MULTI-MODAL EMOTION RECOGNITION FROM AUDIO AND VISUAL DATA

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Maridu Laasya Sri (2020BCS0174)
Eenadula Bhanuprakash (2020BCS0011)
Guda Phanish (2020BCS0151)
Gotam Gorabh (2020BCS0173)

to

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
KOTTAYAM-686635, INDIA

November 2023

DECLARATION

I, **MARIDU LAASYA SRI** (**Roll No: 2020BCS0174**), **EENADULA BHANUPRAKASH** (**Roll No: 2020BCS0011**), **GUDA PHANISH** (**Roll No: 2020BCS0151**), **GOTAM GORABH** (**Roll No: 2020BCS0173**), hereby declare that, this report entitled **"MULTI-MODAL EMOTION RECOGNITION FROM AUDIO AND VISUAL DATA"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** in **COMPUTER SCIENCE AND ENGINEERING** is an original work carried out by us under the supervision of **DR. SIVAIAH BELLAMKONDA** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. We have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Kottayam-686635	**MARIDU LAASYA SRI**
November 2023	**EENADULA BHANUPRAKASH**
	GUDA PHANISH
	GOTAM GORABH

CERTIFICATE

This is to certify that the work contained in this project report entitled **"MULTI-MODAL EMOTION RECOGNITION FROM AUDIO AND VISUAL DATA"** submitted by **MARIDU LAASYA SRI** (**Roll No: 2020BCS0174**), **EENADULA BHANUPRAKASH** (**Roll No: 2020BCS0011**), **GUDA PHANISH** (**Roll No: 2020BCS0151**), **GOTAM GORABH** (**Roll No: 2020BCS0173**), to the Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** has been carried out by them under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635	Dr. SIVAIAH BELLAMKONDA
November 2023	Project Supervisor

ABSTRACT

Multimodal emotion recognition is a vital field of research and technology that strives to enhance the accuracy and comprehensiveness of emotion assessment by integrating information from multiple sensory modalities. Detecting emotions automatically is a complex task because emotions can manifest through diverse channels of expression. The practical uses of this technology span across multiple fields, such as multimedia retrieval and interactions between humans and computers. Deep neural networks have achieved remarkable efficacy in determining emotional states, serving as a source of inspiration for our endeavor. We introduce an emotion recognition system that leverages both auditory and visual modalities. Beyond the mere extraction of characteristics, the ability to effectively manage atypical data points and capture contextual information holds paramount significance. To confront this challenge, we employ Long Short-Term Memory (LSTM) for audio and CNN networks for the image and video related. we used JAFFE, CK+48, CFE for the image and TESS for the speech emotion recognition.

Contents

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

1.1 History 3

1.2 Existing Techniques 4

1.2.1 Deep Learning 5

1.2.2 Machine Learning 6

1.2.3 Convolutional Neural Networks (CNNs) 7

1.2.4 Long Short-Term Memory (LSTM) 7

1.3 Evolution 8

1.3.1 Evolution of CNN for Facial Expression Recognition (FER) 8

1.3.2 LSTM in Audio 9

2 Literature Review **10**

3 Problem Statement **19**

4 Architecture **21**

5 Proposed Methodology **24**

5.1 Proposed Architecture 24

5.1.1 Image Model 24

5.1.2 Audio Model 27

5.2 Model Training 29

5.2.1 Image Model 29

5.2.2 Audio Model 31

5.3 Dataset 31

5.3.1 Toronto emotional speech set (TESS) 31

5.3.2 Jaffe Dataset 32

5.3.3 CK+48 Dataset 32

5.3.4 CFE Dataset 33

6 Results and Discussion **34**

6.1 Image Model 34

6.1.1 Discussion 35

6.1.2 Conclusion 36

6.2 Audio Model 37

7 Future work **39**

Bibliography **40**

List of Figures

2.1 Microexpression Recognition Methods 13

2.2 An overall block diagram of the proposed emotion recognition system 15

4.1 Architecture of multi-modal emotion recognition system 21

5.1 Architecture of audio 29

5.2 Architecture of audio 30

6.1 Training and Validation Accuracy 36

6.2 Training and Validation Loss 37

6.3 Training and Validation accuracy of audio model 38

6.4 Training and Validation Loss of audio model 38

List of Tables

1.1 Comparison of Machine Learning and Deep Learning 5

Chapter 1

Introduction

Emotion recognition in making human-computer interaction more effective. It enables computers to understand and respond to human emotions, enhancing various applications. However, emotions are incredibly subjective and can differ greatly among individuals. People often experience mixed emotions, adding complexity to accurately categorizing emotional states.[2] Consider fatigue monitoring, for instance, where emotion states can be employed to monitor and predict one's fatigue level. In fields like speech recognition, emotion recognition finds utility in call centers, helping detect the emotional state of callers and offering feedback on service quality.[3] Yet, recognizing emotions is not easy due to the nature of human emotions. The lack of distinct temporal boundaries and the varied ways individuals express emotions further complicate the process. While traditional efforts in emotion recognition have mostly focused on deducing emotional states from speech, more recent approaches integrate visual cues, such as facial gestures. The last decade has witnessed significant strides in pattern recognition, courtesy of deep neural networks. These breakthroughs extend beyond well-established domains like object recognition and speech analysis to hybrid problem-solving methods, such as audio-visual recognition. The emerging field of paralinguistics has also benefited from these advancements.[4] Creating an emotion recognition model solely based on one type of data has limitations because emotions are intricate and tied to numerous nonverbal signals. Recent studies are pivoting towards multimodal approaches, considering multiple types of data to build more comprehensive and accurate emotion recognition models. This shift acknowledges the interconnected and nuanced nature of human emotions, opening new avenues for exploration in the field of affective computing.

Facial Expression Recognition (FER) is a significant area within Artificial Intelligence (AI) that focuses on the accurate identification of facial expressions, which are essential for human-computer interaction [5].

Facial Expression Recognition (FER) systems can be used in many different areas like healthcare, education, checking who someone is, and making daily tasks easier. FER systems help in taking care of patients in a better way and it also help teachers to teach the students more effectively. In education, it allows teachers to understand how learners are feeling by observing their facial expressions, which helps in using better teaching techniques. FER is also applied in teaching robots to adjust their behavior in everyday situations.

Speech Emotion Recognition (SER) can be said as computers or machines detecting emotions from human speech. It like an empathetic machine listening to us. we can say that a machine is listening to us with empathy. Emotions are vital in human intelligence, decision-making, social interactions, perception, memory, learning, and creativity Speech emotion recognition has a wide range of practical application scenarios, such as depression diagnosis, call center, online classroom etc.

So in this paper We introduce an emotion recognition system that combines both auditory and visual modalities. Apart from the mere extraction of features, it is of paramount importance to adeptly manage unconventional data points and capture contextual information with efficacy. To address this challenge, we used Long Short-Term Memory (LSTM) for audio in which The model is trained using the ”fit” method ,where we input the features extracted by the MFCC and their corresponding target labels. When this model tested on TESS(Toronto emotional speech set) our audio model showed accuary of 98 percentage . We set aside 20 percentage % of the data for validation, which helps keep track of how well the model is doing during training. Convolutional Neural Networks (CNN) is used for our image and video-related data. We tested our propsed image model on different benchmark datasets like JAFFE, CK+48, and CFE individually. To show the benefit of our proposed image model we combined datasets JAFFE, CK+48, and CFE to form a new mixed dataset.

