DOI: 10.21917/ijsc.2024.0477

EMOTION RECOGNITION THROUGH DEEP LEARNING IN VARIOUS MODES

Bharat Gupta[1] and Manas Gupta[2]
[1]*Ministry of Electronics and Information Technology, Government of India, India*
[2]*Department of Engineering Physics, Indian Institute of Technology Banaras Hindu University, Varanasi, India*

Abstract

The man-machine interface encompasses a crucial area—emotion recognition through facial expressions. Despite its significance, emotion recognition faces challenges such as facial accessories, non-uniform illuminations, pose variations, audio speeches, text conversations, and hand and facial gestures. Understanding emotions like happiness, anger, anxiety, joy, and shock, along with their varying degrees and overlaps, is essential for accurate recognition. These nuances, inherent to humans, pose difficulties and costs in achieving standard results through facial recognition. Recognizing someone's mood through facial expression, conversation, voice modulation, and gestures is a skill humans excel at. However, replicating this ability through facial recognition has proven challenging and costly. This paper addresses these challenges by proposing diverse approaches to emotion detection. By exploring various modes, including facial expressions, conversation analysis, voice modulation, and gestures, the paper tackles current research problems and holds practical applications in public experiments and exhaustive sentiment analysis. The paper presents a good combo of various modes of emotion recognition on multiple datasets (tried and tested widely before amalgamating all to produce an excellent optimal result as an output of the model).

Keywords:

FER, ASR, MFCC, Multimodal Deep Learning

1. INTRODUCTION

Human emotion recognition from visual cues [1] has become increasingly pivotal in computer vision over the last two decades, fuelled by its diverse applications in psychiatric treatment, social robotics [2], and education. The intricacies of emotional expression, marked by complexity and diversity, necessitate a multimodal approach to achieve accurate recognition. Recognition of human emotion extends beyond the confines of a singular modality. This insight is emphasized by studies in multimodal emotion recognition [3] [4] [5] [27] [28], which stress the inadequacy of relying solely on a single source for precise predictions. Emotion recognition encompasses explicit and implicit cues, incorporating facial expressions, eye movements, speech patterns, actions, and physiological signals [13] [17]. The challenges inherent in multimodal learning underscore the limitations associated with unimodal approaches.

The landscape of multimodal emotion recognition research unfolds across two key domains: the representation of raw data modalities and the fusion of these modalities preceding the prediction layer [6] [8]. While explicit affective cues and observable human changes yield valuable insights, exploring implicit affective stimuli in digital media introduces additional dimensions for analysis.

Overcoming challenges in representation, translation, alignment, fusion, and co-learning [7] becomes imperative, underscoring the intricacies of amalgamating information from diverse modalities. Recognizing the limitations inherent in unimodal approaches, researchers advocate for more modalities to improve accuracy [5] [28]. Leveraging complementary information from facial expressions, speech, and physiological signals contributes to a more comprehensive understanding of human emotion.

Despite the complexities involved, the fusion of modalities [8] emerges as a crucial strategy, enabling a more nuanced and accurate prediction of emotional states. Co-learning further emphasizes knowledge transfer between modalities, particularly valuable in scenarios with limited resources, thereby enhancing recognition accuracy.

Upon referencing previous papers, it became evident that leveraging multiple modalities is crucial for achieving enhanced accuracy in emotion recognition compared to individual modalities. The integration of video, audio, and text modalities allows for a holistic understanding of human emotion, compensating for the limitations inherent in each particular modality. By combining visual [26], auditory, and textual cues, the multimodal system [27] gains a more robust and nuanced representation of the user's emotional state, thereby achieving superior accuracy in emotion recognition compared to relying on any single modality in isolation. This comprehensive approach aligns with the broader trend in research, emphasizing the synergistic benefits of multimodal systems for more accurate and reliable emotion recognition.

After extensive research and deliberation, a strategic decision has been made to employ distinct datasets for three models: Video/Visual Emotion Recognition, Audio Emotion Recognition, and Text Emotion Recognition. The selected datasets for model training are as follows: FER2013 plus [29], ExpW cleaned [30], [31], RAF-DB [32], [33] for Video Recognition, SAVEE [10], RAVDESS [34], TESS [35], CREMA-D [36] for Speech Emotion Recognition, and CARER [11] for Text Emotion Recognition. This meticulous dataset selection aligns with each modality's specific requirements and characteristics, ensuring optimal model training and performance across diverse domains of emotion recognition.

