Project Report

On

“Emotion Recognition through Audio and Facial Analysis”

Submitted by

Sarvesh Bhaspale (333005)

Omkar Halpatrao (333021)

Jayesh Jaiswal (333025)

In partial fulfillment for the award of

Third Year of Engineering

(Information Technology)

Guided by

Prof. Swati Patil

Information technology

Vishwakarma Institute of Information Technology, Pune.

(2023- 2024)

CERTIFICATE

This is to certify that, the project "Emotion Recognition through Voice and Facial Analysis" submitted by

Sarvesh Bhaspale (333005)

Omkar Halpatrao (333021)

Jayesh Jaiswal (333025)

is a Bonafide work completed under supervision and guidance of **Prof. Swati Patil,** and it is submitted towards the partial fulfillment for award of Third year of engineering (IT) of Savitribai Phule Pune University Aurangabad.

Prof. Swati Patil
GUIDE
Information Technology

Prof. Swati Patil
PROJECT COORDINATOR
Information Technology

DR. Pravin Futane
Information Technology

Dr. Vivek Deshpande
Vishwakarma Institute of Information Technology, Pune.

Place: PUNE SEAL

Date: 2/12/2023

Examination Approval Sheet

The project report entitled

" Emotion Recognition through Voice and Facial Analysis"

Submitted by

Sarvesh Bhaspale (333005)

Omkar Halpatrao (333021)

Jayesh Jaiswal (333025)

is approved for the degree of Third year of engineer (Information Technology) of Vishwakarma Institute of Information Technology, Pune.

.

Internal Examiner:

External Examiner:

Place: VIIT, Pune

Date: 2/12/2023

DECLARATION

I hereby declare that I have completed and written the project entitled **"Emotion Recognition through Voice and Facial Analysis"**. It has not previously been submitted for the basis of the award of any degree or diploma or similar title of this for any other diploma/examining body or university.

Sarvesh Bhaspale (333005)

Omkar Halpatrao (333021)

Jayesh Jaiswal (333025)

Date: 2/12/2023 TY(IT)

Abstract:

Emotion Detection through a Fusion of Audio and Video Modalities

This project presents an innovative approach to real-time emotion detection by combining audio and video modalities. Leveraging deep learning models, the system processes audio input from a microphone and video input from a camera simultaneously to accurately identify and classify human emotions. The video component utilizes a convolutional neural network (CNN) to analyze facial expressions, while the audio component employs a machine learning model trained on audio features extracted from emotional speech samples.

The video processing module incorporates a Haar Cascade Classifier for real-time face detection, capturing facial regions for subsequent emotion prediction. A pre-trained CNN model assesses facial features, providing insights into the emotional state of the individual. Concurrently, the audio processing module captures and analyzes emotional cues from the user's voice, utilizing features such as Mel-frequency cepstral coefficients (MFCC), chroma, and mel spectrograms. These audio features are fed into a pre-trained machine learning model, enabling the system to discern emotional patterns from speech.

The fusion of audio and video outputs results in a comprehensive emotion prediction, offering a more nuanced understanding of the user's emotional state. The system employs a user-friendly graphical interface built with Tkinter, allowing users to initiate emotion detection from both video and audio sources, providing real-time feedback. The incorporation of parallel processing threads enhances system responsiveness and ensures a seamless user experience.

This project contributes to the field of affective computing by demonstrating the effectiveness of multi-modal emotion detection. The fusion of audio and video modalities enhances the robustness and accuracy of emotion prediction, paving the way for applications in human-computer interaction, virtual assistants, and emotion-aware technologies.

Introduction

1.1 Introduction

This research paper investigates techniques for detecting emotions in people by focusing on facial and voice expressions as the main channels for emotional transmission. It explores a number of research projects that make use of recurrent models, feature extraction methods, and convolutional neural networks (CNNs) for both visual and audio data. The paper discusses the difficulties in identifying complex emotions, the limitations imposed by the environment, Considering the need for substantial datasets to enhance the generalization and accuracy of facial emotion recognition systems. Analogously, Speech Emotion Recognition (SER) research is essential, utilizing several datasets such as CREMA-D, RAVDESS, SAVEE, and IEMOCAP. The analyses compare spectrograms and Mel-spectrograms as input data to illustrate how feature extraction methods and data splitting affect model accuracy. The study also includes feature extraction techniques like Mel Frequency Cepstral Coefficients (MFCC). The review aims to consolidate advancements, challenges, and future directions in these two domains, emphasizing the need for more comprehensive evaluation in real-world settings, ethical considerations, and the potential for multi-modal recognition to enhance emotion detection accuracy across diverse datasets and scenarios.