1.1 History

In the early 2000s, scientists concentrated their studies on understanding and identifying emotions through individual channels like facial expressions or speech. As we entered the mid-2000s, scientists started looking into combining different sources like facial expressions, speech, and physiological signals. The idea was to address the limitations of using just one type of information and develop more advanced systems that could better understand and interpret human emotions accurately. This shift aimed at creating more well-rounded approaches to emotion recognition by considering multiple aspects simultaneously This shift towards a multi-modal approach sought to enhance the overall effectiveness and reliability of emotion recognition technologies.

The advancement of multi-modal emotion recognition was significantly helped by having large datasets that included both audio and visual information Datasets like AVEC and IEMOCAP, which were well-labeled, provided researchers with valuable resources to train and assess their models using real-world data.

1.2 Existing Techniques

The foundation of FER research is built upon **Ekman's** work, which established six fundamental human expressions: anger, happiness, sadness, surprise, disgust, and fear [6].

Mehrabian and Russell's research has underscored the importance of facial expressions in human communication, suggesting that as much as 55 % of expressed information is transmitted through these expressions [7].

Currently, FER methods can be broadly categorized into two types: traditional Machine Learning and deep learning.

Traditional Machine Learning	**Deep Learning**
Doesn't include deep learning. These machine learning methods include: **Pre-processing**, **Feature Extraction**, and **Classification**.	Increasing network layers allows for better learning of object features. Deep neural networks efficiently extract facial features and recognize expressions with high accuracy.
Feature extraction from image is the key issue of image processing in all directions.The quality of the features can affect the classification of the model.	Deep neural network methods eliminate the need for manual feature extraction, relying on convolution operations for efficient feature extraction .

Table 1.1: Comparison of Machine Learning and Deep Learning

1.2.1 Deep Learning

Deep learning has found applications in various domains, including clinical care, immunization coverage and monitoring, medical image classification, and even COVID-19 diagnosis. However, Facial Expression Recognition (FER) remains a challenge due to the similarity in features of certain facial expressions, such as fear and surprise[8][9].

Factors like the angle of view and illumination conditions have a significant impact on FER accuracy[10]. Overcoming these challenges in FER is crucial for advancing human-computer interaction[11].

In the world of Facial Emotion Recognition (FER), both regular Machine Learning and advanced Deep Learning techniques follow a process that's kind of like how humans understand emotions from faces. First, there's the prep phase, where data gets ready. Then, it's like picking out important facial signals in the feature extraction step. Finally, in the classification phase, emotions are tagged based on those picked-up features. It's sort of like how we humans naturally figure out feelings by looking at someone's face. It's a journey that echoes the human intuition of perceiving emotions through facial expressions, bringing a touch of humanity to machine understanding.. The pre-processing stage employs techniques such as histogram equalization and homomorphic filtering, selected according to the dataset and available computational resources.

For feature extraction, methods vary based on the type of input data and may include approaches like geometric or statistical extraction (e.g., PCA, LBP).

1.2.2 Machine Learning

As we see that Machine learning methods brought significant advancements to Speech Emotion Recognition (SER), a technology that discerns emotions in spoken language. Semantic-based methods are crucial in SER, as they are a key category within speech signal processing.

Bimodal SER model that leverages both acoustic information related to speech sound and linguistic information related to words and their meanings for more accurate emotion recognition.

However, SER faces challenges when dealing with different training and test data from various sources, leading to variations in feature distributions that can negatively impact recognition performance. Emotions in speech can be complex, often involving the simultaneous experience or expression of multiple emotions, sometimes not aligning with actual feelings. This complexity presents challenges for cross-corpus SER.

To enhance the accuracy of classification, models such as Capsule Networks (CapsNet) are skilled at recognizing spatial relationships within essential speech details derived from spectrograms.

Stacked generalization is employed to create ensemble classifiers that combine predictions from both CapsNet and RNN classifiers, further improving SER performance.

1.2.3 Convolutional Neural Networks (CNNs)

Imagine Convolutional Neural Networks (CNNs) as these super-smart computer wizards that have a knack for looking at pictures and making sense of them. They're like the tech-savvy detectives in the digital world, helping us recognize and sort photos into different groups. CNNs are built to autonomously and flexibly acquire layered representations of visual information. In this section, we'll delve into the fundamental ideas behind CNNs and how they are pertinent to our emotion recognition model.

1.2.4 Long Short-Term Memory (LSTM)

The LSTM, It's like a smarter version with customized memory blocks that have special gates to control how information flows and connects within itself. The input, output, and forget gates in its structure allow for adaptive memory storage and retrieval. Thanks to a forget gate and peephole connections, the LSTM can smoothly handle continuous input streams without needing to break them into segments. This means it's excellent at capturing and using long-term contextual information through the repeated computation of a mapping from input to output sequences [12].

1.3 Evolution

1.3.1 Evolution of CNN for Facial Expression Recognition (FER)

Convolutional Neural Networks (CNNs)-based facial expression recognition (FER) has advanced significantly over time. An outline of the major advancements in CNN architectures for FER is given in this section.[13] Initially, face feature extraction and emotion classification were the main uses of conventional computer vision algorithms. These methods could not, however, automatically extract hierarchical features from raw pixel data. CNN introduction in FER The addition of CNNs to FER was the game-changer. After CNNs shown remarkable performance in image classification tasks, scientists started modifying these structures for facial expression analysis. For the purpose of identifying intricate patterns in facial photos, CNNs' capacity to automatically learn spatial hierarchies of features proved vital. [13] Deeper CNN architectures were investigated by researchers as computing power grew. Deeper networks—like VGGNet, GoogLeNet, and ResNet—have shown how beneficial it is to learn complex aspects from facial expressions. Facial landmarks, or important spots on the face, were added to CNN-based models to improve FER accuracy. The network was able to concentrate on important facial regions because to the additional spatial information these markers offered. Utilizing pre-trained CNN models on sizable datasets for transfer learning has become standard procedure in FER. With limited labeled data, models that were pre-trained on broad picture datasets (like ImageNet) were refined for facial expression detection tasks, resulting in better performance. The focus of current CNN-based FER research is on interpretability of model decisions, ensembling methods, and attention mechanisms. Overcoming dataset biases, enhancing robustness to position and illumination fluctuations, and creating models that can generalize to a variety of demographic groups are among the challenges. CNNs have evolved from classical techniques to deep learning architectures in the field of facial expression recognition. The goal of ongoing research is to improve CNNs for FER in practical applications by increasing their accuracy, efficiency, and interpretability.

1.3.2 LSTM in Audio

LSTMs revolutionized audio processing, initially excelling in speech recognition. Evolving for acoustic modeling, they advanced in multimodal emotion recognition. Recent trends focus on end-to-end learning from raw waveforms, addressing scalability with distributed training. Ongoing innovations aim to overcome challenges and enhance LSTM performance in diverse audio applications.[13]

Chapter 2

Literature Review

Seema Choudhary et al [14] proposed about Capsule Networks, which is an alternative to CNNs for computer vision. It fixes the limitations of CNNs in handling spatial relationships and recognizing emotions and is also effective for classifying and segmenting images with overlapping objects. Its dynamic routing mechanism enhances the feature representation. Capsule Networks are good at working with smaller sets of data because they're great at understanding how smaller parts fit into the bigger picture. CapsNet also comes with drawbacks primarily in terms of computational intensity.

