TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)
31 Oct - 3 Nov 2023. Chiang Mai, Thailand

Integration of Facial and Speech Expressions for Multimodal Emotional Recognition

Thammaros Ruangdit
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, 10900, Thailand
Email: thammaros.r@ku.th

Tonkla Sungkhin
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, 10900, Thailand
Email: tonkla.s@ku.th

Weerapat Phenglong
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, 10900, Thailand
Email: weerapat.phe@ku.th

Ekachai Phaisangittisagul
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, 10900, Thailand
Email: fengecp@ku.ac.th

***Abstract*—The detection and interpretation of human emotions using a variety of sensory cues, including visual, auditory, body language, and physiological indicators, is commonly referred to as emotional recognition. This technology has a broad range of applications, including medical diagnosis, customer service satisfaction, and intelligent call center routing. While facial expressions represent one of the most conspicuous emotional cues utilized in emotional recognition, the integration of additional expressions may enhance its efficacy. To this end, this study proposes an integration of facial and speech expressions as emotional features for the prediction of emotions in machine learning models. The emotional predictions from facial and speech recognition models are subsequently combined to generate the final emotional prediction. Several facial and speech benchmark datasets are employed to evaluate the proposed methodology. The experimental results demonstrate significant promise and the continued development of emotional recognition as a supplementary system for assisting doctors in clinical and research studies appears feasible.**

***Index Terms*—Emotion recognition, facial expression, face mesh, speech expression, speech features**

I. Introduction

An emotional recognition system (ERS) is a kind of artificial intelligence technology that is designed to detect and interpret human emotions based on a range of visual, auditory, body language, and physiological indicators. The system is often used in fields such as healthcare, marketing, and human-computer interaction to assess mental state and diagnose mental health conditions for improvement of personalize content or services.

ERS technology has the potential to revolutionize many industries by providing insights into consumer behavior and improving communication between humans and machines. For example, in healthcare, ERS can help detect early signs of mental health disorders and monitor patient progress. In marketing, ERS can help companies tailor their products and services to better meet customers' emotional needs.

One area where emotional recognition technology is particularly useful is in the diagnosis and treatment of mood disorders, such as depression and anxiety. Research has shown that individuals with depression and anxiety exhibit distinct patterns of facial expressions, speech patterns, and other nonverbal cues that can be identified using emotional recognition technology [1]. By analyzing these patterns, emotional recognition technology can provide an objective measure of an individual's emotional state, which can be used to diagnose mood disorders and monitor treatment progress.

Various techniques have been used in ERS, including computer vision, natural language processing, and biometric sensors. Computer vision techniques analyze facial expressions, body language, and eye movements to determine emotional states, while natural language processing algorithms analyze speech patterns and tone to infer emotions. Biometric sensors measure physiological signals like heart rate, skin conductance, and brain activity to infer emotional states.

The objective of this research is to construct an emotional recognition framework that can categorize emotional states into four distinct categories, namely happiness, sadness, anger, and neutral. In this study, the integration of facial and speech expressions has been utilized as the primary input modality, obviating the need for specialized physiological sensors. The proposed framework holds the potential to augment conventional subjective measures, thereby elevating the precision and dependability of psychological evaluations. The remainder of this manuscript is arranged as follows. Section II introduces the conventional process of emotional recognition. Section III presents a proposed methodology for emotional recognition. Section IV expounds on the experimental outcomes obtained from varied benchmark datasets and discusses the performance of the proposed approach. Finally, in Section V, conclusions are drawn based on the findings of this research.

II. Related works

The study of emotional recognition has become increasingly important in various fields such as psychology, human-

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

computer interaction, and artificial intelligence. This introduction will provide an overview of the current research on emotional recognition using facial, speech, and other expressions.

Facial expressions are one of the most prominent signs used in emotional recognition. Research has shown that humans are highly proficient at recognizing emotions from facial expressions [2]. Studies have shown that certain facial expressions are associated with specific emotions, such as the raised eyebrows and open mouth of surprise or the furrowed brow and downturned mouth of sadness [3]. Moreover, facial expressions have been found to be consistent across different cultures [4]. With the advancement of computer vision technology, automated facial expression recognition systems have been developed to recognize and classify facial expressions [5]. Machine learning algorithms have been developed to detect and classify these expressions, achieving high accuracy rates in some cases [6]. However, challenges remain in dealing with variations in facial expression due to individual differences, cultural factors, and context. Despite these challenges, recent research has made significant progress in improving the accuracy and robustness of facial expression recognition. For example, deep learning techniques such as convolutional neural networks (CNNs) have been shown to outperform traditional machine learning approaches in some cases [7]. Additionally, some studies have explored the use of facial landmarks or action units (AU) to improve the accuracy of facial expression recognition.

