International conference on Advanced Techniques in Communication Networking and Automation

RESEARCH ARTICLE OPEN ACCESS

# MULTIMODAL EMOTION RECOGNITIONUSING DEEP LEARNING

Dr. S. Padmapriya
Professor - CSE
A.V.C College of Engineering
Mannampandal, Mayiladuthurai
padmapriyas@avccengg.net

Deepika R
Monika T
Samyuktha G U
Sindhuja M K
B.E Computer Science and Engineering
A.V.C College of Engineering
Mannampandal, Mayiladuthurai
gusamyuktha@gmail.com

*Abstract*- **Multimodal emotion recognition in real-time involves the integration of multiple sources of information such as facial expressions, speech, text and psychological signals, to recognize and interpret human emotions in real-time. This approach has gained significant attention in recent years due to its potential applications in various fields, including healthcare, gaming and human computer interaction. Existing system focuses more on emotion recognition in unimodalities. Our project aims to optimize the performance of the emotion recognition system and presented a model for multimodal emotion recognition from audio and image data. This system introduces an improved Multimodal Emotion Recognition system that uses CNN as chosen model to examine a person's voice and facial expression and identify the corresponding emotion. The datasets were collected from Kaggle. These algorithms tend to recognize basic emotions like happy, fear, sad and angry.**

**Keywords - Multimodal emotion, CNN, MFCC**

## I. Introduction

Emotion recognition is a field of study that involves using technology to identify, analyze, and interpret human emotions. Multimodal emotion recognition is the process of recognizing and interpreting human emotions by integrating multiple sources of information, such as facial expressions, speech, physiological signals, and body language. Our approach aims to improve the accuracy of emotion recognition by combining information from multiple modalities that may complement or reinforce each other. Human emotions are complex and can be expressed in different ways through various modalities. For instance, a person may express sadness through a facial expression, a change in voice tone, and physiological changes in the body, such as changes in heart rate and breathing. Multimodal emotion recognition aims to capture these complex and dynamic expressions of emotions by combining information from different modalities. In our project we recognize emotions in real-time by using deep learning techniques.

## II. Related Work

In Reference[1] The authors investigate the ability of Inception-Resnet v2, CNN-LSTM deep learning networks to classify the human emotion. In this paper, The output of the system combines audio, text, and video data, taking into consideration their shared characteristics. The audio data is processed using a CNN-LSTM network, which incorporates both long-term and short-term memory to extract relevant features. The video data is processed using an Inception-Res Net-v2 network to extract facial expressions. The extracted features from both audio and video are combined using a SoftMax classifier in an LSTM network for emotion recognition. Additionally, a CNN-LSTM network is utilized to learn audio-emotional characteristics. The text data is processed using a Bi-LSTM network, and the merged features are classified using SoftMax. The final classification is generated by merging the results from all modalities using a logistic regression model. The proposed method achieves an 82.9% recognition accuracy on the IEMOCAP dataset.

In Reference[2] The paper aimed to investigate cutting-edge models for multimodal emotion recognition, focusing on textual, audio, and video inputs. The authors used XCeption architecture for finding facial emotion, Time distributed Convolutional Neural Network for speech emotion then Recurrent Neural Network and LSTM for text emotion. The authors develop an ensemble model that integrates information from all three modalities to present it in a transparent and understandable manner. They utilize the RAVDESS database, which contains speeches for seven

distinct emotions, but they exclude "Surprise" due to the classifier's difficulty in distinguishing it from other emotions. The final outcomes are highly satisfactory, as they achieve an accuracy score of nearly 75%.

Reference[3] The author suggests an asynchronous feature-level fusion method that combines separate signal measurements into a single hybrid feature space, which can be used for classifying or clustering multimedia content. They apply this approach to identify fundamental affective states using speech prosody and facial expressions. The experimental findings on two audiovisual emotion databases, which included 42 and 12 subjects, demonstrate that the proposed system outperforms unimodal systems based solely on facial expressions or speech, as well as synchronous feature-level and decision-level fusion methods.

Reference[4] The author of the paper has developed a system that classifies facial expressions and/or audio signals into one of seven emotions: anger, disgust, fear, happy, neutral, sad, and surprise. Both facial expressions and audio signals are easily obtainable from the user and are crucial in determining the emotional state. The two models are then combined based on the probabilities of different emotions to arrive at the final result. This emotion detection system has a wide range of practical applications. The project was implemented using Python, Keras, and Librosa and can determine the emotions of a person in real-time using either audio, video, or both. This project stands out from other similar applications as it utilizes ResNet50 and CNN, and we have compared its accuracy percentage of 65% to serverless, CNN with VGGFace, ResNet18, and IR50. To our knowledge, no other application has utilized ResNet50 with CNN. The technologies used in this project include ResNet50 CNN, Python, Keras, and Librosa.