Kunxia Wang et al [15] proposed about Capsule Networks, which is introduced for Facial Expression Recognition (FER) but It failed to pay attention to efficiency. To tackle this, the **Efficient-CapsNet** was introduced to balance efficiency and performance by using special types of convolution and batch normalization techniques. It also use a self-attention system to find out how to pass information through the network. To categorize different emotions, it uses a margin loss function. They tried out these Efficient-CapsNets on various datasets like JAFFE, CK+, and FER2013, each containing seven emotional categories.

B. Annappa et al [16] said that Facial Expression Recognition (FER) is important for teaching machines about emotions. Facial expressions are an essential part of human how we communicate without words. But it's not easy for computers, there are challenges like how people express emotions, things covering parts of the face, and different lighting. To understand these expressions, they are using the **Facial Action Coding System (FACS)** and **Action Units (AUs)**. When parts of the face are covered or someone's head is in different positions, it makes recognizing expressions difficult. They have tried different ways to teach computers this skill, using traditional machine learning methods and more advanced deep learning approaches.

Mohammed Abo-Zahhad et al [17] suggested that Capsule networks is good for simpler datasets such as MNIST (numbers) and recognizing emotions, but It struggled with more complicated sets like CIFAR10 (images). **DeepCaps** did better on CIFAR10, both CapsNets and DeepCaps are slower compared to the faster Convolutional Neural Networks (CNNs). To fix this, FECapsNet and Deep-FECapsNet were created to make Capsule Networks better. They made two changes: they figured out how information moves around and a new method to transform this information. These changes helped them to perform as well as made them faster and less complicated. **Specifically, they reduced the number of things the system had to learn by 58 % and made training take 64% less time for each round**.

Hasan Deeb et al [18] suggested a new FER system is designed for recognizing seven different emotions in faces. It does a few things to process the information—it gets the data ready by cleaning it up, find the face in the picture, and making it understandable for the computer. Then, it figures out important features using some fancy methods. After that, it sorts the emotions using a system called an Extreme Learning Machine with the Improved Black Hole algorithm. They tested this system on a few different sets of data with faces from Japan, Sweden, and a larger one called CK+. It is getting over 90% of the emotions right in all the tests. This new system is really good because it combines the strengths of different methods to make it accurate, less complicated, and better at understanding different types of faces. But, it has a drawback that it's a bit complicated to put into action, and it only works well when faces are looking straight ahead. This is a problem lots of systems have. When faces are at different angles, it's tough for computers to catch the emotions accurately.

Wenjuan Gong et al [19] discussed about Microexpressions, These are super quick facial expressions that happen so fast and our eyes usually miss them. With the help of Deep Learning algorithms, they are now trying to understand these microexpressions. Different methods for recognizing these microexpressions fall into five groups, based on what kind of info they look at. The few available datasets for microexpressions, like Polikovsky's, USFHD, SMIC, CASME, CASME II, and SAMM, are pretty small. Because these datasets are small and the expressions are not very strong, preparing the data before using it is a big task. How the data is handled really matters to make sure the recognition is accurate. People show many different expressions at once—these are called compound expressions. Datasets focusing on these could be super helpful for testing recognition systems.

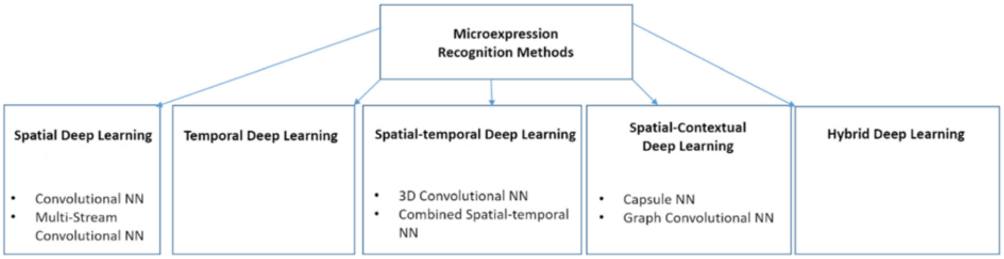
Microexpression Recognition Methods
Spatial Deep Learning
Convolutional NN
Multi-Stream Convolutional NN
Temporal Deep Learning
Spatial-temporal Deep Learning
3D Convolutional NN
Combined Spatial-temporal NN
Spatial-Contextual Deep Learning
Capsule NN
Graph Convolutional NN
Hybrid Deep Learning

Figure 2.1: Microexpression Recognition Methods

Gang Liu et al [20] discussed about the Understanding emotions in speech (**Speech Emotion Recognition**) is a big deal in studying how we get emotions through what people say. Some new models, like transformer-based and wav2vec2-based is really good at picking up personality traits and even transferring that learning to recognize emotions in speech. However, there aren't enough datasets and not enough info on how we perceive emotions through speech. This lack makes computers unable to do speech-emotion recognition accurately. To tackle this, there's a new idea that tries to mimic how humans naturally understand emotions and does multiple learning tasks at once. They've used a Convolutional Neural Network to sort the speech emotions into five categories that are anger, calm, anxiety, happiness, and sorrow.

Tian Y et al [21] discussed about the Facial expression recognition, which is a growing field that could change how people interact with machines. Deep convolutional neural networks are being used for facial expression recognition. Studies have compared how networks like AlexNet, GoogleNet, VGGNet, and ResNet perform. Machine learning and deep learning based methods are better than traditional methods for facial expression recognition, especially for finding both basic and complex emotions.

Wang Z et al [22] discussed about Emotion recognition, which is becoming important as we go for more intelligent and natural human-machine interaction. To recognize emotions from the human face, the geometry of facial organs is represented by cubic spline coefficients and facial texture is represented by Histogram of Gradients (HOG). Support Vector Machine (SVM) classifier is trained to recognize facial expressions using image data provided by CHEAVD 2.0 and a voice-based emotion classifier is trained using SVM with acoustic features extracted from the accompanying video voices. The recognition results from both classifiers are combined at the decision level using the Bayesian rule.

M Shamim Hossain et al [23] describes the method for emotion recognition that uses deep learning and fusion techniques. They first transform speech signals into Mel-spectrograms, which are treated as images. For video, key frames are extracted from the video and processed. Here, 2D CNN is used for speech and 3D CNN is used for video to extract features from both modalities. After that, fusion methods, including 'max,' 'product,' Bayesian sum rule, and an Extreme Learning Machine (ELM) based fusion, are used to compare their performance in emotion recognition. The ELM-based fusion combines features from both modalities in a non-linear manner which leads to improved accuracy. And, then the Support Vector Machine (SVM) is used as the final classifier to make emotion predictions based on the fused features.