2. RELATED WORKS

Prior research in human emotion understanding predominantly employed unimodal [12] or bimodal approaches [13], focusing primarily on facial and auditory modalities. Studies by Noroozi et al. [14], Bandela et al. [15], and Zamil et al. [16] notably concentrated on bimodal systems, incorporating visual and auditory signals with features such as MFCC, FBEs, and spectral characteristics for emotion classification. However, the limitations of these approaches in providing a comprehensive description of human emotions became evident.

A critical assessment of the current research landscape underscores a need for more papers integrating multiple modalities in emotion recognition, namely Video/Visual, Speech Audio, and Text emotion recognition. Existing studies often combine facial emotion recognition with EEG or gestures [16], imposing constraints on practical applications. Upon scrutinizing approximately 200 datasets, the research team identified a scarcity of comprehensive options encompassing all three modalities, leading to the strategic selection of IEMOCAP [18], MOSEI [19], and MOSI [20] datasets for a multimodal approach.

The imperative shift towards multimodal deep learning is underscored by the shortcomings of preceding unimodal and bimodal studies. While bimodal approaches using visual and auditory signals have shown promise [13], expanding to include more modalities, as depicted in Figure 1, significantly enhances accuracy. Multimodal deep learning [22], [28], processing information across linguistic, acoustic [24], [27], and visual modalities, aligns with the nuanced complexity of human reactions. This comprehensive approach is pivotal for achieving superior accuracy in emotion recognition tasks, transcending the limitations of unimodal or bimodal frameworks and advancing the field substantively.

3. METHODOLOGY

This study concurrently employs three models to predict emotions in short video clips, enhancing accuracy through parallel comparison and normalization. Optimal models from audio, images, video, text, and gestures literature are integrated for improved accuracy and interpretability. The model allows the labelling of unlabelled data via fusion across modalities and aims to develop lightweight models with performance comparable to memory-intensive ones. The model's block diagram is referenced below.

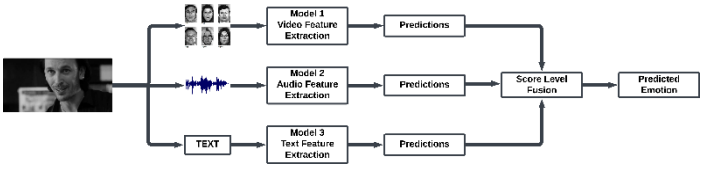

Fig.1. Flowchart depicting the sequence of models being used and how the final result has been concluded using the fusion technique

If we are assigned the emotion recognition task of an actor in a particular short video, then to solve it, we will use three unlike models trained on different datasets for predicting specific outputs as per the above image. Those models shall be:

- Emotion recognition model 1 by learning text as the outcome from ASR or subtitling
- Emotion recognition model 2 by learning speech waveforms
- Emotion recognition model 3 by learning facial expressions

It is imperative to extract specific details from the short video. This process involves retrieving the number of frames from the video, which will be forwarded to model 3. Model 3 will then analyse each frame to identify the emotion, and the output from all frames will be compiled for a comprehensive result. Simultaneously, we will extract text using ASR or subtitling techniques. This extracted text will be directed to model 1, while the speech waveform will be channelled to model 2 for emotion identification. It is crucial to note that these three tasks, frame analysis, text extraction, and speech waveform analysis, are distinct and operate independently.

Consequently, our approach involves deploying these models in parallel, allowing for simultaneous processing of each task. We propose two evaluation methods for our multimodal engine to ensure a confident output. The first involves aggregating the probabilities of different emotions from each model and selecting the maximum value. Alternatively, a voting mechanism can be employed, considering each model's output for a particular emotion. This dual evaluation strategy enhances the robustness and reliability of our multimodal emotion prediction system.

Above is an example of Multimodal deep learning where different types of neural network-based models trained for specific tasks are used to extract the features for appropriate output. This approach will give equal importance to all the unlike modalities by combining high-level embeddings from the different kinds of input by concatenating them and then applying a SoftMax to get an output. The following are the sub-modules of the project:

3.1 SPEECH AUDIO

From the audio data, we have extracted three critical features used in this study: MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram, and Chroma. The Python implementation of the Librosa package was used in their extraction. MFCC, crucial for analysing time signals, represents the vocal tract envelope by applying the Fourier Transform on a spectrogram. Mel Spectrogram is obtained through Fast Fourier Transform on signal segments, mapping amplitude on a Mel scale. Chroma, a 12-element vector, indicates pitch class energy. For training, a multilayer perceptron is used to generate specific outputs. This approach enhances our understanding of SER [23] models and their integration based on user input.