Reference[5] In this research paper, the author proposed a hybrid deep convolutional neural network that utilizes both audio and visual information for emotion recognition. The visual network is based on Census-Transform and is capable of extracting facial expression features from videos with high discrimination ability. The audio data is first transformed into an image representation and then input into a 2D-Convolutional Neural Network (CNN) for feature extraction. The visual data is processed through a Census-Transform (CT) based on CNN. The audio and visual features are then fused, reduced through Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). Emotion recognition is achieved using K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB) classifiers. The experimental results on RML, eNTERFACE05, and BAUM-1s datasets show that the proposed model achieves competitive recognition rates compared to other state-of-the-art approaches

Reference[6] The authors examines the use of audiovisual information to identify human emotions, utilizing a cross-corpus evaluation with three different databases (SAVEE, eNTERFACE'05, and RML) as the training set and AFEW (a database that simulates real-world conditions) as the testing set. To represent emotional speech, commonly known audio and spectral features, as well as MFCC coefficients, are used. Classification is carried out using the SVM algorithm. Regarding facial expressions, the Viola-Jones face recognition algorithm is used to identify faces in key frames, and facial image emotion classification is conducted through CNN (AlexNet). Multimodal emotion recognition is achieved through decision-level fusion. The accuracy of the emotion recognition algorithm is compared with human decision makers' validation.

Reference[7] This paper explores the effectiveness of audio-visual emotion recognition, which combines facial expressions and affective speech. Local binary patterns (LBP) features are extracted for single facial expression recognition, while three typical acoustic features, including prosody features, voice quality features, and Mel-Frequency Cepstral Coefficients (MFCC) features, are extracted for single speech emotion recognition. The two modalities are then fused, and audio-visual emotion recognition is performed at the feature-level. Support vector machines (SVM) are used for all emotion classification. The study uses the eNTERFACE'05 emotional audio-visual database and reports experimental results that demonstrate the effectiveness of the presented method, achieving an accuracy of 66.51%, which outperforms mono-modality recognition.

## III. Proposed methodology

Our proposed approach involves training and implementing a microservice that uses camera and microphone inputs to detect the current emotion of a user based on their voice and facial expression. Real-time prediction of emotions is made possible by two separate models that analyze audio and video data simultaneously using Convolutional Neural Networks (CNNs). Initially, the bubble chart displays the real-time probabilities of each emotion for both video and audio separately (without considering multimodality). Then, by using a combined value of 70-30, the system responds to the user by displaying an emoji of the predicted multimodal emotion. The final prediction is a combined "multimodal" output that

is obtained by weighing and combining the two separate predictions.(figure 1)

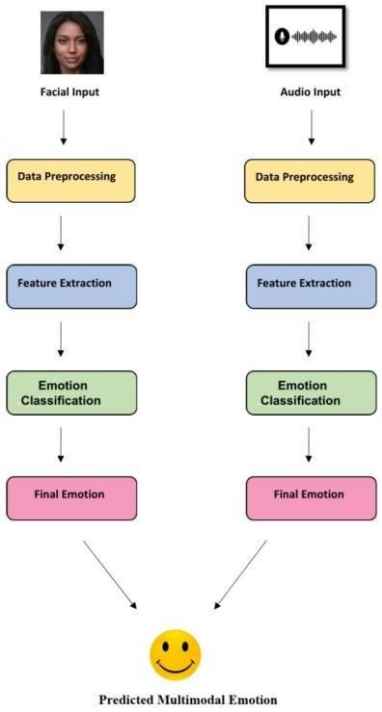


Figure 1: Architecture of proposed model

Datasets:

For Audio

- 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) Database
- Berlin Database of Emotional Speech (emo-db)

| Emotion | Number of Audio files |
|---|---|
| Angry | 652 |
| Fear | 652 |
| Happy | 652 |
| Sad | 652 |

Table 1

For Video

- CK+48 5 emotions (used 4 of them)

| Emotion | Number of Audio files |
|---|---|
| Angry | 135 |
| Fear | 75 |
| Happy | 207 |
| Sad | 84 |

Table 2

(i) Facial Emotion

A Convolutional Neural Network (CNN) was utilized to analyze facial expressions. The OpenCV library was employed to capture frames from the camera, detect faces in the frames using the Haar Cascade classifier, crop and resize the detected faces, and apply a pre-trained Keras model to predict the emotion of the person in the face. If the script detects one or more faces, it crops and resizes each face, converts it to a 4D tensor, and passes the resulting tensor to the pre-trained Keras model. The predicted emotion is subsequently written to a JSON file.

