International Journal of Research Publication and Reviews, Vol (6), Issue (5), May (2025) Page – 8933-8937

WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Facial and Voice Emotion Detection via Deep Learning Methods

Ms. KavyaL[1], Ms.Merlin Mary James[2]

[1]PG Student, Department of Computer Science Engineering, Mangalam College of Engineering, Ettumanoor, kerala, India.
e-mail:
[2]AssistantProfessor,Department of Computer Science Engineering, Mangalam College of engineering, Ettumanoor, kerala,
E-mail: [1]lkavya805@gmail.com, [2]merlin.james@mangalam.in

Abstract

The project presents a multimodal emotion recognition system uses deep learning techniques to analyze and interpret human emotions through facial expressions and voice signals. The system integrates two primary modals: Facial Emotion Recognition (FER) and Voice Emotion Recognition (VER), to achieve a more comprehensive and accurate assessment of emotional states. FER is implemented using both a custom Convolutional Neural Network (CNN) and a pre-trained ResNet model to classify emotions from facial images. VER utilizes hybrid deep learning architectures, including CNN-LSTM and CNN-GRU, to extract and analyze temporal and spectral features from audio data. After evaluating the performance of each model, the most effective architectures are integrated into a unified system. The final solution is deployed as a web application, with a React-based front-end and a Python-powered back end, enabling real-time emotion detection through a user-friendly interface.

Keywords: Facial expressions, Voice analysis, CNN- LSTM, CNN-GRU.

I. Introduction

Comprehending human emotions impacts human-computer interaction, spanning sectors, such as healthcare, education, customer services, etc. Emotion recognition systems recognize and discern human emotions based on different human signals, such as facial signals and vocal prosody. With the overall improvement of deep learning techniques, subsequent technical improvements have resulted in subsequent increases in performance and robustness in emotion recognition systems. This presents a proposed development of a multi-modal human emotion recognition system that combines facial and voice to recognize human emotions. The system consists of two central components: Facial Emotion Recognition (FER) as well as Voice Emotion Recognition (VER). Facial Emotion Recognition (FER) is implemented using a custom Convolutional Neural Network (CNN) and pre-trained ResNet model, when emotion classification is made based on a facial dataset (e.g., expression). Voice Emotion Recognition (VER) is addressed using hybrid deep learning architecture, including CNN-LSTM and CNN-GRU, in which, spectral and temporal features are extracted from the audio signal recordings. After evaluating and discovering which models perform the best, these selected suitable models will join as one unified emotion recognition framework. Finally, to make the system accessible and usable, it is a web application, which utilizes a React-based web application front-end with a python back-end petition. This way, the web application demonstrates real-time emotion detection, while being an interactive and usable software application. People are using computers in an ever-increasing number of ways in daily life -and when inferring emotions from human-computer interaction becomes critical for shared understanding, there will be demand for systems that are able to both recognize and respond appropriately to different emotional states. Emotion recognition technology could apply in many domains - conversational agents, personal / mental health applications, customer service automation, education - and many overlap each other, as multiple possibilities for avatars and virtual agents that respond to human emotion also overlap. Current emotion recognition technology ignores the use of multiple modalities, either facial expressions or voice, by using one source of individually people's way of expression, that can cause measurement error in context, such as ambiguity, environmental disturbances, and so on. Tools of multimodal emotion recognition are great! Their multimodal aspects are beneficial in terms of emotional cues from facial expressions and vocalizations can be complementary; and the individual modality systems enhanced reliability, and human-human discussions quite often complement and reinforce facial expressions with the use of their voices. Deep learning models (CNNs; CNN-LSTM; CNN-GRU) provide opportunities to learn more complex features automatically and with relatively large datasets of multimodal observation to help improve performance toward recognizing emotional human states. The intention of the current project is to exploit the use of deep learning in multimodal analysis of emotion to create an emotion recognition system that will be more accurate and robust than current emotion recognition systems. The implementation of the facial and voice-based models in a web application can demonstrate the user-centered platform of advanced emotion recognition systems, as well as a technology solution for deploying new multimodal deep learning advances for real time human-computer interaction.