3.2 VIDEO

The benchmark paper chosen for this study is 'Facial Emotion Recognition: State of the Art Performance on FER2013' by Yousif Khaireddin and Zhuofa Chen [9]. This paper reported a 73% accuracy on the FER2013 dataset, surpassing the state-of-the-art accuracy of 68% from a Kaggle contest. The primary model utilized in this study is VGGNet [25], [26], and the official source code was employed to achieve an accuracy of 79.6% in 50 iterations. This accuracy is deemed sufficient, given the intended use of this model in conjunction with other models.

Recognizing the non-uniform distribution of data across the seven emotions, we opted to train the model on multiple datasets. We implemented a thorough process to address the non-uniformity and enhance the quality of the data. Which involved cleaning the images by hand and ensuring they met high standards. For balancing, we performed automatic augmentation, and to further refine the dataset, we engaged in both up-sampling by adding new augmented data and down-sampling by manually removing images from larger categories. This meticulous approach was applied to merge and normalize the FER2013 plus [29], ExpW [30], [31], and RAF-DB [32], [33] datasets, resulting

in a more refined, balanced, and augmented model training dataset.

3.3 TEXT RECOGNITION

Implementing text emotion recognition involves Bidirectional LSTM, a sequence processing model comprising two LSTMs, one processing input in a forward direction and the other in reverse. This bidirectional approach enhances the algorithm's contextual understanding by considering preceding and succeeding words in a sentence. BiLSTMs excel in sequence classification [21], training two LSTMs on the input sequence simultaneously, one on the original and the other on the reversed sequence. This bidirectional context addition accelerates results. In text recognition, Bidirectional Recurrent Neural Networks (BRNNs) utilize information from past and future time steps, offering a comprehensive understanding of the entire text context. The network processes time steps individually but simultaneously move through the sequence in both directions, optimizing contextual interpretation.

3.4 MULTIMODAL EMOTION RECOGNITION (INTEGRATED MODELS)

For experimenting, we have taken the following steps:

- Taken YouTube video clips of different emotions as mp4 video.
- Divided test video into frames and saved it as images under the output folder.
- Converted mp4 file to wav file for extracting out speech spectrogram.
- Further, through MFCC, Mel spectrogram and chroma, converted wav output into embeddings to pass user given input.
- Get text output from speech by using Google ASR API.
- Pass those three outcomes [images, speech spectrogram embeddings, text] into our trained models for predicting emotions.
- Used voting mechanism for facial emotion recognition. For example, if a video has been divided into five frames as images then the emotion of each image is identified and the maximum occurrence of emotion from the output array is chosen, and the image emotion is announced.
- Similarly, we took speech and text emotions from our trained models and again applied a voting mechanism for deciding video emotion.

4. DATA COLLECTION

4.1 VISUAL EMOTION RECOGNITION (VIDEO)

We carefully compiled and refined datasets from three primary sources: FER2013, ExpW, and RAF-DB. The goal was to create a comprehensive and high-quality dataset that would facilitate precise emotion detection during the initial training phase.

4.1.1 Dataset Compilation:

- FER2013: It is a well-known facial expression dataset that contains 35,887 grayscale, 48x48 pixel images of faces, each labeled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Despite its extensive use, FER2013 [29] contains some images with ambiguous or misleading facial expressions that can hinder the model's learning process.
- ExpW: Expression in the Wild dataset [30],[31] offers a more diverse range of expressions captured in real-world scenarios. This dataset includes around 91,793 images, each labeled with one of seven emotions. The variety in background, lighting, and facial poses makes it a valuable resource for training models intended for practical applications.
- RAF-DB: Real-world Affective Faces Database [32],[33] is another critical dataset containing 29,672 facial images, also labeled with seven basic emotion categories. The images in RAF-DB are collected from the internet, covering a wide range of ages, races, and occlusions, providing a diverse training ground for emotion recognition models.