Su Zhang et al [24] discusses a technical report for the ABAW5 challenge, which aims to recognize continuous valence-arousal using visual, audio, and linguistic information. They outline two multimodal models, each with four

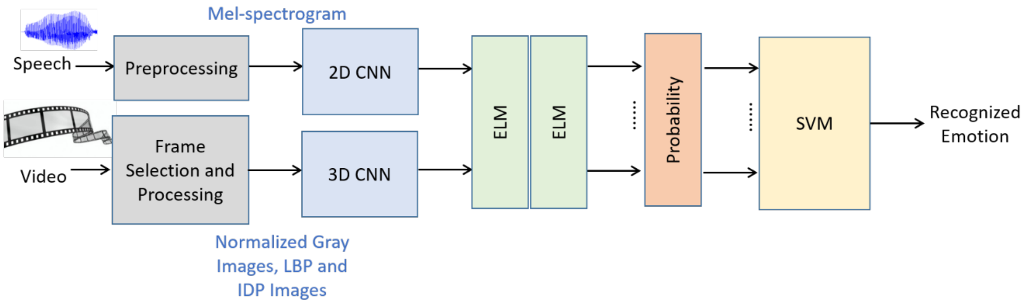

Figure 2.2: An overall block diagram of the proposed emotion recognition system

main blocks: visual, audio, linguistic, and co-attention. Each block is designed to extract features from its respective modality. The co-attention block fuses these extracted features. The models use the Aff-Wild2 database which includes data for valence-arousal estimation. After the processing is done, techniques such as CCC-centering, which is used to merge results from multiple folds of cross-validation, and clipping to ensure the predicted values fall within a specified range.

Zhongjie Li et al [25] proposed that multimodal fusion methods enhances the recognition accuracy and robustness of the model when combining the EEG and audio data. The proposed model architecture combines CNN for EEG data and BiLSTM for audio data. The evaluation metrics include accuracy and F1-score which are used to enhance the performance of the emotion recognition model. The multimodal fusion model surpasses the single-modal models and achieves good performance. Handling of EEG signals and trial signals as well as the conversion of EEG data from 1D to 2D data frames are important considerations in this field.

Yasser Alharbi et al [26] suggested that Voice emotion recognition is a challenging task. A single machine learning method can give acceptable results but not yet meet the desired result. Combining spectrogram analysis, which extracts low-level features, and deep learning has made it easier to recognize emotions in voice. Capsule Networks (CapsNet) and Recurrent Neural Networks (RNNs) are combined to develop an ensemble classifier model. CapsNet is used to identify spatial correlations of vital speech information in spectrograms using a pooling technique and RNNs are used for processing time-series datasets. Perceptual Linear Prediction Cepstral Coefficients (PNCC) are used for feature extraction and Capsule Networks (CapsNet) and Recurrent-Long Short-Term Memory (R-LSTM) are used for training. The results suggest that this approach is highly effective, with 96.05 overall accuracy on combined datasets eNTERFACE'05 and SAVEE, particularly in detecting the 'FEAR' emotion with an accuracy of 98%.

Anant Singh et al [27] proposed that the Transformer-based models have a significant impact on speech processing. However, there is limited research in evaluating transformer-based models for SER (speech emotion recognition) across multiple languages. The datasets used in such research exhibit variations in several aspects, including the number of utterances, the number of speakers, class distribution, and the number of emotion classes. Three speech models, wav2vec2, XLSR, and HUBERT, are used for feature extraction. The single-layer probing models consistently outperformed aggregation models in terms of accuracy and variability. It is found that the middle layers of speech representation models capture the most important features for SER. By extracting features from the optimal layer, It reports state-of-the-art results for German and Persian languages in SER. This suggests that selecting the appropriate layer is crucial for maximizing the effectiveness of speech representation models in SER tasks.

Panagiotis Tzirakis et al [28] proposed about the research on combining the auditory and visual modalities. Convolutional Neural Network (CNN) is used to extract features from speech data for the auditory modality and Deep Residual Network (ResNet) with 50 layers for the visual modality. The outputs are fused and passed to a Long Short-Term Memory (LSTM) to handle the problem of being insensitive to outliers and modeling context effectively. It is trained in an end-to-end fashion by taking advantage of correlations between the auditory and visual data. The database used is Remote Collaborative and Affective (RECOLA) database. The proposed model achieves significantly better performance compared to other models when tested with the RECOLA database.

Hui Ma et al [29] proposed that Emotion Recognition in Conversations (ERC) focuses on extracting human emotions from conversations or dialogues having two or more interlocutors. It has potential applications in opinion mining, healthcare, generating emotion-aware dialogues, and more. It also focuses on context and speaker-sensitive dependencies within conversations. Many previous methods focus on textual conversations but recent research has extended this to audio and video cues. A transformer-based model called ”self-distillation (SDT)” includes intra- and inter-modal transformers is used to capture interactions and uses a hierarchical gated fusion strategy to learn the different modalities dynamically. The model is implemented using PyTorch and datasets that are used are IEMOCAP dataset, MELD dataset, and others. ERC has shown better results compared to the existing approaches. Transformer-based fusion methods can be computationally expensive due to the self-attention mechanism.

Chapter 3

Problem Statement

Building a Multi-Model emotion recognition from audio and visual data using the deep-learning.

In the realm of human-computer interaction, affective computing, and diverse applications spanning from virtual assistants to mental health monitoring, the accurate detection of emotions from both auditory and visual data presents a genuine challenge. Thus far, the predominant focus has been on emotion recognition (EMR) utilizing a singular model, whereby the techniques concentrate on capturing either the auditory qualities of speech or the facial expressions exhibited by individuals. Nevertheless, it is worth noting that human emotions are intricate and multifaceted, thereby rendering these unidimensional approaches are difficult. The primary drawback associated with singular-model EMR lies in its inherent limitation of providing a comprehensive depiction of an individual's emotional state. Emotions encompass a complex interplay of explicit and implicit cues, necessitating a more comprehensive approach. While audio-based EMR can discern the subtleties in speech intonation, it may inadvertently overlook the minute facial expressions that convey an entirely different narrative. Conversely, visual-based EMR may prove inadequate in discerning the nuances of vocal inflections. Recognizing the limitations of sticking to just one approach in emotion recognition (EMR), there's a clear call for systems that bring together both audio and visual information. These multi model concepts aims to give more accurate picture of emotional states. The idea here is that audio and visual cues complement each other, and by merging them, we can grasp a deeper understanding of emotions – catching both the obvious and the subtle aspects of emotional expression. This research zeroes in on three main goals. First off, it wants to whip up an AI model that can nail down emotions from both audio and visual data. Secondly, it's out to fix the gaps left by single-model EMR by tapping into the teamwork of audio and visual modalities. And thirdly, the aim is to boost the accuracy and resilience of emotion recognition by smartly blending information from various sources. Now, why do we need all this?. Well, first and foremost, this research isn't just about upping the accuracy game; it's about diving deep into the intricate world of human emotions. This heightened accuracy has some promising implications, especially for applications that need a finely tuned understanding of how we express ourselves emotionally. Plus, it's not just a techy experiment – the impact is real and spans across fields like human-computer interaction, virtual reality, and keeping tabs on mental health. Lastly, by exploring the potential of multi-modal EMR, this research isn't just about improving the status quo; it's pushing the envelope in affective computing, breaking new ground in emotion recognition technology.

Chapter 4

Architecture

This chapter provides a thorough explanation of our multi-modal emotional identification architecture, which uses picture, video, and audio data to identify people's emotional states. Three main models serve as the foundation for the overall architecture: the audio, video, and image models. After that, these separate models are combined into a single multi-modal model.