II RelatedWork

[1] AayushiChaudhari (2023) proposes a multimodal emotion recognition system that combines facial expression and speech analysis using deep learning. Gabor filters are applied to facial images to extract spatially sensitive features, which are then processed by a Convolutional Neural Network (CNN). For speech, Mel-Frequency Cepstral Coefficients (MFCCs) are used to capture emotional tone, also fed into a CNN. To reduce dimensionality, Principal Component Analysis (PCA) is employed, and Support Vector Machines (SVM) are used for final emotion classification. The model is trained on diverse datasets, including JAFFE, EMOTIC, RAVDESS, and TESS, resulting in improved accuracy across visual and audio inputs.

[2] In the study "Automatic Recognition of Student Emotions from Facial Expressions During a Lecture," G. Tonguç (2020) investigates student engagement by analyzing facial expressions using webcam footage during live lectures. The system, developed in C and integrated with Microsoft's Emotion Recognition API, enables real-time emotion detection in classroom settings. Emotional data from 67 students was analyzed using MANOVA to assess variations across different lecture phases. While the study offers practical insights into student attention, it faces limitations such as data loss due to poor visibility, lack of long-term emotion tracking, and the absence of qualitative student feedback for deeper analysis.

[3] "Facial Emotion Recognition Using Deep Learning: Review and Insights," WafaMellouk (2020) offers a detailed overview of deep learning techniques used in facial emotion recognition (FER). Published in Procedia Computer Science, the review emphasizes the importance of preprocessing steps like resizing and normalization to improve model performance. Mellouk analyzes various deep learning architectures, including CNNs, CNN-LSTMs, and 3D CNNs, with a focus on their ability to capture both spatial and temporal facial features. The paper also highlights key FER datasets such as FER2013 and AffectNet, and explores data augmentation methods to address limitations in training data and enhance model robustness.

[4] "Emotion Recognition on Speech Processing Using Machine Learning," Vikrant Chole explores the classification of emotional states from speech using machine learning techniques. The system processes audio by segmenting it into 60-millisecond frames with 10-millisecond overlaps, preserving continuity in the signal. Features are extracted using Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Coefficients (MFCCs), which capture key spectral and perceptual aspects of speech. The K-Nearest Neighbor (KNN) algorithm is used for classification due to its effectiveness with clearly defined feature clusters. Trained on the Berlin and Spanish speech emotion databases, the model shows potential but faces limitations, including overfitting, limited emotional nuance detection, insufficient training data, and the need for better noise filtering and feature selection to improve real-world performance.

[5] In the 2023 article "Speech Emotion Recognition Based on Emotion Perception," Gang Liu introduces a biologically inspired framework for speech emotion recognition (SER), published in the EURASIP Journal on Audio, Speech, and Music Processing. Drawing from neuroscience, Liu employs a multi-task learning approach to mimic human-like, intuitive emotion perception, allowing the model to implicitly classify emotions without explicit labeling. Trained on the IEMOCAP dataset, the system effectively captures nuanced affective cues, leading to improved recognition accuracy. However, the study notes limitations, including the lack of diverse, high-quality datasets and challenges in generalizing SER models across different languages, accents, and speaking styles.

III.EXISTING SYSTEM

Emotion recognition systems are considered to have a future in the monitoring of psychological health and are receiving a surge of interest currently. However, to date, there are still few systems that focus on continuous and real-time psychological health assessment, which emotion recognition is primarily focused on to date. There are two key stages in emotion recognition, that are usually leading: Facial emotion recognition, which uses a video retained image, often using a Gabor filter that captures spatial frequency characteristics or using a Convolutional Neural Network (CNN) that automates the learning and classification of facial image complexities. On the other hand, voice emotion recognition is concerned with the acoustic features of speech. Perhaps the best known technique is the Mel-Frequency Cepstral Coefficients (MFCC) that mimics human auditory perception and ability to discriminate emotional tone in speech via connected acoustic features. In order to train and validate these systems, researchers often find large databases such as the Surrey Audio-Visual Expressed Emotion (SAVEE) database useful as it combines complementary audio visual data into a single performance modified with emotion label annotation. Similar to SAVEE, there are a number of databases and standards produced by emotional speech database researchers to promote certain stability and generalizability between (1) speakers, (2) languages, and (3) emotions in practice with emotion recognition models. To address the challenge of high-dimensional data, dimensionality reduction techniques such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) are used. These techniques simplify the data while preserving the most informative features, leading to improved processing efficiency and classification accuracy.