4.1.2 Handpicking and Cleaning the Data:

To ensure the accuracy of the emotion detection model, we meticulously identified and removed ambiguous images from each dataset. The process involved the following steps:

- Manual Review: Each image was carefully examined to identify and remove those that did not clearly represent the labeled emotion. This step was crucial in eliminating potential sources of confusion for the model.
- Cross-Validation: Images were compared across the three datasets to ensure consistency in labeling and to remove any duplicates or mislabeled images.
- Consensus Labeling: In cases where the emotion label was uncertain, a consensus labeling approach was employed, requiring agreement on the emotion displayed in the image.

4.1.3 Data Augmentation:

After cleaning the datasets, we augmented the data to achieve a uniform distribution of images across all emotion categories. This step was essential to enhance the model's learning process and mitigate overfitting. Our data augmentation techniques included:

- Rotation: Slightly rotating images to create variations in the dataset, helping the model learn to recognize emotions from different angles.
- Flipping: Horizontally flipping images to increase the diversity of facial expressions.
- Scaling and Cropping: Randomly scaling and cropping images to introduce variability in face sizes and positions within the frame.
- Brightness and Contrast Adjustment: Modifying the brightness and contrast of images to simulate different lighting conditions.

4.1.4 Ensuring Uniform Data Distribution

To prevent any single emotion category from dominating the dataset and causing the model to overfit, we carefully balanced

the number of images in each category. This uniform distribution was achieved by:

- Oversampling: Increasing the number of images in underrepresented emotion categories through augmentation techniques.
- Under sampling: Reducing the number of images in overrepresented categories to maintain balance.
- Synthetic Data Generation: Creating synthetic images for certain to ensure adequate representation without compromising the quality of the dataset.

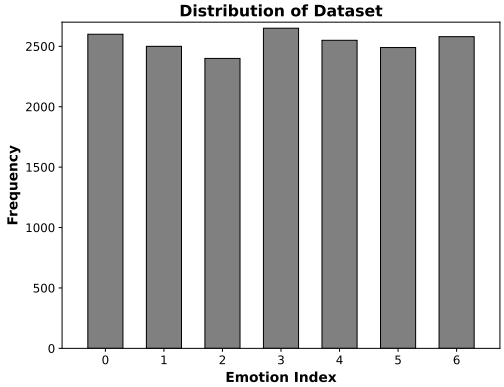

Fig.2. Data distribution of emotions after augmentation

4.2 AUDIO EMOTION RECOGNITION

We collected and processed datasets such as SAVEE, RAVDESS, TESS, and CREMA-D. These datasets were specifically chosen for their comprehensive and varied recordings of emotional speech. We used Mel-Frequency Cepstral Coefficients (MFCCs) to extract relevant features from the audio samples, enabling accurate emotion detection. The library used for processing, Librosa, provided the necessary tools to analyze and transform the sound signals effectively.

4.2.1 Datasets collected

- SAVEE: The Surrey Audio-Visual Expressed Emotion database [10] contains recordings of male actors expressing seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The recordings are clean and high-quality, providing a solid foundation for emotion recognition.
- RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song [34] includes 24 professional actors, male and female, vocalizing two lexically matched statements. The recordings cover eight different emotions and are captured under controlled conditions.
- TESS: The Toronto Emotional Speech Set [35] consists of 200 target words spoken by two actresses aged 26 and 64. The dataset covers seven emotions and is designed to study how different age groups express emotions vocally.
- CREMA-D: The Crowd-Sourced Emotional Multimodal Actors Dataset [36] features 91 actors, male and female, representing diverse ages and ethnic backgrounds. The dataset includes recordings of 12 sentences spoken in six different emotions.

4.2.2 Feature Extraction Using Librosa:

Librosa is a powerful Python library for audio and music analysis. It provides a comprehensive suite of tools to process and analyze sound signals. One of the key features of Librosa is its ability to compute MFCCs, which are widely used in speech and audio processing for their ability to capture the timbral aspects of audio signals.

4.2.3 Processing Audio Files:

To prepare the audio data for training, we used the following process:

- Loading the Audio File: The audio files were loaded using librosa.load, which allowed us to resample the audio to a consistent sample rate and trim the duration for uniformity.
- Extracting MFCCs: We computed the MFCCs for each audio file, capturing the essential features required for emotion recognition.
- Padding the MFCCs: The extracted MFCC features were padded to ensure uniform dimensions across all samples, facilitating efficient processing by the machine learning model.