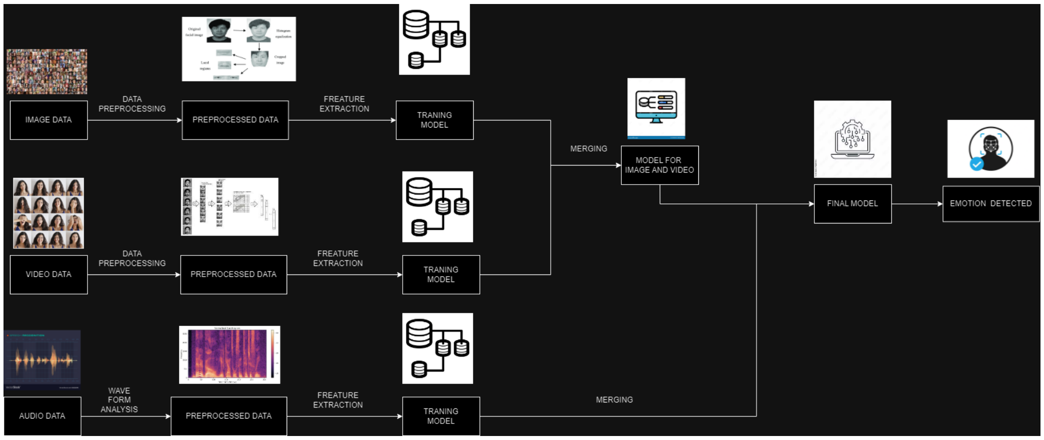

Figure 4.1: Architecture of multi-modal emotion recognition system

Overview of Each Individual Model

1.Image Model:We start the journey with the picture model, where we carefully prepare the input data to guarantee the best possible compatibility with our next stages of feature extraction. Following pre-processing, we extract features from the visual data by selecting important features. After then, the model's parameters are adjusted using a labeled dataset during training so that it can correctly classify incoming data into seven distinct emotional categories.

2. Audio Model: To improve data quality, the audio model, like the picture model, goes through a rigorous pre-processing stage. After that, important audio features required for emotion recognition are isolated by feature extraction. Using a labeled dataset as training data, the model learns to categorize audio inputs into the specified emotional classes. Its effectiveness is then thoroughly tested using a specific test dataset.

3. Video Model: The image and audio models' workflows are replicated in the video model. Preprocessing makes the input data more optimal, and feature extraction finds patterns that are significant in the visual data. The model is able to categorize video data into the specified emotional categories after being trained on a labeled dataset. Extensive testing verifies its functionality with omitted data.

Multi-Modal Fusion:The picture and video models are combined to create an integrated visual model when each model has finished training and testing. More training is applied to this combination model to improve its comprehension of multi-modal visual data. Testing later on guarantees that the fusion process doesn't impair the model's overall performance.

Multi-Modal Integration with Audio:The multi-modal visual model and the audio model are combined in the last stage to create a multi-modal emotional recognition system. An extensive dataset with inputs in the form of audio, video, and images is used to train an all-encompassing model. Its ability to precisely identify and categorize emotions in a variety of modalities has been extensively tested.

Chapter 5

Proposed Methodology

In this part, we'll dive into how we approached the issue and take a closer look at the models we applied for both images and videos.

5.1 Proposed Architecture

5.1.1 Image Model

We're diving into an exciting research project aimed at creating a powerful emotion recognition system. These CNN networks will be the backbone of our approach to thoroughly analyze facial expressions and neatly classify them into predefined emotion categories. It's all about making technology truly understand and interpret human emotions in a robust and meaningful way!

Dataset

The set of data used for both training and assessment includes a varied assortment of facial pictures sorted into seven different emotions: anger, disgust, fear, happiness, neutrality, sadness, and surprise. Beforehand, the dataset undergoes processing to ensure consistency in image sizes, improving the model's capability to adapt across various facial expressions.

Data Preprocessing

Image Loading and Conversion

Images are loaded from the dataset, initially in BGR format, and subsequently converted to RGB to ensure consistency in color representation.

Image Resizing

To facilitate efficient model training, all images are resized to a consistent target size of (48, 48) pixels.

Data Normalization

Normalizing the pixel values in the picture data to fall between 0 and 1 improves convergence when training.

Data Augmentation

We're spicing things up by artificially expanding our dataset using a technique called data augmentation. This not only beefs up our dataset but also boosts our model's knack for generalization. Thanks to the Keras ImageDataGenerator, we're throwing in rotations, zooms, and flips—both horizontally and vertically—to make sure our model gets a taste of diverse examples. It's like giving our model a broader perspective, helping it learn and adapt better.

Convolutional Neural Network Architecture

The proposed CNN architecture is designed to extract hierarchical features from facial expressions for subsequent emotion classification. The architecture consists of:

- An initial Conv2D layer with 64 filters and ReLU activation.
- MaxPooling2D layers to downsample the spatial dimensions.
- Additional Conv2D layers with increasing filter sizes (128 and 256) and accompanying MaxPooling2D layers.
- A Flatten layer to convert 3D feature maps into a 1D vector.
- Dense layers with 512 neurons and ReLU activation.
- A final Dense layer with 7 neurons for the output, representing the seven emotion classes, and a softmax activation for multi-class classification.

Model Compilation and Training

We kick things off by compiling the model, using the Adam optimizer and categorical crossentropy as our go-to loss function. The training process unfolds on the prepped and augmented dataset, embracing a batch size of 32 over 15 epochs.

Evaluation

Once the model is well-trained, we put it through its paces on a separate test set. We're not just stopping at accuracy; we're crunching numbers on various metrics. And for a visual treat, we've got nifty charts showcasing the model's learning journey, both in training and validation.

Conclusion

Our grand plan is to take emotion recognition to the next level. How? By blending a carefully curated dataset with a top-notch convolutional neural network. The upcoming sections will spill the beans on experimental results, discussions, and how we stack up against the competition. Stick around!

5.1.2 Audio Model

Our audio model is the maestro of emotion classification, specifically crafted for deciphering feelings from audio data using MFCC features. Here's the breakdown of its architectural symphony:

LSTM Layer: At the heart of the model, we've got an LSTM layer flaunting 123 memory units and a linear activation function. This layer is the maestro, conducting the orchestra of temporal dependencies in sequential MFCC data. It's designed to output a singular prediction for each sequence, with an input shape of (40, 1), indicating its prowess in handling sequences with 40 time steps and one feature each.

Dense Layer 1: This layer, with 64 units and a ReLU activation function, adds a dash of non-linearity to the mix. By turning negative values into zeros, it tunes the model to catch subtle patterns, enhancing its knack for capturing nuanced features.

Dropout Layer 1: With a dropout rate of 0.2, this layer plays bodyguard, preventing overfitting by randomly benching 20% of input units during training. This touch of randomness enhances the model's street smarts, making it more adaptable and robust.

Dense Layer 2: Featuring 32 units and a ReLU activation function, the second dense layer builds upon the wisdom gained from the previous layer. The ReLU activation injects more non-linearity, helping the model dive even deeper into the intricate patterns within the data.

Dropout Layer 2: Marching in line with its predecessor, this dropout layer (with a rate of 0.2) introduces a calculated dose of randomness during training. Think of it as the model's workout routine, keeping it agile and in top shape for real-world challenges.