IV. PROPOSED SYSTEM

A. Facial Emotion Recognition

[i] Image Preprocessing

Preprocessing is a critical step in facial emotion recognition, ensuring that image data is properly formatted for deep learning models. Datasets such as FER2013 or CK+, which include labeled facial expression images are first loaded. All images are resized to a standard resolution commonly 224×224 pixels to ensure consistency across batches and compatibility with model input requirements.

Facial regions are detected using Multi-Task Cascaded Convolutional Networks (MTCNN), a robust deep learning-based face detector that identifies facial landmarks and generates bounding boxes. Detected faces are then cropped and normalized—scaling pixel values to a range like [0, 1] or [-1, 1] to match CNN input expectations.

The dataset is then split into training, validation, and testing subsets. The training set is used to fit model parameters, the validation set to tune hyperparameters, and the test set to evaluate generalization on unseen data. This preprocessing pipeline ensures high-quality, standardized inputs for optimal model performance.

[ii] Model Development

Two main strategies are typically used to build facial emotion recognition models. The first involves developing a custom Convolutional Neural Network (CNN) tailored to emotion classification. This network consists of multiple convolutional and pooling layers that extract spatial features from facial images, followed by fully connected layers for final classification. Model parameters such as filter size, kernel shape, activation functions, and dropout rates are optimized to enhance performance and reduce overfitting.

The second strategy uses transfer learning, commonly with pre-trained models like ResNet-50. Originally trained on large datasets such as ImageNet, ResNet provides strong feature extraction capabilities. In this approach, earlier layers of ResNet are frozen, while the final layers are fine-tuned using the facial emotion dataset. This not only reduces training time but also improves accuracy, even with limited labeled data, making it a practical choice for real-world applications.

[iii] Model Training and Evaluation

Training begins with compiling the model using key configurations such as the optimizer, loss function, and evaluation metrics. The Adam optimizer is widely used for its ability to adapt learning rates dynamically. For multi-class emotion classification, the categorical cross-entropy loss function is commonly used to measure the difference between predicted probabilities and actual labels.

While accuracy gives a general sense of model performance, it may not fully reflect success in cases of class imbalance. Therefore, additional metrics are used:

- Precision: Measures the proportion of correct positive predictions.
- Recall: Indicates how well the model identifies all instances of a specific class.
- F1 Score: The harmonic mean of precision and recall, offering a balanced evaluation, especially useful in imbalanced datasets.

[iv] Model Deployment

Once the model—either the custom CNN or fine-tuned ResNet—achieves satisfactory performance based on metrics like accuracy and F1-score, it is selected for deployment. The trained model is saved using formats such as HDF5 (.h5), TensorFlowSavedModel, or PyTorch's .pth, allowing easy integration into applications without retraining. These formats enable deployment across various platforms, supporting real-time emotion recognition in fields like healthcare, education, and customer service.

B. Voice Emotion Recognition

[i] Audio Preprocessing

Audio preprocessing is a crucial first step in voice emotion recognition, aimed at preparing and enhancing raw audio data for deep learning. Datasets like RAVDESS are used, offering not only speech samples but also metadata including speaker identity and labeled emotional categories. To increase robustness and generalizability, data augmentation techniques such as pitch shifting, time stretching, and adding background noise are applied. These techniques simulate real-world variations in speech, helping the model become resilient to diverse acoustic conditions.

After augmentation, feature extraction is performed to convert audio waveforms into formats suitable for machine learning. The most commonly used features are Mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs). These representations effectively capture critical speech characteristics like pitch, tone, and energy distribution, all of which are closely linked to emotional expression in voice signals.

[ii] Model Development

To accurately recognize emotions from voice, a hybrid CNN-LSTM architecture is employed. Convolutional Neural Networks (CNNs) are used first to process input spectrograms and extract localized spatial features related to pitch, frequency shifts, and vocal intensity—key indicators of emotion. These spatial features are then passed to Long Short-Term Memory (LSTM) layers, a type of Recurrent Neural Network (RNN) designed to retain temporal information over long sequences.