Speech patterns are another important signal used in emotional recognition. Research has shown that emotional states can be conveyed through various aspects of speech such as pitch, intensity, and duration [8]. Research on emotional recognition based on speech expressions has primarily focused on three types of approaches: acoustic-based, linguistic-based, and multimodal approaches [9]. Acoustic-based approaches rely on extracting prosodic and spectral features from speech signals to classify emotions. Linguistic-based approaches rely on analyzing the lexical and semantic content of speech to infer emotional states. Multimodal approaches combine both acoustic and linguistic features to improve the accuracy of emotion recognition. In recent years, automatic speech emotion recognition (ASER) systems have been developed to recognize emotions from speech. These systems have been used in various applications such as affective computing and mental health diagnosis.

Other non-verbal indicators such as body language, gestures, and physiological signals have also been used in emotional recognition. For example, Kreibig *et al.,* [10] has shown that physiological signals such as heart rate variability can be used to detect emotions. Moreover, recent studies have shown that machine learning techniques can be used to classify emotions based on electroencephalogram (EEG) signals [11].

III. Proposed Methodology

The proposed method for emotional recognition is presented in Fig. 1 which can be divided into face and speech emotional recognitions. Independent datasets containing facial and speech expressions, along with corresponding emotional labels, are used to construct the face and speech emotional recognition models. In practice, facial and speech expressions are extracted from video data. In the face emotional recognition algorithm, the initial step is to detect the facial object in the image, for which various algorithms such as Haar cascades, Multi-Task Cascaded Convolutional Neural Network (MTCNN), and face mesh can be employed. In this study, the face mesh method is used for face detection. The face mesh represents a 3D model of the human face consisting of interconnected vertices, edges, and faces, which enables accurate tracking and analysis of facial expressions including movements. Results of different face detection algorithm (Haar cascade, MTCNN, and Face mesh) are shown in Fig. 2. To train the face emotional recognition model, the detected facial objects undergo image alignment to position the faces centrally, as illustrated in Fig.3.

In this study, speech from the video is extracted, and sentence segmentation is performed before extracting useful speech features. Various speech feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) [12], Mel spectrogram [13], and Chroma-based representation [14] can be employed. The MFCC is a robust method for speech analysis, which relies on the human auditory system's perception of sound. It involves taking the logarithmic transformation of the power spectrum of the speech signal, followed by a discrete cosine transform (DCT) to derive a set of coefficients representing the spectral envelope of the speech signal. On the other hand, the Mel spectrogram combines the Mel frequency scale with the spectrogram, simulating the human auditory system's frequency resolution, and thus serving as a useful tool for speech analysis. Furthermore, the Chroma Short-Time Fourier Transform (STFT) is one such method that has been widely used for music analysis and has shown promising results in various applications. The Chroma STFT represents a signal in terms of its chromatic content, which is obtained by mapping the signal's frequency content onto the 12-tone chromatic scale. This is achieved by first applying a short-time Fourier transform (STFT) to the signal to obtain a time-frequency representation, and then summing the energy in each frequency band that corresponds to a particular chroma bin. The resulting chroma representation including MFCC and Mel spectrogram are used as speech features to train the speech emotional recognition model.

Regarding face emotional recognition, the proposed model predicts emotions every second using the average of emotion predictions of each frame from the input video. In contrast, the speech emotion recognition model predicts emotions based on the entire length of each sentence. The final emotion prediction (P_e - Predicted probability of each emotional state) is obtained by integrating the emotion predictions from both facial (P_f) and speech (P_s) expressions as follows:

$$P_e = w \times P_f + (1 - w) \times P_s \tag{1}$$

where w is a parameter to weight the predicted emotion

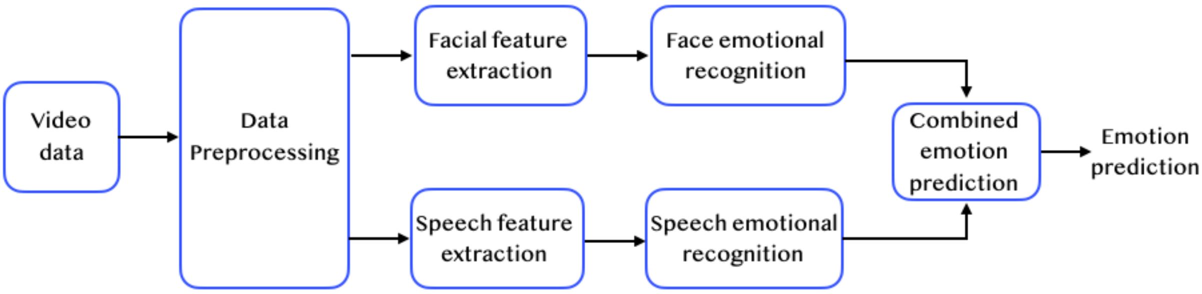

Video data
Data Preprocessing
Facial feature extraction
Face emotional recognition
Speech feature extraction
Speech emotional recognition
Combined emotion prediction
Emotion prediction

Fig. 1: The proposed framework of emotional recognition

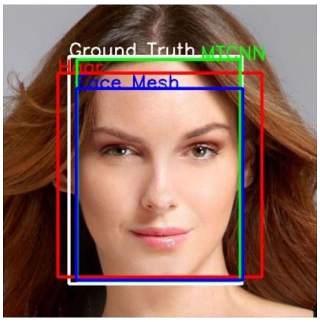


Fig. 2: Illustration of different face detection algorithms.

Fig. 3: Examples of face image before alignment (left) and after alignment (right).

probability between face and speech expressions. In this study, trial and error is used to determine the value of w.

IV. Experiments

The diagrammatic representations of the models for recognizing emotional expressions in faces and speech are illustrated in Fig.4-5, respectively. In the facial recognition model, a pre-trained convolutional neural network, specifically Inception-v3, which has been trained on the ImageNet dataset, is utilized to extract facial features. These features are then processed by a feedforward neural network for recognizing facial emotions. For speech emotion recognition, a feedforward neural network with four layers is constructed. This network predicts emotions based on Mel-frequency cepstral coefficients (MFCC), Mel spectrogram, and Chroma short-time Fourier transform (STFT) as the input features.

The performance of the proposed facial recognition model is evaluated by the AffectNet-HQ dataset [14], which includes high-quality images of human faces that represent eight distinct emotional categories: neutral, happy, sad, angry, surprised, fearful, disgusted, and contemptuous. However, this study decides to focus on four emotions, namely neutral, happy, sad, and angry, with a total of 3,074, 4,327, 3,310, and 4,107 images, respectively. The reason for choosing only four emotions is that they might be the most relevant and actionable for decision-making in the diagnosis of depression, which is the main focus of the future application. Figure 6 displays example facial expressions associated with each of these emotions. Additionally, the speech emotion recognition model was trained using multiple datasets, including the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D), Toronto Emotional Speech Set (TESS), Surrey Audio-Visual Expressed Emotion (SAVEE), the Audio, Speech, and Vision Processing Lab Emotional Sound Database (ASVP-ESD), and Emotional Speech Database (ESD). These datasets collectively contain 17,274 files of neutral, 18,947 files of happy, 17,351 files of sad, and 18,442 files of angry expressions.

Table I demonstrates the classification results of face and speech emotions based on 5-fold cross-validation and the confusion matrix of the emotion recognition system based on the facial and speech expression is provided in Fig. 7. Clearly, facial emotion recognition achieved better performance than speech emotion recognition. In addition, the findings reveal that the proper value of the weight parameter (w) in the experiment is 0.6, indicating that the face recognition model has a more significant impact than the speech recognition model. It should be noted that the value of w can be adjusted according to the datasets. Although the availability of ground truth data for emotional recognition in the video database is limited, we have created hand-made videos showcasing a range of emotions to evaluate the proposed model. Figure 8 depicts a real-time comparison of a video clip between the predicted and actual emotional state, including the error. The proposed model shows promising results and exhibits potential for employment in emotional classification tasks. The occurrence of errors can be attributed to the rapid changes in emotional states and may involve a wide range of expressions from the dataset. For instance, individuals may exhibit varying facial expressions, pitch, and speech intensity even while expressing the same emotional state, leading to difficulties in recognizing emotions by the model.

