involved in the experiment, with their ages ranging from 22 to 27. The participants were categorized by gender, Innovativeness, and Personality. This table offers a comprehensive overview of the composition of the participant group. Table 3 presents the list of questions which cover a range of topics and were designed to evaluate the recognition of emotions. The questions are presented in a clear and organized format.

4.3 Real-time implementation

To achieve a real-time implementation for multimodal emotion recognition, different platforms are set up such as webcam-based, audio-based, and cEEGrid-based. The webcam-based system, audio-based system, and the cEEGrid-based system is set up to capture the facial emotions, voice recording, and brain signals of the user. Emotion recognition is performed in real-time for each modality separately before fusion is taken place. The overall implementation setup used in this work for real-time emotion recognition: unimodal and multimodal is presented in Figure 6. The devices comprise of webcam, microphone, speaker, and EEG sensor grid. During the experiment, the user is sited in front of a camera and a microphone, while wearing an EEG sensor on his ear. Once the model is executed, the user demonstrates different types of emotions using facial expressions, voice tone, and thoughts. Overall, the desired settings for real-time emotion recognition are demonstrated successfully.

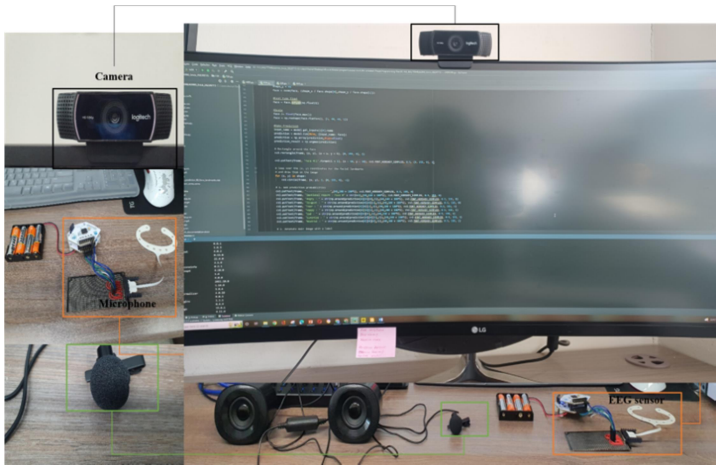


Figure 6. Overall implementation setup for real-time emotion recognition: unimodal and multimodal

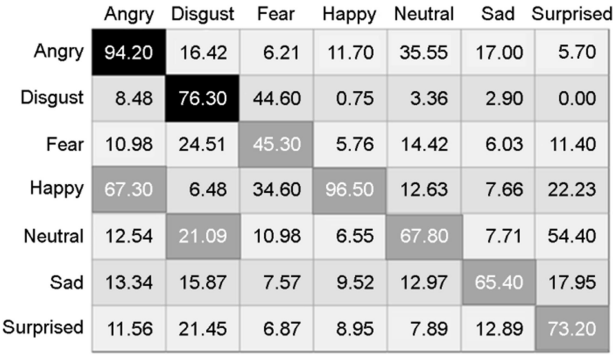
4.4 Performance evaluation

In this section, results obtained from emotion recognition: unimodal and multimodal are presented. We perform each emotion recognition separately first before combining all three. The test is conducted by recording the test for each modality in real time. We report the result obtained from the study. This work displays seven kinds of emotions for each modality namely: "happy", "angry", "sad", "fear", "disgust", "surprise", and "neutral". Table 4 reports the summary of the emotion recognition results on both unimodal and multimodal emotion recognition in real time. The multi-threaded system is built and executed on separate threads for each modality before the main thread is used for all three. The weights are assigned to select predictions with the highest accuracy results. We calculate the average for each emotion of each modality and average the result for the final prediction score. Table 4 demonstrates that the combination of the unimodality yields the optimum result. The numbers in the table show how well each emotion was predicted. When compared to multimodal emotion identification, we noticed that each modality's

accuracy performance differs just marginally. The combined RAVDESS and TESS datasets may explain why overall accuracy for unimodality is greater for SER in particular. From the results of the offline classification matrix's confusion matrix, the SER had a better performance as compared to FER and EEG. Figure 7 shows the confusion matrices obtained in an offline setting for face, speech, and EEG emotion recognition. All emotion categories have been identified with a high level of class performance as indicated by the diagonal elements of all matrices.