For training, the script reads all the images from a designated dataset directory and converts them into arrays. The training data is randomly shuffled to prevent any learning sequence. Next, the features and labels are separated from the training data, and the features array is reshaped to have four dimensions (batch size, height, width, and channels). The features data is then normalized by dividing each pixel value by 255. The script then loads a pre-trained MobileNetV2 model and removes the last classification layer. A new classification layer is appended to the model with a final output of four classes, representing the four facial emotions (angry, happy, sad, and neutral). The new model is compiled with sparse categorical cross-entropy loss and Adam optimizer. The model is trained for 25 epochs with the provided features and labels data. Then the emoji represents the facial expression in real-time.

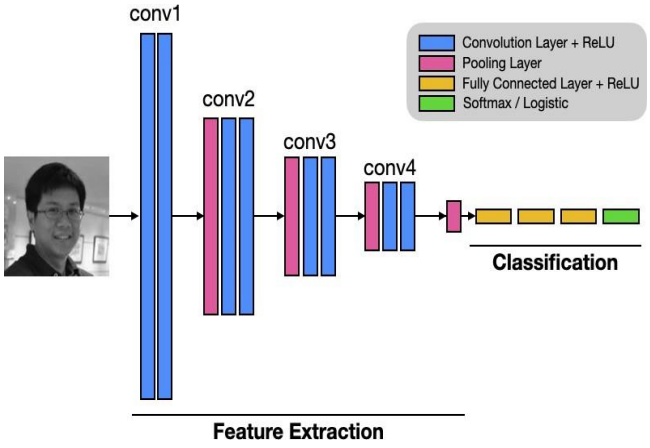


Figure 2: CNN for facial emotion recognition

(ii) Speech Emotion

Using the librosa library, audio processing is performed to compute various audio characteristics such as Mel-frequency cepstral coefficients (MFCC), root mean square (RMS) value, zero-crossing rate (ZCR), and chroma features. MFCC, a popular technique for speech and audio processing applications, is obtained through the Mel scale, which is a listener-based pitch perceptual scale. The librosa library is utilized to determine the MFCC. To increase the amount of training data for the neural network, the audio data is expanded by introducing noise, time-stretching, and

pitch-shifting. A pre-trained neural network is employed to classify the audio, which is loaded from the disk through the Keras library. The neural network takes in the audio features extracted from the audio data as inputs and predicts the emotion label. In summary, the process includes recording audio using the PyAudio library, saving it to a WAV file, extracting audio features through the methods mentioned above, inputting the audio features to the pre-trained neural network for prediction, and ultimately displaying the predicted emotion label.

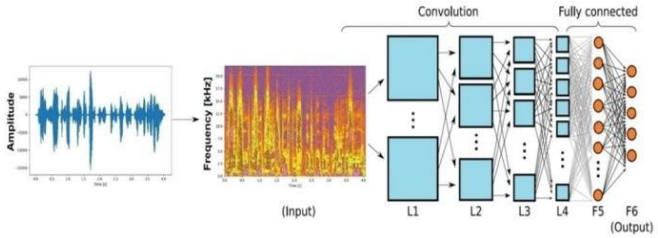


Figure 3 : CNN for speech emotion recognition

Zero Crossing Rate - The zero-crossing rate (ZCR) is **the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.**

Zero Crossing Rate

$$ZCR(A) = \frac{1}{2}\sum_{i=2}^{n} |sign(a_i) - sign(a_{i-1})| \quad (1)$$

Root Mean Square - The square root of the mean of the square. RMS is **a meaningful way of calculating the average of values over a period of time**.

Root Mean Square

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} \quad (2)$$

(iii) Multimodal Emotion

To enhance the accuracy and dependability of emotion recognition systems, a combined emotion recognition approach (70-30) is employed. This approach merges the results obtained from both video and audio analysis, with 70% weightage given to emotions identified through video analysis and the remaining 30% assigned to emotions detected through audio analysis. By integrating information from both sources, this method leverages the complementary data provided by video-based systems, which rely on facial expressions, head movements, and body language, and audio-based systems, which analyze speech patterns, tone, and other acoustic features.

This approach overcomes the limitations of relying on a single modality, such as facial occlusion, poor lighting conditions, or speech impediments. The final combined emotion score is calculated by taking a weighted average of emotions detected through both modalities, with the video-based score contributing 70% and the audio-based score contributing 30%.

Overall, this combined emotion recognition method has been proven to increase the accuracy of emotion recognition systems, particularly in scenarios where one modality is unreliable or noisy.

## IV. Results and Discussion

The dataset used for training is organized into 4 categories of audio and facial features, each corresponding to one of the four fundamental emotions: Happiness, Fear, Anger, and Sadness. The dataset has a roughly equal representation of all four emotions, as verified by human evaluators. The model underwent 25 epochs of training, and the outcomes are shown in Tables 3 and 4, which display the results of the convolutional neural network after being trained with the facial and audio data sets.

Furthermore, we conducted an analysis of audio-based emotions encompassing eight different categories, including happiness, sadness, anger, calmness, disgust, fear, neutrality, and surprise. The outcomes of this analysis are presented in Table 5.