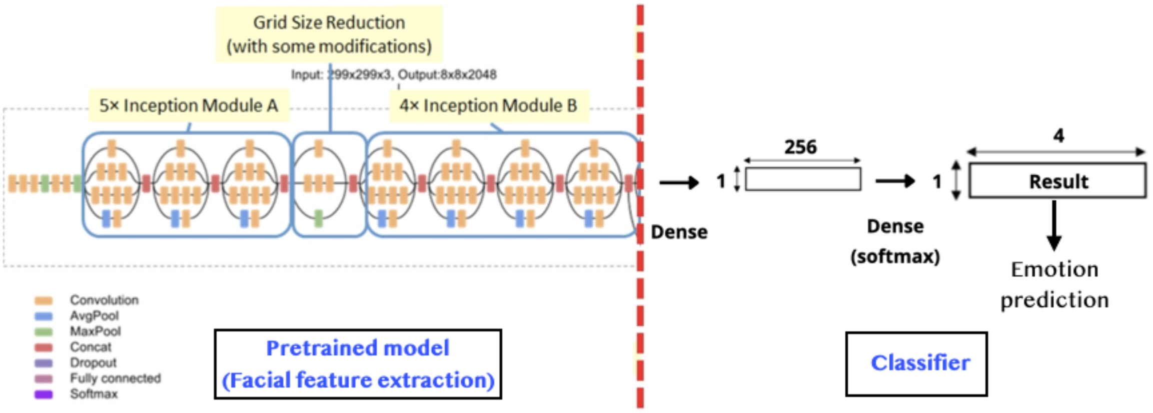


Fig. 4: Structure of face recognition model.

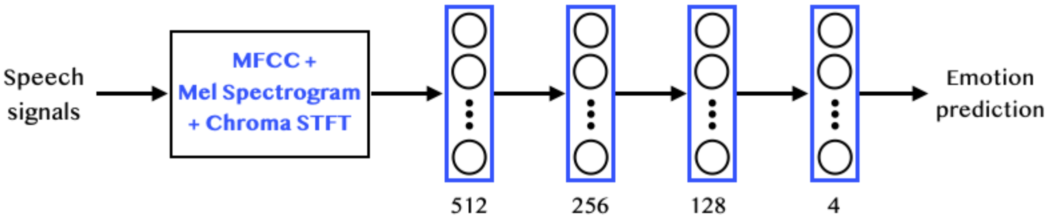
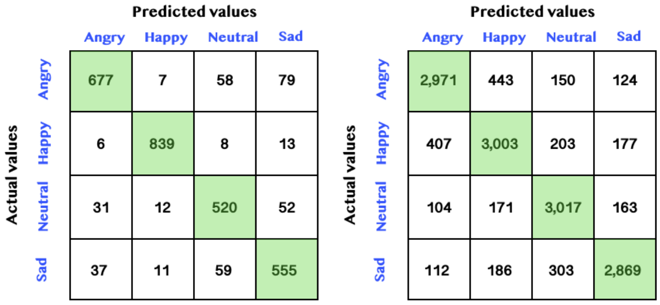


Fig. 5: Structure of speech recognition model.

Fig. 6: Sample images from AffectNet-HQ of four facial emotion expressions.



	Angry	Happy	Neutral	Sad
Angry	677	7	58	79
Happy	6	839	8	13
Neutral	31	12	520	52
Sad	37	11	59	555

	Angry	Happy	Neutral	Sad
Angry	2,971	443	150	124
Happy	407	3,003	203	177
Neutral	104	171	3,017	163
Sad	112	186	303	2,869

Fig. 7: Confusion matrix of the emotion recognition system: Facial emotion recognition (left) and Speech emotion recognition (right)