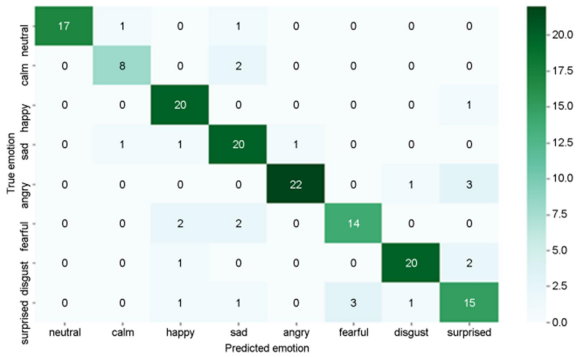
Table 4. Accuracy performance on unimodal and multimodal emotion recognition

Modality	Face	Speech	EEG	F+S+E
Happy	96.50%	86.50%	98.50%	98.60%
Angry	94.20%	78.20%	67.30%	97.80%
Sad	65.40%	77.90%	83.70%	91.60%
Fear	45.30%	54.60%	72.30%	78.90%
Disgust	76.30%	45.30%	57.40%	88.60%
Surprise	73.20%	67.50%	71.60%	75.30%
Neutral	67.80%	46.30%	84.70%	89.70%



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprised
Angry	94.20	16.42	6.21	11.70	35.55	17.00	5.70
Disgust	8.48	76.30	44.60	0.75	3.36	2.90	0.00
Fear	10.98	24.51	45.30	5.76	14.42	6.03	11.40
Happy	67.30	6.48	34.60	96.50	12.63	7.66	22.23
Neutral	12.54	21.09	10.98	6.55	67.80	7.71	54.40
Sad	13.34	15.87	7.57	9.52	12.97	65.40	17.95
Surprised	11.56	21.45	6.87	8.95	7.89	12.89	73.20

(a)



(b)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprised
Angry	67.73	16.42	6.21	5.44	3.79	2.18	5.70
Disgust	8.48	57.40	5.90	0.75	3.36	2.90	0.00
Fear	10.98	26.71	72.30	5.76	14.42	6.03	11.40
Happy	67.30	9.48	10.79	98.85	12.63	7.66	2.14
Neutral	12.54	21.09	10.98	6.55	84.70	7.71	4.17
Sad	13.34	15.87	7.57	9.52	12.97	83.70	17.95
Surprised	11.56	21.45	6.87	8.95	7.89	12.89	71.60

(c)

Figure 7. Confusion matrices for FER, SER and EEG in an offline setting: (a) FER, (b) SER, (c) EEG

4.5 Comparison with other methods

For the recognition of multimodal emotions, various techniques are employed. Every MER model that is now in use takes a distinct approach in terms of characteristics retrieved, modalities employed, and performance evaluation techniques. The user has implemented the proposed multimodal emotion identification system in real-time to demonstrate its effectiveness. The final recognition performance of our method is compared with several baselines and the results are shown in Table 5. The recognition rate of the system suggested in Cimtay et al. (2020) achieved the highest result when compared to that of other methods, as shown in Table 6. That suggested in Liu et al. (2022) came in second. The third strategy with the highest accuracy is the one this study suggests. It is significant to note that each of these investigations used a variety of modalities, which may have an impact on the recognition rate. Another factor that may also affect the recognition performance rate is the dataset used. The table shows the distinction between the dataset used for all studies. For instance, Liang et al. (2018) advocated real-time multimodal emotion identification in conversational videos using reinforcement learning and domain expertise. The baseline's inability to detect emotions in real time caused the level of the entire discussion to fluctuate constantly. Zhang et al. (2022) proposes an emotion recognition system using audio and visual data that is built on a convolutional neural network architecture. For the multimodal emotion identification system, the real-time experimental result is not given, though. For offline multimodal emotion recognition using audio, visual, and text modalities, a multi-channel weight-sharing autoencoder based on cascade multi-head attention is presented (Cimtay et al., 2020). Liu et al. (2022) proposed a multimodal emotion integrating text, audio, and visual modalities with the use of weighted average and F1 score for accuracy performance with the exploitation of different classes but no real-time implementation was achieved. The method utilized in this work has been successful in overcoming the difficulty of multimodal emotion recognition since it proposes multimodal emotion recognition in real-time.