Output Layer: Tailored for the seven emotion classes, this layer boasts 7 units and a softmax activation function. It's the storyteller, crafting a probability distribution. The softmax function transforms raw output into probabilities, making it a perfect fit for multi-class classification adventures.

To tie it all together, we compile the model with categorical cross-entropy loss and the Adam optimizer. The model summary spills the beans on each layer's configuration, output shape, and the number of parameters in this symphony of emotion decoding.

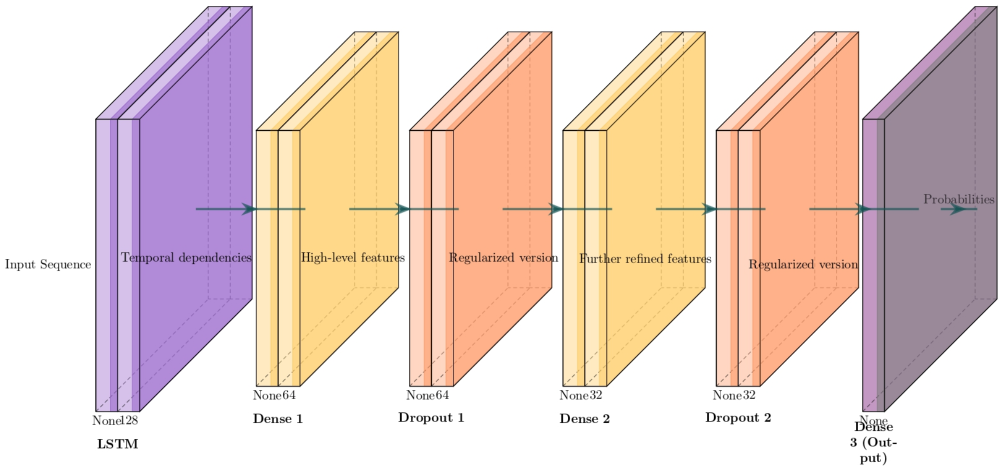

Figure 5.1: Architecture of audio

5.2 Model Training

5.2.1 Image Model

In the Python script provided, the Convolutional Neural Network (CNN) constructed using the Keras API follows a sequential architecture. The model kicks off with a MaxPooling layer for downsampling, paving the way for a convolutional layer with 64 filters of size (3,3) and a dash of Rectified Linear Unit (ReLU) activation. The subsequent layers consist of more convolutional and max-pooling magic, with 128 and 256 filters taking the stage in each duo. Following this, the network guides the one-dimensional vector crafted by the convolutional layers through a duo of dense layers. The grand finale is a dense layer employing a softmax activation for multi-class classification, boasting 7 units representing the emotion classes. The first dense layer is a powerhouse

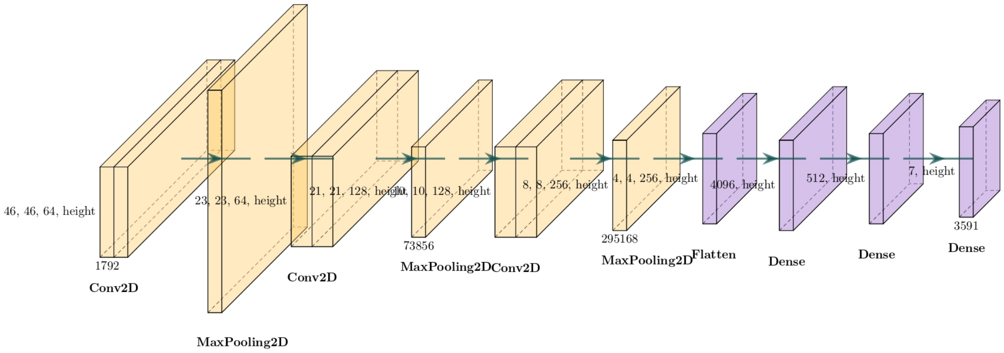

Figure 5.2: Architecture of audio

with 512 units, fuelled by the ReLU activation function.

As part of the methodology, images are loaded and gracefully preprocessed from various emotion files. They undergo a resizing ceremony, transforming into a desirable size of (88, 88) pixels. To align with the RGB vibe, the images gracefully switch from BGR to RGB. Adding a touch of flair, Keras' ImageDataGenerator steps in, introducing data augmentation like rotation, zooming, shifting, and flipping—a symphony to enhance generalization. The dataset plays a key role, with 80

For the model's report card, accuracy takes center stage as the evaluation metric. The model gets its act together using the Adam optimizer and dances to the categorical cross-entropy loss function. Training for a solid 15 epochs, with a backstage crew of 32 in each batch, the model then faces the music with a test set assessment. To add a visual touch, Matplotlib makes a cameo, showcasing the accuracy and loss spectacle during both training and validation. It's not just code; it's a script for a dazzling performance!

5.2.2 Audio Model

The model is trained using the ”fit” method , where we input the feature extracted by the mfcc and corresponding target labels.A validation split of 20 percentage is employed, dedicating a portion of the training data to monitor the model’s performance during training. The training process unfolds over 100 epochs, with each epoch processing the entire dataset with batch size of 512.For randomly reordered data before epoch we use shuffle parameter.

5.3 Dataset

5.3.1 Toronto emotional speech set (TESS)

The TESS dataset is a unique resource comprising high-quality audio recordings featuring two female actresses (aged 26 and 64) uttering 200 target words in the carrier phrase ”Say the word ....” across seven distinct emotions. This dataset, unlike many others, exclusively represents female speakers, providing a valuable counterbalance to the prevalent skew towards male speakers in emotion classification datasets. With a total of 2800 audio files, each emotion and actress is meticulously organized within dedicated folders, facilitating seamless exploration. The TESS dataset, recorded in WAV format, stands out for its potential to enhance emotion classifier generalization due to its balanced representation of female voices and nuanced emotional expressions.

5.3.2 Jaffe Dataset

Data Preprocessing: At the heart of our endeavors [30] lies the Jaffe dataset—an emotive treasure trove featuring anger, disgust, fear, happiness, neutral, sadness, and surprise. Guided by the grayscale brush, images undergo a transformation, resizing gracefully to a standardized 48 by 48 resolution. The harmony of data representation is achieved through a meticulous normalization to the [0, 1] range.

Convolutional Neural Network Structure: Enter the protagonist of our narrative—the Jaffe model [31] powered by a Convolutional Neural Network (CNN). The plot unfolds with initial layers orchestrating the ballet of max pooling and convolution to extract intricate features. As the crescendo builds, a dense layer takes center stage. The final act unfolds as the model, guided by the enchanting softmax activation, categorizes expressions into seven types.

Model Training and Evaluation: The crafting of our model is a tale of optimization and loss. An Adam optimizer and the sparse categorical cross-entropy loss weave the fabric of our creation, which takes center stage through 20 epochs. The performance, measured by the metrics of recall, accuracy, precision, and the revealing confusion matrix, is scrutinized in the spotlight of evaluation measures. It's not just a model; it's a performance under the critical gaze of thorough assessment.

5.3.3 CK+48 Dataset

Data Preprocessing: Emotions translated to numerical values for CK+48 include surprise, sadness, fear, happiness, disgust, contempt, and anger [32].

Photos are prepared for the model by resizing and normalizing them.

CNN Structure Regarding CK+48: The CK+48 model ends with dense layers after two convolutional layers with max pooling and dropout. Accuracy and loss charts are used to visualize training over ten epochs.