TABLE I: Emotion classification performance of face and speech expressions

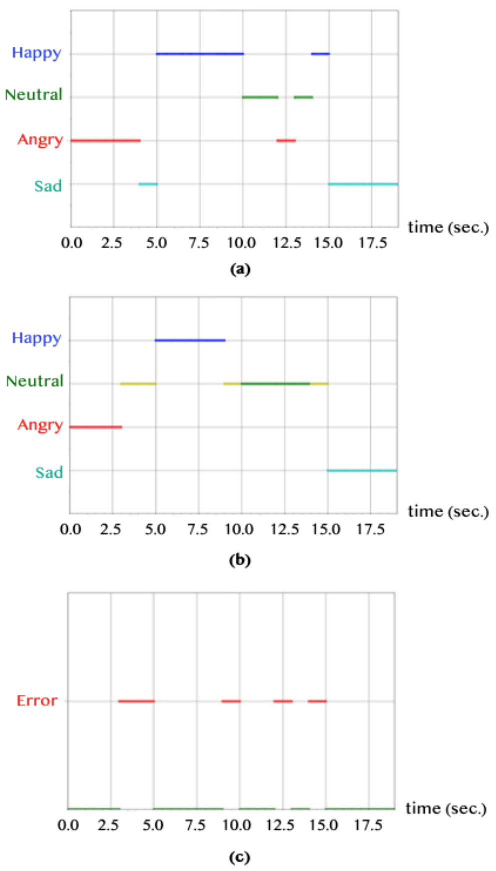
	Face emotional recognition	Speech emotional recognition
Accuracy	85.82	83.10
Precision	85.80	83.09
Recall	85.88	83.10
F1-score	85.84	83.09

V. Conclusion

This research proposes a method for predicting emotions based on facial and speech expressions. The approach involves integrating facial and speech recognition models, each of which is trained on independent datasets. The facial recognition model is implemented using Inception-v3, while the speech recognition model is developed by a feedforward neural network. The proposed method aims to classify four emotions, namely neutral, happy, sad, and angry, on various databases. Our experimental findings indicate that the facial recognition model outperforms the speech recognition model. The proposed emotional recognition method has the potential to serve as an auxiliary system to support clinical decision-making by healthcare professionals. In future studies, the inclusion of additional physiological data such as heart rate, blood pressure, respiration rate, and body temperature could further enhance the accuracy of emotion recognition.

References

[1] D. Kliemann and R. Adolphs, ”The social neuroscience of mentalizing: challenges and recommendations,” Current Opinion in Psychology, 24, 2018, pp. 1–6.



(a)

(b)

(c)

Fig. 8: Real-time comparison results between the ground truth and model prediction: (a) Ground truth (b) Model prediction (c) Error

[2] P. Ekman and W. V. Friesen, ”Constants across cultures in the face and emotion,” Journal of Personality and Social Psychology, 17(2), 1971, pp. 124–129.

[3] P. Ekman, ”An argument for basic emotions,” Cognition and Emotion, 6(3-4), 1992, pp. 169–200.

[4] D. Matsumoto and P. Ekman, ”American-Japanese cultural differences in intensity ratings of facial expressions of emotion,” Motivation and Emotion, 13(2), 1989, pp.143–157.

[5] I. Kotsia, I. Buciu, and I Pitas, ”An analysis of facial expression recognition under partial facial image occlusion,” Image and Vision Computing. 26, 2018, pp. 1052–1067.

[6] A. Dhall, R. Goecke, J. Joshi, and K. Sikka, ”Emotion recognition using facial expressions, speech and multimodal information,” ACM Transactions on Multimedia Computing, Communications, and Applications, 15(1), 2019, pp.1–20.

[7] A. Mollahosseini, D. Chan, and M. Mahoor, ”Going deeper in facial expression recognition using deep neural networks,” 2016 IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–10.

[8] K. R. Scherer, ”Vocal affect expression: A review and a model for future research,” Psychological Bulletin, 99(2), 1986, pp. 143–165.

[9] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, and S. Narayanan, ”IEMOCAP: Interactive emotional dyadic motion capture database,” Journal of Language Resources and Evaluation, 42(4), 2008, pp. 335–359.

[10] S. D. Kreibig, ”Autonomic nervous system activity in emotion: a review,” Biol. Psychol. 84, 2010, pp. 394–421.

[11] A. Alnafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil, ”Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review,” Applied Sciences. 7, 2017, pp. 1239.

[12] S.B. Davis and P. Mermelstein, ”Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing, 28, 1980, pp. 357–366.

[13] P. Mermelstein, ”Distance measures for speech recognition, psychological and instrumental,” in Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., 1976, pp. 374–388.

[14] R. Shepard, ”Circularity in judgments of relative pitch”. Journal of the Acoustical Society of America. 36 (212) 1964, pp. 2346–2353.

[15] A. Mollahosseini, B. Hasani and M. H. Mahoor, ”AffectNet: A database for facial expression valence and arousal computing in the wild”, IEEE Trans. Affective Computing, vol. 10, no. 1, 2019, pp. 18–31.