The CNN captures fine-grained acoustic patterns, while the LSTM models how these features evolve over time. This combination allows the system to understand both the immediate and sequential nature of emotional cues in speech, making the hybrid model highly effective in speech emotion recognition.

[iii] Model Training and Evaluation

The model training process begins with compilation, where key settings like the loss function and optimizer are defined. For multi-class emotion classification, categorical cross-entropy is typically used as the loss function, measuring the difference between predicted and actual emotion classes. Optimizers like Adam or Stochastic Gradient Descent (SGD) help adjust model weights efficiently during backpropagation.

The model is trained on a designated training set, while performance is monitored on a validation set to fine-tune hyperparameters and prevent overfitting. Once training is complete, the model is evaluated on a separate test set to assess generalization. Evaluation metrics include accuracy, precision, recall, and F1 score. While accuracy offers a broad success rate, precision and recall reveal how well specific emotion categories are identified. The F1 score, balancing both, is especially important when dealing with imbalanced emotion distributions in the dataset.

[iv] Model Deployment

After evaluation, the best-performing model—either the CNN-LSTM or an alternative like CNN-GRU—is selected for deployment. The chosen model is saved in a standard format (e.g., .h5, .pth, or TensorFlowSavedModel) to retain trained weights and configurations. This enables seamless integration into applications such as virtual assistants, call center analytics, or mental health monitoring systems, where real-time voice-based emotion detection can provide valuable user insights.

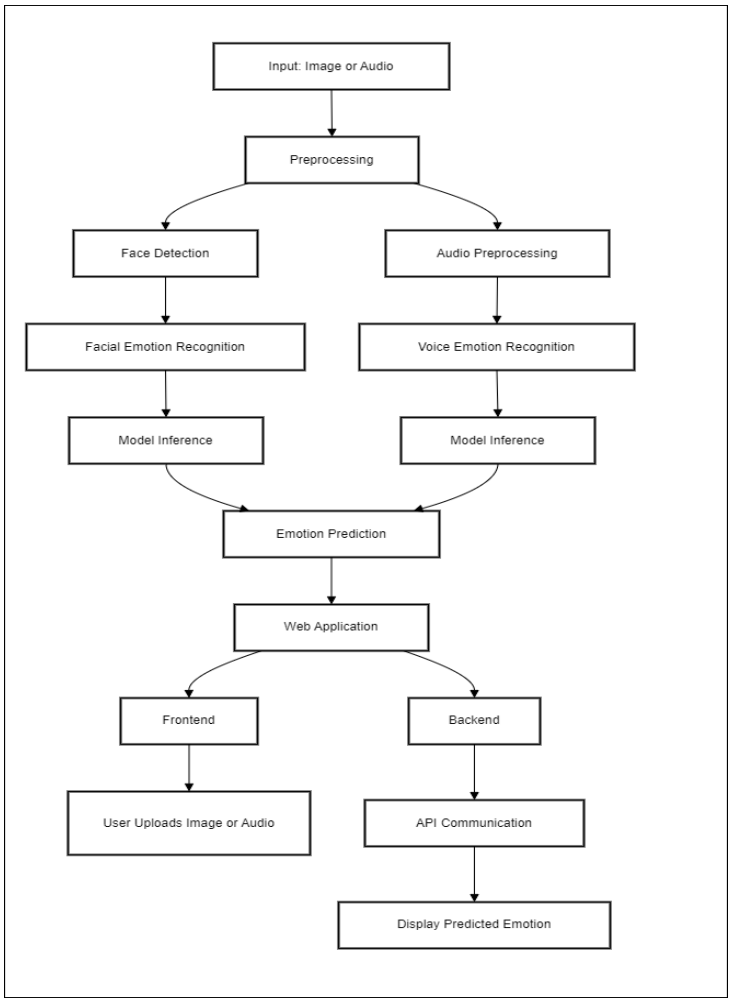

Fig a. Block Diagram for proposed system

C. System Architecture and Hardware Requirements

Theoverallsystem isstructuredaroundtwocoremodules: Facial Emotion Recognition and Voice Emotion Recognition. These modules function by capturing real- timevideoinputfromadevice'scameraorwebcam,which is then processed in sequential stages. Each module is organized into four fundamental phases: Image Preprocessing, Model Development, Model Training and Evaluation, and Model Deployment.