Model Training and Evaluation: The accuracy of the CK+48 model is confirmed by evaluation [32]. To evaluate its ability to forecast in real-world scenarios, external images are supplied.

Prediction on External and Random Test photos: The model is tested on random and external photos to demonstrate its practical usefulness. True labels are compared with predicted emotions.

5.3.4 CFE Dataset

Data Preprocessing: According to Merghani et al. (2019), preprocessing in CFE entails loading, labeling, resizing, and normalization.

Transfer Learning with VGG16: According to [33], the CFE dataset uses VGG16 for transfer learning. Custom dense layer structure and frozen pretrained layers are used.

CNN Model for CFE: To improve variety and generalization, combine an alternate CNN model with generator-based data augmentation.

Assessment and Prognosis: Both models are assessed, and their adaptability is tested using external imagery.

Since the three aforementioned datasets did not produce the anticipated results, we manually combined them and developed a new model for the mixed dataset, which consists of all the samples from the three aforementioned datasets.

Chapter 6

Results and Discussion

6.1 Image Model

We took our neural network through a rollercoaster of emotions while training. We covered 7 emotions in the dataset. But we didn't settle for the basics. To give our model a broader perspective, we added some spice during training. We played with the data, throwing in rotations, zooming, and flips—both horizontal and vertical. It's all about boosting those generalization skills.

Training Metrics

We guided the model through 15 epochs, employing a batch size of 32. Throughout this training journey, we vigilantly monitored the accuracies and losses in both the training and validation stages.

```
Epoch 15/15
```

```
69/69 [==============================] - 29s 392ms/step - loss: 0.7801 -
accuracy: 0.7645 - val_loss: 1.0621 - val_accuracy: 0.6691
```

Evaluation on Test Set

We tested the model on a test set it hadn't seen during training to evaluate how well it generalizes.

```
Test accuracy: 66.91%
```

6.1.1 Discussion

Model Performance

The achieved test accuracy of 66.91% suggests that the model has learned to classify emotions effectively. The utilization of data augmentation techniques during training has probably played a crucial role in enabling the model to adeptly handle diverse facial expressions.

Overfitting Considerations

While the training accuracy reached 76.45%, a slight performance drop was observed on the test set. Further investigation into potential overfitting is warranted. Fine-tuning the model complexity or employing additional regularization techniques may be explored to improve generalization.

Class-wise Performance

An analysis of class-wise performance may provide insights into the model's strengths and weaknesses across different emotions. Further investigations

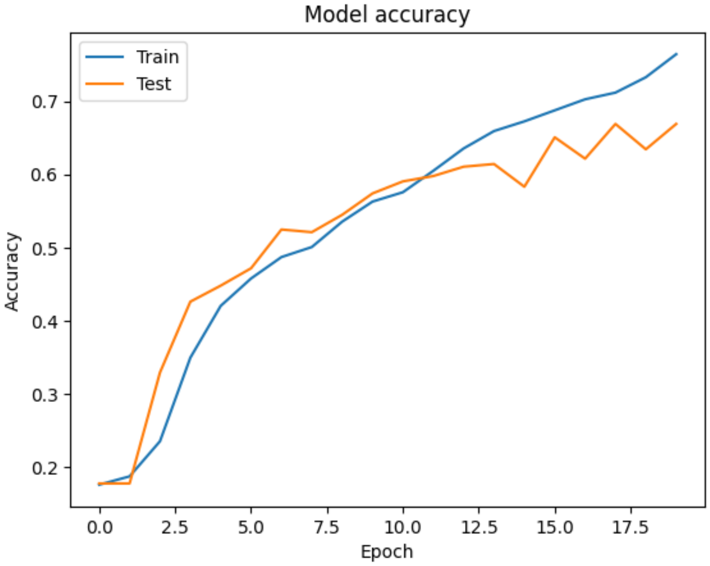

Figure 6.1: Training and Validation Accuracy

into misclassifications, confusion matrices, and class-wise precision-recall curves could reveal areas for improvement.

Training Dynamics

Examining the training dynamics is essential to evaluate the model's convergence speed, determine the suitability of the chosen learning rate, and assess whether additional training epochs could result in enhanced performance.

6.1.2 Conclusion

Wrapping up, the applied CNN model showcases promising outcomes in classifying mixed emotions. Subsequent actions could involve refining the model architecture, optimizing hyperparameters, and delving into a comprehensive analysis of misclassifications to elevate overall performance.

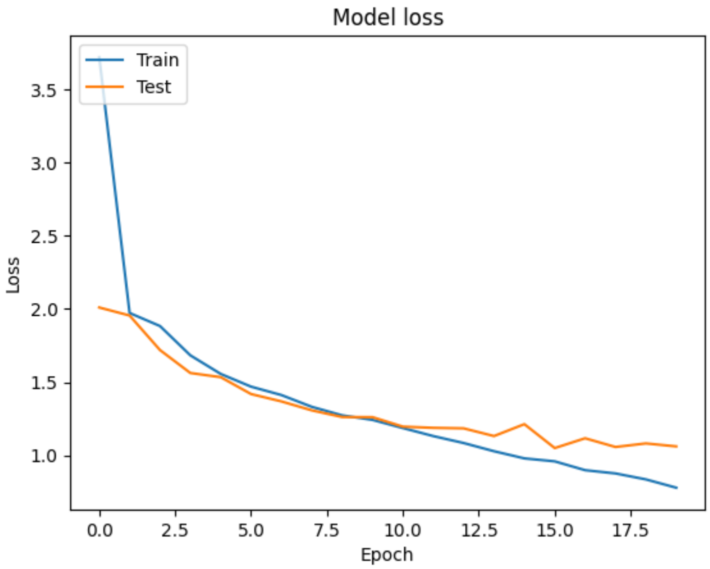

Figure 6.2: Training and Validation Loss

6.2 Audio Model

The overall accuracy on the test set reached an impressive 98 percent, affirming the efficacy of our approach. To sum up, the domain of emotion recognition is progressing rapidly, fueled by an increasing enthusiasm for promoting intelligent and natural human-machine interactions These approaches encompass a diverse set of modalities, including facial expressions, human voices, and even their combination. However, the continual progress in this field promises a future where machines can intuitively understand and respond to human emotions, opening up new possibilities for human-computer interaction. For the next phase, we are planning to complete the 1st modal, which is emotion recognition from the image with efficient methods for pre-processing, feature extraction and classification.