4.2.4 MFCC Dimensions:

In this process, we extracted 13 MFCCs from each audio sample. The choice of 13 coefficients is standard practice in speech processing, as it captures the most significant features of the speech signal. The padded MFCC array had a length of 216, ensuring consistency across all samples and simplifying the input structure for the machine learning model.

4.3 TEXT EMOTION RECOGNITION

The CARER dataset was sourced from publicly available repositories known for high-quality text data labeled with emotions. It includes a variety of text samples, such as social media posts, news articles, and customer reviews, each tagged with specific emotions like joy, anger, sadness, surprise, and fear. The collection process involved scraping data from these sources using Python scripts and APIs, ensuring a diverse and comprehensive dataset.

Preprocessing is a multi-step procedure to transform raw text data into a suitable format for model training. Here's how we approached it:

4.3.1 Initial Data Exploration and Cleaning:

- Inspection: We began by inspecting the dataset to understand its structure and identify any anomalies.
- Cleaning: This involved removing duplicate entries, correcting misspelled words, and standardizing text formats.

4.3.2 Handling Missing Values

Any missing entries, especially those without emotion labels, were either filled using imputation techniques or removed to maintain data integrity.

4.3.3 Text Normalization Techniques

- Lowercasing: Converting all text to lowercase to ensure uniformity.

- Removing Punctuation and Special Characters: Eliminating unnecessary characters that do not contribute to the emotion detection process.
- Tokenization: Splitting sentences into individual words or tokens to facilitate analysis.

4.3.4 Stop Words Removal:

Common words that do not carry significant meaning (e.g., "and", "the", "is") were removed to focus on the more informative words.

4.3.5 Lemmatization and Stemming:

Converting words to their root or dictionary form (e.g., "running" to "run").

4.3.6 Data Preparation for Model Training:

Once preprocessed, the text data needs to be converted into a numerical format understandable by machine learning models. We use the following vectorization method:

- Dense and Fixed-Length Vectors: Word embeddings like Word2Vec, and GloVe provide dense, fixed-length vector representations of words, which are suitable for CNNs.
- Semantic Information: These embeddings capture semantic relationships and contextual information, which are essential for understanding emotions in text.

5. RESULTS

5.1 SPEECH EMOTION RECOGNITION

After 50 epochs, Validation Accuracy: 60%

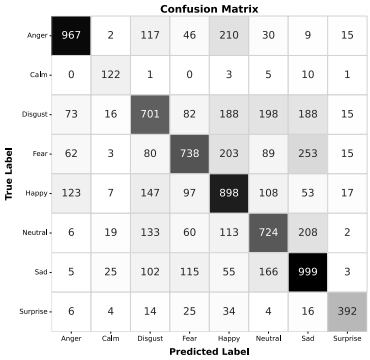

Fig.3. Confusion matrix between true label vs predicted label

5.2 FACE EMOTION RECOGNITION MODEL

After 60 epochs, Validation Accuracy: 79.6%

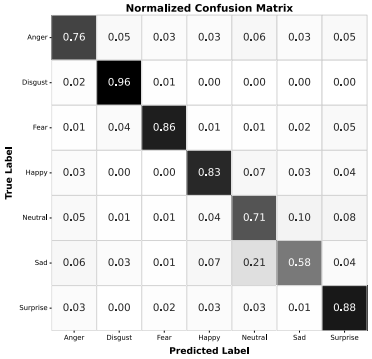

Fig.4. Confusion matrix between true label vs predicted label

5.3 TEXT RECOGNITION MODEL:

After 15 epochs, the Validation Accuracy: 91%.

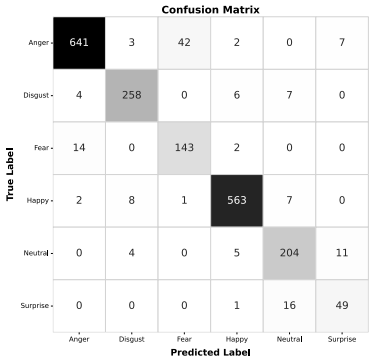

Fig.5. Confusion matrix between true label and predicted label

5.4 MULTIMODAL EMOTION RECOGNITION (INTEGRATED)

Following are sample output:

Result 1 : when of God's secret, it's my coffee
Speech Emotion: Happy: 0.88
Facial Emotions:
1st image- Happy: 0.84, 2nd image- Happy: 0.94, 3rd image- Happy: 0.76
Text Emotion: Surprise: 1.0 on text "when of God's secret it's my coffee"
Overall Emotion using voting mechanism: **HAPPY**

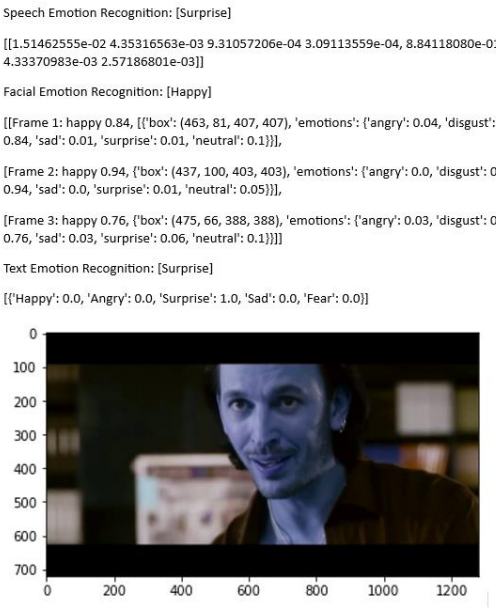

Fig.6. Testing our Multi Modal Emotion Recognition Model

5.5 TESTING INTEGRATED MODEL

5.5.1 Data Collection and Labelling:

We began by manually collecting 300 video clips, each ranging from 3 to 4 seconds in duration. These clips were sourced from various environments to ensure diversity in emotional expressions. The primary goal was to capture a wide range of emotions in natural settings. After gathering the clips, we proceeded with the labeling process. Each video was meticulously reviewed and assigned an emotion label.

5.5.2 Extracting Modalities:

To effectively analyze the emotional content of the videos, we extracted three different modalities from each clip:

- Image Frames: We extracted individual frames from each video clip. This step involved selecting key frames that captured the essence of the emotion displayed.
- Audio: The audio track from each video was isolated and processed. We focused on features such as pitch, tone, and volume, which are often indicative of emotional states.
- Text: If present, any spoken words in the video were transcribed to text. This was done using speech-to-text algorithms to convert spoken language into textual data.

5.5.3 Model Architecture:

For our analysis, we implemented three separate models, each tailored to one of the extracted modalities:

- Deep CNN for Image Frames: This model was designed to process the visual information from the frames. It utilized convolutional layers to capture spatial hierarchies and features that are indicative of emotional expressions.
- Encoder-Decoder for Audio: The architecture was configured to analyze sound patterns. It employed spectrograms as input, transforming audio signals into a visual representation that the model could process effectively. The encoder part of the architecture consisted of several convolutional layers that extracted spatial features from the spectrograms, followed by recurrent layers like LSTMs to capture the temporal dynamics of the audio signals. The decoder then used these extracted features to predict the emotion labels, often incorporating attention mechanisms to focus on the most relevant parts of the input sequence.
- Embedding-based Text Classifier for Text: The text CNN focused on the transcribed speech. Using embeddings, this model captured the semantic meaning of the words, which often convey significant emotional context.

5.5.4 Fusion of Indivisual Results

Once each CNN model processed its respective modality, we employed a fusion technique to integrate the outputs. The fusion mechanism combined the features learned from each modality into a unified representation.

5.5.5 Evaluation and Results

After training the models, we evaluated their performance by comparing the predicted emotions against the true labels. The results were promising, with our system achieving an accuracy of 81%. It Is important to note that our sample size for testing the model was relatively small, which can influence the variability of the accuracy. With a larger dataset, we anticipate that the accuracy could be more stable and potentially higher.

6. FUSION TECHNIQUE

This project explored multimodal emotion recognition, revealing three critical scenarios that shaped our fusion technique:

6.1 CASE 1: CONSISTENT PREDICTIONS BY AT LEAST TWO MODALITIES

In the presence of consistent predictions from two modalities and a divergence from the third, the agreement between the two congruent models determined the final result. This approach prioritized concordance to enhance the model's robustness.

6.2 CASE 2: PREDICTIONS WITH CLOSE SECONDARY EMOTIONS

A confidence system was implemented in scenarios where all three models predicted primary and secondary emotions with closely matched probabilities. Examining the top two predicted emotions and their probabilities for each model, a confidence score was assigned. To enhance precision, a fixed threshold of 0.67 was set. Only predictions exceeding this threshold were considered accurate, providing a stringent criterion for final emotion determination. This approach effectively addressed instances where primary and secondary predicted emotions were highly competitive across all modalities.