For development, the system leverages tools such as Python IDLE, Anaconda, and Visual Studio Code, all of which contribute to an efficient and manageable coding environment. Theplatformruns on Windows 10, offering compatibility with various software libraries and frameworks. It supports both mobile and desktop application development—Flutter is utilized for frontend design, while backend development is handled in Python usingwebframeworkslikeFlaskorDjango.Thesystemis operated on hardware equipped with an Intel i5 processor and 12GB of RAM, which provides reliable performance for programming, model training, and application testing.

V. Conclusion and Future Work

Through the integration of facial expression and voice signal analysis with deep learning techniques, this project made a holistic contribution to multimodal emotion recognition. A custom Convolutional Neural Network (CNN) combined with a pre-trained ResNet model enabled efficient and accurate classification of facial emotion with high training and validation accuracy. The CNN-LSTM and CNN-GRU architectures exploited the spatial and temporal features in voice data effectively. Although there are limitations, the high training accuracy, and train and validation accuracy of the models overall support the conclusion that it was successful in generalizing across s forms of emotional inputs. The obvious application of a web-based application using React for the front end and Flask/Django backend clearly demonstrates potential deployment in a real-time scenario, such as mental health monitoring, virtual learning environments or interactive AI systems. The use of databases like SAVEE and numerous emotion-labeled audio sets were instrumental in supporting model training and evaluation, ensuring the robustness and consistency of performance.

Future Recommendations:

As for further research and validity testing, there are some improvements that can be made to the study. First: expanding the dataset we used to include more emotions, languages, ages and ethnicities will increase the generalizability of the model. Second: it would add value to the model if the predictions were intelligently fused in a multimodal approach, to improve the truthfulness and accuracy of outcomes. Third: optimizations can be made to the performance of the model in real-time and also use model compression techniques like quantization or pruning to deploy the model on a mobile or edge device. Fourth: the model can also incorporate adaptive learning techniques such as online or continual learning to allow the system to learn with new user data, as it becomes available. Lastly, if physiological properties (i.e. heart rate, EEG signals) were to be incorporated into the system it may improve the understanding of human emotional states and facilitate the move towards a more complete and continuous psychological monitoring system.

REFERENCES

1. V. Sreenivas, V. Namdeo, and E. Vijay Kumar, "Modified deep belief network-based human emotion recognition with multiscale features from video sequences," Software: Practice and Experience, vol. 51, no. 6, pp. 1259-1279, Jun. 2021.
2. H. Liao et al., "Deep learning enhanced attributes conditional random forest for robust facial expression recognition," Multimedia Tools and Applications, vol. 80, no. 19, pp. 28627- 28645, 2021.
3. F. Wang, H. Chen, and L. Kong, "Real-time facial expression recognition on robots for healthcare," in 2018 IEEE International Conference on Intelligence and Safety for Robotics, Shenyang, China, pp. 402-406, 2018.
4. L. M. Martinez and R. Valenti, "Deep learning for non-human emotion recognition: A review," Artificial Intelligence Review, vol. 53, no. 1, pp. 1-20, 2020.
5. S. Yin, Y. Wang, and J. Luo, "Recognizing animal emotions via deep learning: A survey," Pattern Recognition Letters, vol. 131, pp. 128-135, 2020.
6.Aayushi Chaudhari,Chintan Bhatt, ThanhThi Nguyen, NisargPatel,KirtanChavda, KalindSarda. “Emotion Recognition System via Facial Expressions and Speech Using Machine Learning and Deep Learning Techniques”. Vol 2023
7. Sonawane B, Sharma P. “Deep learning based approach of emotion detection and grading system. Pattern Recognit Image Anal”. Vol 2020.
8. Prof. Vikrant Chole, Ms. NamrataYerande, Ms. AartiShinde,” EMOTION RECOGNITION ON SPEECH PROCESSING USING MACHINE LEARNING”,Vol 2023
9. GürayTonguç, BetulOzaydınOzkara,” Automatic recognition of student emotions from facial expressions during a lecture ”, Volume 2020
10. Gang Liul, Shifang Cai1 and Ce Wang,” Speech emotion recognition based on emotion perception”, volume 2023