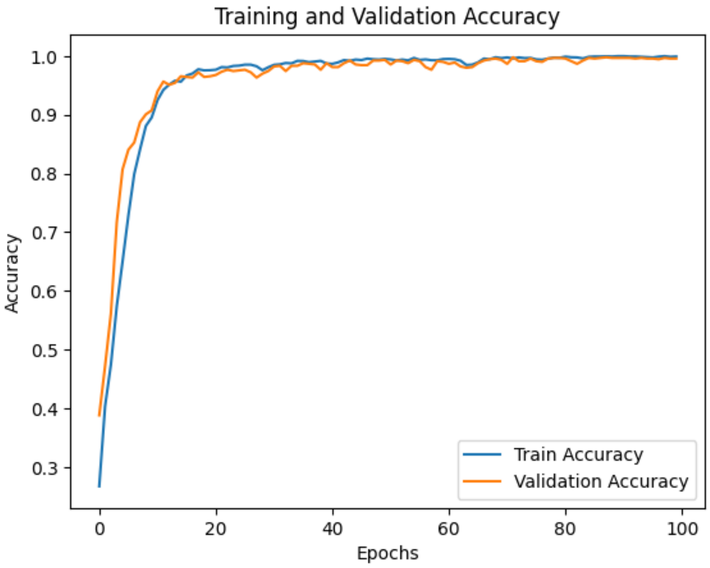

Figure 6.3: Training and Validation accuracy of audio model

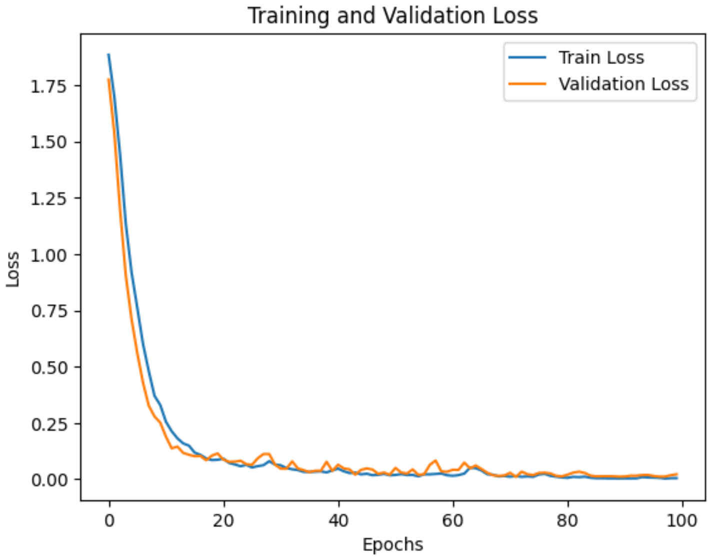

Figure 6.4: Training and Validation Loss of audio model

Chapter 7

Future work

The current research is grappling with the challenge of finding the right dataset to nail down multi-modal emotion recognition. There's a push to put together a dataset using video snippets, a crucial move to fill in the gaps. Looking ahead, the plan is to beef up this dataset, making sure it covers a wide range of emotions, various demographics, and different environmental conditions. The goal? To make the emotion recognition model sturdy and capable of handling a variety of situations. The models we've cooked up so far have been trained on the videos we gathered, tweaking the settings to find the sweet spot. We're even thinking of trying out different neural network setups and training methods to see what works best. In the future, we're planning to dig deep into the system's effectiveness by comparing it against existing benchmarks and the latest and greatest approaches out there. we're not stopping there. We'll be stacking our system up against established datasets and industry standards to make sure it's top-notch.

Bibliography

[1] Ji, Q., Zhu, Z., Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. IEEE transactions on vehicular technology, 53(4), 1052-1068.

[2] Burkhardt, Felix, Jitendra Ajmera, Roman Englert, Joachim Stegmann, and Winslow Burleson. ”Detecting anger in automated voice portal dialogs.” In INTERSPEECH. 2006.

[3] Anagnostopoulos, C. N., Iliou, T., Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review, 43, 155-177.

[4] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8), 1301-1309.

[5] Guo X, Zhang Y, Lu S, Lu Z. Facial expression recognition: a review. Multimedia Tools and Applications. 2023 Aug 17:1-47.

[6] Hong, J., Lee, H. J., Kim, Y., Ro, Y. M. (2020). Face tells detailed expression: Generating comprehensive facial expression sentence through facial action units. In MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26 (pp. 100-111). Springer International Publishing.

[7] Mehrabian, A., Russell, J. A. (1974). An approach to environmental psychology. the MIT Press.

[8] He, Y., Chen, S. (2020). Person-independent facial expression recognition based on improved local binary pattern and higher-order singular value decomposition. IEEE Access, 8, 190184-190193.

[9] Khatri, N. N., Shah, Z. H., Patel, S. A. (2014). Facial expression recognition: A survey. International Journal of Computer Science and Information Technologies (IJCSIT), 5(1), 149-152.

[10] Peng, X., Yu, X., Sohn, K., Metaxas, D. N., Chandraker, M. (2017). Reconstruction-based disentanglement for pose-invariant face recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1623-1632).

[11] Han, Z., Huang, H. (2021). Gan based three-stage-training algorithm for multi-view facial expression recognition. Neural Processing Letters, 53, 4189-4205.

[12] Van Houdt, G., Mosquera, C., Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53, 5929-5955.

[13] Borgalli, M. R. A., Surve, S. (2022). Deep learning for facial emotion recognition using custom CNN architecture. In Journal of Physics: Conference Series (Vol. 2236, No. 1, p. 012004). IOP Publishing. for rest

[14] Choudhary, S., Saurav, S., Saini, R., Singh, S. (2023). Capsule networks for computer vision applications: a comprehensive review. Applied Intelligence, 1-28.

[15] Wang, K., He, R., Wang, S., Liu, L., Yamauchi, T. (2023). The Efficient-CapsNet model for facial expression recognition. Applied Intelligence, 53(13), 16367-16380.

[16] Adyapady, R. R., Annappa, B. (2023). A comprehensive review of facial expression recognition techniques. Multimedia Systems, 29(1), 73-103.

[17] Abo-Zahhad, M., Eldifrawi, I., Abdelwahab, M., El-Malek, A. H. A. (2023). Deep and shallow fast embedded capsule networks: going faster with capsules. Analog Integrated Circuits and Signal Processing, 114(3), 315-324.

[18] Deeb, H., Sarangi, A., Mishra, D., Sarangi, S. K. (2022). Human facial emotion recognition using improved black hole based extreme learning machine. Multimedia Tools and Applications, 81(17), 24529-24552.

[19] Gong, W., An, Z., Elfiky, N. M. (2022). Deep learning-based microexpression recognition: a survey. Neural Computing and Applications, 34(12), 9537-9560.

[20] Liu, G., Cai, S., Wang, C. (2023). Speech emotion recognition based on emotion perception. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1), 22.

[21] Tian, Y., Kanade, T., Cohn, J. F. (2011). Facial expression recognition. Handbook of face recognition, 487-519.

[22] Xu, F., Wang, Z. (2018, October). Emotion recognition research based on integration of facial expression and voice. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

[23] Hossain, M. S., Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49, 69-78.

[24] Zhang, S., An, R., Ding, Y., Guan, C. (2022). Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2376-2381).

[25] Li, Z., Zhang, G., Dang, J., Wang, L., Wei, J. (2021). Multi-modal emotion recognition based on deep learning of EEG and audio signals. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

[26] Alharbi, Y. (2022). Effective ensembling classification strategy for voice and emotion recognition. International Journal of System Assurance Engineering and Management, 1-12.

[27] Singh, A., Gupta, A. (2023). Decoding Emotions: A comprehensive Multilingual Study of Speech Models for Speech Emotion Recognition. arXiv preprint arXiv:2308.08713.

[28] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8), 1301-1309.

[29] Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., Xu, B. (2023). A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations. IEEE Transactions on Multimedia.

[30] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In Proceedings Third IEEE international conference on automatic face and gesture recognition (pp. 200-205). IEEE.

[31] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[32] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society conference on computer vision and pattern recognition-workshops (pp. 94-101). IEEE.

[33] Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.