6.3 CASE 3: DIVERGENT PREDICTIONS ACROSS ALL MODALITIES

In scenarios of divergent predictions, a specialized voting mechanism incorporated individual model confidence scores based on historical accuracy. This dynamic strategy enabled nuanced decision-making, considering the reliability of each modality.

By concurrently evaluating these three cases, our algorithm ensures trustworthy and precise emotion recognition, adapting its decision-making process to the characteristics of the input data. This comprehensive approach is essential for advancing the field of multimodal emotion recognition, enabling the model to perform effectively in diverse contexts.

7. CONCLUSION

This paper introduces a sophisticated multimodal emotion recognition system, leveraging speech, facial, and text models to discern emotions, including Anger, Sadness, Happiness, and neutrality. The successful outcomes in our initial experiment validate the effectiveness of our integrated trained models. Looking ahead, our project's evolution involves incorporating gesture emotion recognition, broadening the spectrum of emotional cues for a nuanced understanding of user sentiment. Furthermore, we plan to implement voting mechanisms to refine the accuracy of emotion predictions, drawing insights from each model. The unavailability of the IEMOCAP dataset in our current study is a noted limitation. Upon its acquisition (a 16.5 GB zipped file comprising speech, facial, and text emotions as videos), future research endeavours will harness its diverse emotional

expressions, contributing to a more comprehensive and impactful multimodal emotion recognition system. This commitment underscores our dedication to advancing the field, with applications extending to human-computer interaction and affective computing.

REFERENCES

[1] P.P. Filntisis, N. Efthymiou, G. Potamianos and P. Maragos, "Emotion Understanding in Videos through Body, Context, and Visual-Semantic Embedding Loss", *Proceedings of International Conference on Computer Vision*, pp. 1-6, 2020.

[2] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar and P. Dario, "Emotion Modelling for Social Robotics Applications: A Review", *Journal of Bionic Engineering*, Vol. 15, No. 2, pp. 1-14, 2018.

[3] K. Ezzameli and H. Mahersia, "Emotion Recognition from Unimodal to Multimodal Analysis: A Review", *Information Fusion*, Vol. 99, pp. 1-11, 2023.

[4] J. Zhang, Z. Yin, P. Chen and S. Nichele, "Emotion Recognition using Multi-Modal Data and Machine Learning Techniques: A Tutorial and Review", *Information Fusion*, Vol. 59, pp. 1-17, 2020.

[5] N. Ahmed, Z. Aghbari and S. Girija, "A Systematic Survey on Multimodal Emotion Recognition using Learning Algorithms", *Intelligent Systems with Applications*, Vol. 17, pp. 1104-1113, 2023.

[6] Y. Cimtay, E. Ekmekcioglu and S. C. Ozhan, "Cross-Subject Multimodal Emotion Recognition based on Hybrid Fusion", *IEEE Access*, Vol. 8, pp. 11119-11125, 2020.

[7] A. Rahate, R. Walambe, S. Ramanna and K. Kotecha, "Multimodal Co-Learning: Challenges, Applications with Datasets, Recent Advances and Future Directions", *Information Fusion*, Vol. 81, pp. 1-18, 2022.

[8] G. Papandreou, A. Katsamanis, V. Pitsikalis and P. Maragos, "Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition", *Proceedings of IEEE Workshop on Multimedia Signal Processing*, pp. 1-7, 2007.

[9] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-11, 2021.

[10] P.J.B. Jackson and S.U. Haq, "Survey Audio-Visual Expressed Emotion (SAVEE) Database", Available at http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/ , Accessed in 2011.

[11] E. Saravia, H.C.T. Liu, Y.H. Huang, J. Wu and Y.S. Chen, "Carer: Contextualized Affect Representations for Emotion Recognition", *Proceedings of IEEE International Conference on Empirical Methods in Natural Language Processing*, pp. 1-5, 2018.

[12] P.S. Tomar, K. Mathur and U. Suman, "Unimodal Approaches for Emotion Recognition: A Systematic Review", *Cognitive Systems Research*, Vol. 77, pp. 1-13, 2023.