1.2 Objective/Purpose

The primary objective of this project is to develop a robust and real-time emotion detection system that leverages information from both audio and video modalities. The key objectives include:

1. Multi-Modal Emotion Detection: Implement a system that simultaneously processes facial expressions from video inputs and speech features from audio inputs to achieve a more comprehensive understanding of the user's emotional state.

2. Real-Time Processing: Ensure the system's responsiveness by employing parallel processing threads for video and audio, allowing for instantaneous emotion detection without perceptible delays.

3. Facial Feature Extraction: Utilize a Convolutional Neural Network (CNN) for real-time face detection and feature extraction from facial expressions, contributing to accurate and dynamic emotion recognition.

4. Audio Feature Extraction:Implement audio feature extraction techniques, such as Mel-frequency cepstral coefficients (MFCC), chroma, and mel spectrograms, to capture emotional cues present in the user's speech.

Literature Review:

The literature review delves deeply into various methodologies and advancements in Facial Emotion Recognition (FER) through neural networks. It extensively covers studies employing diverse datasets and techniques, aiming to enhance emotion recognition from facial expressions. These studies utilized approaches such as local binary patterns (LBP), histogram of oriented gradients (HOG), convolutional neural networks (CNNs), support vector machines (SVM), k-Nearest Neighbour (k-NN), and transfer learning for feature extraction and classification from datasets like CK+, JAFFE, and FER2013. They highlighted challenges related to environmental conditions, dataset biases, and ethical considerations in FER technology. The importance of applications across healthcare, user experience design, and human-computer interaction was underscored.

Moreover, detailed preprocessing techniques were explored in another set of studies, focusing on noise reduction, edge detection, and feature extraction before integrating CNNs for emotion recognition. While these studies demonstrated performance improvements, they lacked comprehensive evaluations in diverse scenarios and fell short in addressing dataset diversity and ethical considerations regarding privacy and biases. Challenges were identified in handling larger datasets, preventing overfitting, and improving classification accuracy across various emotional expressions.

Additionally, a study emphasized the use of a self-cure relation network for nuanced emotion recognition despite label noise in training data. This research showcased potential contributions in human-computer interaction, mental health, and behavioral analysis. Challenges revolved around validating the model's efficacy with noisy training data and ensuring its generalization in real-world scenarios.

Further investigations incorporated dual-channel face emotion detection utilizing Gabor features extraction and channel attention networks, aiming to improve accuracy. Challenges remained in handling larger datasets, preventing overfitting, and enhancing classification accuracy across diverse emotions.

Lastly, the review extensively covered neural network architectures (CNNs, RNNs, ResNet, VGG, etc.) and their roles in emotion recognition, highlighting advancements in transfer learning and addressing challenges in real-time applications. The review emphasized the importance of interpreting contextual cues and individual variations, acknowledging the limitations in accurately understanding human emotions in real-world scenarios.

In summary, the comprehensive review provided insights into evolving methodologies, strengths, and challenges within FER utilizing neural networks. While these approaches show promise across various domains, further refinements, including dataset diversity, ethical considerations, and model robustness, are crucial for reliable and accurate emotion recognition in practical settings.

Methodology:

For Face Analysis:

The choice between using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks depends on the nature of the data and the characteristics of the task. Let's discuss why CNNs are commonly used for image-related tasks, such as facial expression detection, and why RNNs/LSTMs are more suitable for sequence-related tasks:

CNNs for Image-Related Tasks (e.g., Facial Expression Detection):

Local Hierarchical Feature Extraction:

Spatial Hierarchies: CNNs are designed to capture spatial hierarchies and local patterns in data. In images, features like edges, textures, and shapes are often localized in specific regions. CNNs use convolutional layers to detect these features in a hierarchical manner.

Parameter Sharing:

Weight Sharing: CNNs use weight sharing through convolutional kernels. This allows the network to learn a small set of shared parameters that are applied across the entire image. This is particularly effective for images where the same patterns may appear in different spatial locations.