For Image:

| CLASSES | PRECISION | RECALL | F1 |
|---|---|---|---|
| ANGRY | 0.93 | 1.0 | 0.96 |
| FEAR | 0.93 | 1.0 | 0.97 |
| HAPPY | 1.0 | 1.0 | 1.0 |
| SAD | 1.0 | 0.83 | 0.91 |

Table 3

For Audio:

| CLASSES | PRECISION | RECALL | F1 |
|---|---|---|---|
| ANGRY | 0.97 | 0.80 | 0.88 |
| FEAR | 0.82 | 0.85 | 0.84 |
| HAPPY | 0.85 | 0.77 | 0.81 |
| SAD | 0.81 | 0.97 | 0.88 |

Table 4

For Audio using 8 emotions:

| CLASSES | PRECISION | RECALL | F1 |
|---|---|---|---|
| ANGRY | 0.91 | 0.92 | 0.92 |
| CALM | 0.72 | 0.80 | 0.75 |
| DISGUST | 0.87 | 0.85 | 0.86 |
| FEAR | 0.90 | 0.87 | 0.89 |
| HAPPY | 0.85 | 0.82 | 0.83 |
| NEUTRAL | 0.89 | 0.88 | 0.89 |
| SAD | 0.86 | 0.90 | 0.87 |
| SURPRISE | 0.86 | 0.87 | 0.86 |

Table 5

The model exhibits superior performance when presented with facial input as opposed to audio input. Therefore, for the multimodal system, we assigned a greater weightage of 70% to facial input and a weightage of 30% to audio input.

## V. Conclusion

The purpose of this study was to develop a system capable of recognizing emotions based on both audio and facial input data. Given the distinct characteristics of these two types of signals, we initially explored the efficacy of utilizing either facial or speech information alone for emotion recognition. Our system also recognizes multimodal emotion by using the combined (70-30) weights of face and audio emotion. To accomplish speech-based emotion recognition, we employed the Convolutional Neural Network (CNN) in conjunction with several feature sets, including Zero Crossing Rate, Root Mean Square (RMS), and Mel Frequency Cepstral Coefficients (MFCC). For facial emotion recognition, we employed the Mobilenetv2 architecture, which is a type of Convolutional Neural Network. The model we developed exhibited favorable performance when evaluated using the provided datasets. Moving forward, we intend to improve the accuracy of emotion detection by further refining our model.

## References

[1] Seyad Sadegh Hosseini, Mohammed Raze Yamaghini, Soodabeh poorzaker Arabani- “Multimodal modelling of human emotion using sound, image and text fusion", Research square, 2023.

[2] Anatoli de Bradke, Mael Fabien, Raphael Lederman, and Stephane Reynal - “Multimodal Emotion Recognition”, Telecom ParisTech, Paris 75013, France, 2019.

[3] Muharram Mansoorizadeh, Nasrolah Charkari - "Multimodal information fusion application to human emotion recognition from face and speech", researchgate, 2014.

[4] Tushar Tandon, Divyansh Rastogi, Vikas Gupta, Dr. Sonu Mittal - "Emotion Detection Using Facial and Audio Features Using ResNet50 and CNN", IJRES, 2022.

[5] Jadisha Yarif Ramírez Cornejo and Helio Pedrini - "Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform", researchgate, 2019.

[6] Egils Avots · Tomasz Sapinski · Maie Bachmann · Dorota Kaminska - "Audiovisual emotion recognition in wild", Crossmark, 2018

[7] Shiqing Zhang, Lemin Li, and Zhijin Zhao - "Audio-Visual Emotion Recognition Based on Facial Expression and Affective Speech", springer, 2012

[8] Khadijeh Aghajani - "Audio-visual emotion recognition based on a deep convolutional neural network", JAIDM, 2022.

[9] Kritika Rupauliha, Aman Goyal, Aman Saini, Akshay Shukla, Sridhar Swaminathan - "Multimodal Emotion Recognition in Polish (Student Consortium)", IEEE, 2020.

[10] Hiranmayi Ranganathan, Shayok Chakraborty and Sethuraman Panchanathan - "Multimodal Emotion Recognition using Deep Learning Architectures", IEEE, 2016.

[11] Gaurav Sahu - "Multimodal Speech Emotion Recognition and Ambiguity Resolution", arXiv, 2019.

[12] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, Rada Mihalcea - "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations", arXiv, 2019.

[13] Naveed Ahmed , Zaher Al Aghbari, Shini Girija - "A systematic survey on multimodal emotion recognition using learning algorithms" International Journal of Intelligent Systems and Applications, 2023.

[14] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, Ashutosh Modi - "COGMEN: COntextualized GNN based Multimodal Emotion recognitioN", arXiv, 2022.