Table 5. Model performance using proposed weighted average

Modality	Face	Speech	EEG	F+S+E
Happy	63.3%	82.8%	88.5%	92.6%
Angry	73.2%	78.2%	77.3%	83.8%
Sad	85.4%	87.9%	83.7%	91.6%
Fear	75.3%	64.6%	72.3%	94.9%
Disgust	76.3%	75.3%	77.4%	88.6%
Surprise	83.2%	93.5%	61.6%	95.3%
Neutral	79.8%	86.3%	74.7%	89.7%
Average	78%	74.5%	69.5%	80.8%

Table 6. Comparison with other methods

Dataset	Recognition rate (%)
IEMOCAP+MELD (Liang et al., 2018)	66.8%
CMU-MOSEI + IEMOCAP (Liu et al., 2022)	85.8%
IEMOCAP+MSP-IMPROV (Cimtay et al., 2020)	86.5%
RAVDESS+CREMA-D (Zhang et al., 2022)	78.70%
RAVDESS+TESS+FER2013+DEAP (proposed)	84.47%

5. Conclusion

This paper proposed a real-time multimodal emotion recognition based on multithreaded weighted average fusion motivated by our previous work emotion recognition implementation with multimodalities of face, voice, and EEG (Udurume et al., 2022). In contrast with previous work focusing on multimodal emotion recognition, this paper proposes a novel multithreaded weighted average fusion for the task of real-time multimodal emotion recognition to improve emotion recognition performance. We presented a weighted average fusion approach to overcome this issue. Effective fusion technique selection is crucial to enhancing the accuracy of the emotion identification system. The accuracy findings demonstrate that the weighted average also performs with excellent accuracy on the multimodal emotion recognition task. We also evaluated the performance of the suggested system against other cutting-edge models. Our experimental findings demonstrate that the system can operate with high accuracy for each modality individually and collectively in both offline and online environments. In light of this, the suggested real-time system is able to recognize emotions in real time on the local PC device. Although there have been previous studies on multimodal emotion recognition, there have been very few studies on multimodality application in real time.

Acknowledgement

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) funded by the MSIT (2020R1A4A101777511).

References

Angadi, S. and Reddy, V.S., Hybrid deep network scheme for emotion recognition in speech, *International Journal of Intelligent Engineering and Systems*, 12(3), 59-67, 2019.

Bhattacharya, P., Gupta, R.K. and Yang, Y., Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos, *IEEE Transactions on Affective Computing*, 14(2), 1547-1557, 2021.

Castellano, G., Kessous, L. and Caridakis, G., Emotion recognition through multiple modalities: Face, body gesture, speech, affect and emotion in human-computer interaction, Peter & R. Beale (Eds.), 92-103, 2008. DOI:10.1007/978-3-540-85099-1_8.

Cimtay, Y., Ekmekcioglu, E. and Caglar-Ozhan, S., Cross-subject multimodal emotion recognition based on hybrid fusion, *IEEE Access*, 8, 168865-168878, 2020.

Eskimez, S.E., Zhang, Y. and Duan, Z., Speech driven talking face generation from a single image and an emotion condition, *IEEE Transactions on Multimedia*, 24, 3480-3490, 2021. doi:10.1109/TMM.2021.3099900

Goodfellow, I.J., Erhan, D. and Carrier, P.L., "Challenges in Representation Learning: a report on three machine learning contests", *Proceedings of the International Conference on Neural Information Processing*, (pp. 117-124), 2013.

Guo, K., A hybrid fuzzy cognitive map/support vector machine approach for EEG-based emotion classification using compressed sensing, *International Journal of Fuzzy Systems*, 21(3), 263-273, 2019.

Khare, A., Parthasarathy, S. and Sundaram, S., "Self-supervised learning with cross-modal transformers for emotion recognition", *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, (pp. 381-388), 2021.

Koelstra, S, Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I.Y., DEAP: A Database for Emotion Analysis Using Physiological Signals, *IEEE Transactions on Affective Computing*, 3(1), 18-31, 2012.

Liang, P.P., Salakhutdinov, R. and Morency, L.P., "Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion", *Proceedings of the 1st Grand Challenge and Workshop on Human Multimodal Language*, (pp. 1-23), 2018.

Liu, J., Chen, S., Wang, L., Liu, Z., Fu, Y. and Guo, L., "Multimodal emotion recognition with capsule graph convolutional based representation fusion", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (ICASSP)*, (pp. 6339-6343), 2021.

Liu, W., Qiu, J.L., Zheng, W.L. and Lu, B.L., Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition, *IEEE Transactions on Cognitive and Developmental Systems*, 14(2), 715-729, 2022. doi:10.1109/TCDS.2021.3071170.

Livingstone, S.R. and Russo, F.A., The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic multimodal set of facial and vocal expressions in North American English, *PLoS ONE*, 13(5), 2018.