Translation Invariance:

Shift-Invariance: CNNs inherently have translation invariance. This means they can recognize patterns regardless of their location in the image. This property is beneficial for tasks like object recognition and facial expression detection where the position of the face in the image may vary.

2D Grid Structure:

Image Grids: Images have a 2D grid structure, and CNNs are well-suited to exploit this structure. Convolutional operations scan through local regions of the input, capturing spatial relationships effectively.

RNNs/LSTMs for Sequence-Related Tasks (e.g., Natural Language Processing, Time-Series):

Temporal Dependencies:

Sequential Patterns: RNNs and LSTMs are designed to capture temporal dependencies in sequences. They have a memory mechanism that allows them to remember information from previous time steps, making them suitable for tasks where the order of data matters.

Variable-Length Input:

Dynamic Sequences: RNNs can handle input sequences of variable lengths. This is crucial for tasks where sequences may have different lengths, such as natural language processing where sentences can vary in length.

Long-Term Dependencies:

Memory Cells: LSTMs, a type of RNN, are specifically designed to address the vanishing gradient problem in standard RNNs. They can capture long-term dependencies in data, making them suitable for tasks where understanding context over a more extended period is important.

In summary, for image-related tasks like facial expression detection, CNNs are favored due to their ability to capture spatial hierarchies, local patterns, and translation invariance. On the other hand, RNNs and LSTMs are better suited for tasks involving sequential data where temporal dependencies and long-term context are essential.

For Audio Analysis:

1. Multi-Layer Perceptron (MLP):

Characteristics:

Feedforward neural network.

Consists of an input layer, one or more hidden layers, and an output layer.

Suitable for tasks where data patterns are complex but not necessarily sequential.

Each neuron in a layer is connected to every neuron in the subsequent layer.

Advantages for Speech Emotion Recognition:

Simple and efficient for smaller datasets.

Can capture non-linear relationships in feature space.

Limitations:

May struggle with capturing temporal dependencies and sequential patterns in the data

Conclusion:

This project has successfully implemented a real-time emotion detection system using facial expressions from video and speech features from audio. The integration of CNNs for face detection and feature extraction techniques like MFCC, chroma, and mel spectrograms for audio has resulted in a responsive and accurate emotion recognition system. The GUI provides an intuitive interface, and pre-trained models ensure adaptability and reliability. The project sets the stage for advancements in emotionally intelligent systems with applications in human-computer interaction and mental health monitoring.

References:

[1] R. Abou Zafra, A. A. Abdullah, L. Alaraj, R. Albezreh, and R. Barhoum, “An experimental study in Real-time Facial Emotion Recognition on new 3RL dataset,” 2023. [Online]. Available: https://github.com/

[2] X. Wang, Y. Wang, and D. Zhang, “Complex Emotion Recognition via Facial Expressions with Label Noises Self-Cure Relation Networks,” Comput. Intell. Neurosci., vol. 2023, pp. 1–10, Jan. 2023, doi: 10.1155/2023/7850140.

[3] T. Kumar Arora et al., “Optimal Facial Feature Based Emotional Recognition Using Deep Learning Algorithm,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/8379202.

[4] D. Ammous, A. Chabbouh, A. Edhib, A. Chaari, F. Kammoun, and N. Masmoudi, “Designing an Efficient System for Emotion Recognition Using CNN,” J. Electr. Comput. Eng., vol. 2023, pp. 1–11, Sep. 2023, doi: 10.1155/2023/9351345.

[5] Z. Y. Huang et al., “A study on computer vision for facial emotion recognition,” Sci. Rep., vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-35446-4.

[6] A. Pandey, A. Gupta, and R. Shyam, “FACIAL EMOTION DETECTION AND RECOGNITION,” Int. J. Eng. Appl. Sci. Technol., vol. 7, no. 1, pp. 176–179, May 2022, doi: 10.33564/IJEAST.2022.v07i01.027.

[7] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets,” Electron., vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223831.

[8] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network,” Appl. Sci., vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084750.

[9] J. L. Bautista, Y. K. Lee, and H. S. Shin, “Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation,” Electron., vol. 11, no. 23, Dec. 2022, doi: 10.3390/electronics11233935.

[10] S. Chamishka et al., “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling,” Multimed. Tools Appl., vol. 81, no. 24, pp. 35173–35194, Oct. 2022, doi: 10.1007/s11042-022-13363-4.

.