[13] F. Muhammad, M. Hussain and H. Aboalsamh, "A Bimodal Emotion Recognition Approach through the Fusion of Electroencephalography and Facial Sequences", *Diagnostics*, Vol. 13, No. 5, pp. 1-8, 2023.

[14] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips", *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp. 1-13, 2019.

[15] S.R. Bandela, S.S. Priyanka, K.S. Kumar, Y.V.B. Reddy and A.A. Berhanu, "Stressed Speech Emotion recognition using Teager Energy and Spectral Feature Fusion with Feature Optimization", *Computational Intelligence and Neuroscience*, Vol. 2023, pp. 1-12, 2023.

[16] A.A. Zamil, S. Hasan, S.M. Jannatul Baki, J.M. Adam and I. Zaman, "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames", *Proceedings of International Conference on Robotics, Electrical and Signal Processing Techniques*, pp. 2221-2226, 2019.

[17] Y. Huang, J. Yang, S. Liu and J. Pan, "Combining Facial Expressions and Electroencephalography to Enhance Emotion Recognition", *Future Internet*, Vol. 11, No. 5, pp. 1-22, 2019.

[18] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee and S.S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database", *Language Resources and Evaluation*, Vol. 42, No. 4, pp. 1-13, 2008.

[19] A. Bagher Zadeh, P.P. Liang, S. Poria, E. Cambria and L.P. Morency, "Multimodal Language Analysis in the Wild: CMU-Mosei Dataset and Interpretable Dynamic Fusion Graph", *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1-12, 2018.

[20] A. Zadeh, R. Zellers, E. Pincus and L.P. Morency, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages", *IEEE Intelligent Systems*, Vol. 31, No. 6, pp. 82-88, 2016.

[21] F.A. Acheampong, C. Wenyu and H.N. Mensah, "Text-Based Emotion Detection: Advances, Challenges, and Opportunities", *Engineering Reports*, Vol. 2, No. 7, pp. 78-89, 2020.

[22] W. Mellouk and W. Handouzi, "Facial Emotion Recognition using Deep Learning: Review and Insights", *Procedia Computer Science*, Vol. 175, pp. 1-21, 2020.

[23] H. Aouani and Y.B. Ayed, "Speech Emotion Recognition with Deep Learning", *Procedia Computer Science*, Vol. 176, pp. 1-11, 2020.

[24] B. Schuller, G. Rigoll and M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in A Hybrid Support Vector Machine-Belief Network Architecture", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1-8, 2006.

[25] K. Simoyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1-4, 2015.

[26] G.C. Porusniuc, F. Leon, R. Timofte and C. Miron, "Convolutional Neural Networks Architectures for Facial Expression Recognition", *Proceedings of IEEE International Conference on E-Health and Bioengineering*, pp. 1-7, 2019.

[27] S. Zhang, S. Zhang, T. Huang and W. Gao, "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition", *Proceedings of ACM on International Conference on Multimedia Retrieval*, pp. 331-336, 2016.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A.Y. Ng, "Multimodal Deep Learning", *Proceedings of International Conference on Machine Learning*, pp. 224-229, 2011.

[29] E. Barsoum, C. Zhang, C.C. Ferrer and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution", *Proceedings of ACM International Conference on Multimodal Interaction*, pp. 465-475, 2016.

[30] Z. Zhang, P. Luo, C.C. Loy and X. Tang, "Learning Social Relation Traits from Face Images", *Proceedings of International Conference on Computer Vision*, 2015.

[31] Z. Zhang, P. Luo, C.C. Loy and X. Tang, "From Facial Expression Recognition to Interpersonal Relation Prediction", *International Journal of Computer Vision*, Vol. 126, No. 5, pp. 1-15, 2017.

[32] S. Li, W. Deng and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-11, 2017.

[33] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 28, No. 1, pp. 1-13, 2019.

[34] S.R. Livingstone and F.A. Russo, "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*, vol. 13, no. 5, 2018.

[35] M.K. Pichora-Fuller and K. Dupuis, "Toronto Emotional Speech Set (TESS)", Available at https://borealisdata.ca/dataset.xhtml?persistentId=doi%3A10.5683%2FSP2%2FE8H2MF, Accessed in 2020.

[36] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova and R. Verma, "Crema-D: Crowd-Sourced Emotional Multimodal Actors Dataset", *IEEE Transactions on Affective Computing*, Vol. 5, No. 4, pp. 377-390, 2014.