Lu, Y.F., Zheng, W.L., Li, B.B. and Lu, B.L., "Combining eye movements and EEG to enhance emotion recognition", *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, (pp. 1170-1176), 2015.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M. and Battenberg, E., "Librosa: Audio and music signal analysis in Python", *In Proceedings of the 14th python in Science Conference*, (pp. 18-25), 2015.

Nasim, A.S., Chowdory, R.H., Dey, A. and Das, A., "Recognizing Speech Emotion Based on Acoustic Features Using Machine Learning," *Proceedings of the International Conference on Advanced Computer Science and Information Systems*, (pp. 1-7), 2021.

Nguyen, D., Nguyen, D.T., Zeng, R., Nguyen, T.T., Tran, S.N., Nguyen, T., Sridharan, S. and Fookes, C., Deep Auto-Encoders With Sequential Learning for Multimodal Dimensional Emotion Recognition, *IEEE Transactions on Multimedia*, 24, 1313-1324, 2022. doi:10.1109/TMM.2021.3063612.

Pichora-Fuller, K.M. and Dupuis, K., Toronto emotional speech set (TESS), *Scholars Portal Dataverse*, 2020.

Prieto, L.A.B. and Oplatkova, Z.K., Emotion recognition using autoencoders and convolutional neural networks, *Mendel*, 24(1), 113-120, 2018.

Radoi, A., Birhala, A., Ristea, N.C. and Dutu, L.C., An End-To-End Emotion Recognition Framework Based on Temporal Aggregation of Multimodal Information, *IEEE Access*, 9, 135559-135570, 2021. doi:10.1109/ACCESS.2021.3116530.

Rahman, W., Hasan, M.K., Lee, S., Zadeh, A.B., Mao, C. and Morency, L.P., "Integrating multimodal information in large pretrained transformers", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 2359-2369), 2020.

Ristea, N.C., Dutu, L.C. and Radoi, A., "Emotion recognition system from speech and visual information based on convolutional neural networks", *Proceedings of the International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*,

(pp. 1-6), 2019.

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P. and Salakhutdinov, R., "Multimodal transformer for unaligned multimodal language sequences", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (pp. 6558-6569), 2019.

Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B. and Zafeiriou, S., End-to-end multimodal emotion recognition using deep neural networks, *IEEE Journal of Selected Topics in Signal Processing*, 11(8), 1301-1309, 2017. doi:10.1109/JSTSP.2017.2764438

Udurume, M., Caliwg, A., Lim, W. and Kim, G., "Emotion recognition implementation with multimodalities of face, voice, and EEG", *Journal of Information and Communication Convergence Engineering*, 20, 174, 2022. doi:10.56977/jicce.2022.20.3.174.

Vogt, T., "Real-time automatic emotion recognition from speech: The recognition of emotions from speech in view of real-time applications", Suedwestdeutscher Verlag fuer Hochschulschriften, 2011.

Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E. and Morency, L.P., "Memory fusion network for multi-view sequential learning", *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 1-8), 2018.

Zhang, K., Li, Y., Wang, J., Cambria, E. and Li, X., Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge, *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 1034-1047, 2022. doi:10.1109/TCSVT.2021.3072412.

Author listings

Miracle Udurume: udurumemiracle@kumoh.ac.kr
Highest degree: M.S. Student, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology
Position title: M.S. Student, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology
Areas of interest: HCI, Machine learning, VUI, Embedded system

Erick C. Valverde: ecvalverde@kumoh.ac.kr
Highest degree: M.S. Student, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology
Position title: M.S. Student, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology
Areas of interest: HCI, Machine learning, VUI, Embedded system

Angela Caliwag: a_caliwag@kumoh.ac.kr
Highest degree: M.S., Department of IT Convergence, Kumoh National Institute of Technology
Position title: Researcher, Department of Aeronautics, Mechanical and Electronic Convergence, Kumoh National Institute of Technology
Areas of interest: HCI, Machine learning, VUI, Embedded system

Sangho Kim: kimsh@kumoh.ac.kr
Highest degree: PhD, Pohang University of Science and Technology (POSTECH)
Position title: Professor, School of Industrial Engineering, Kumoh National Institute of Technology
Areas of interest: Ergonomics, Interaction Design, Kansei Engineering

Wansu Lim: wansu.lim@kumoh.ac.kr
Highest degree: PhD, Gwangju Institute of Science and Technology (GIST)
Position title: Professor, Department of Aeronautics, Mechanical and Electronic Convergence Engineering, Kumoh National Institute of Technology
Areas of interest: HCI, Machine learning, VUI, Embedded system