Real-time Facial Emotion Recognition Web Application

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Computer Science and Engineering
by Research

by

Palash Agarwal
20161125
palash.agarwal@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology,
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Real-time Facial Emotion Recognition Web Application"** by **Palash Agarwal**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____

Date

Advisor: Prof. Prakash Yalla

To my family
and friends

Acknowledgements

I would like to take this opportunity to extend my sincere appreciation to all those who not only aided me in completing my research thesis but also contributed to my growth as a researcher.

First and foremost, I want to convey my deepest thanks to my advisor, Prof. Prakash Yalla, for his consistent support throughout the years. His guidance played a pivotal role in refining my research skills, and I can clearly see the progress I've made under his mentorship. This thesis owes its existence to his steadfast dedication to excellence.

Special thanks are due to Khoushik Ananth, who not only imparted crucial domain knowledge but also readily assisted with any inquiries and collaborative brainstorming.

I would also like to express my gratitude to Prof. Sushmita Banerji for her continuous support throughout my research journey.

My heartfelt thanks go to my family, a constant source of inspiration and encouragement throughout my research. Without their unwavering love and support, completing this research would not have been possible. I also want to acknowledge all my friends who made my college experience memorable and served as a constant source of motivation.

Abstract

This thesis addresses the challenge of real-time facial emotion recognition through the lens of Deep Convolutional Neural Networks (DCNNs). While DCNNs have proven to be state-of-the-art for image classification, their efficacy is contingent upon abundant, accurately labeled data. Emotion recognition, a complex and subjective task, suffers from a scarcity of comprehensive datasets. Existing facial expression databases lack the depth required to train deep neural networks effectively, particularly those that have excelled in object recognition tasks.

The limitations are further compounded by high inter-subject variations, including age, gender, ethnic backgrounds, and individual expressiveness levels. Facial images are also subject to occlusions, pose variations, and wearable accessories, intensifying the complexity of the emotion recognition problem. Consequently, the choice of a benchmark dataset plays a pivotal role in the development and evaluation of an effective model.

In this research, we explore the nuances of real-time facial emotion recognition, delving into the challenges posed by limited and diverse datasets. Our approach involves a deep investigation into the intricacies of emotion representation in facial images, leveraging the power of DCNNs. By addressing the dataset limitations and exploring techniques to mitigate the impact of diverse personal attributes, we aim to contribute to the advancement of facial emotion recognition models. The ultimate goal is the development of a real-time facial emotion recognition web application, demonstrating the practical application of our findings in a user-friendly and accessible manner.

Contents

Table of Contents

Introduction 9

1.1 Background 9

1.2 Importance of Facial Emotion Recognition 9

1.3 Challenges in Facial Emotion Recognition 10

1.4 Evolution of Facial Emotion Recognition 11

1.5 Objectives of This Research 11

1.6 Significance of the Study 12

Literature Review 13

2.1 Introduction 13

2.2 Early Approaches to Facial Emotion Recognition 13

2.2.1 Hand-crafted Features 13

2.2.2 Classical Machine Learning Methods 14

2.2.3 Deep Learning Methods 14

2.2.4 Graph Convolutional Networks (GCNs) 16

2.2.5 Multimodal Emotion Recognition 17

2.2.6 Transfer Learning 17

2.3 Datasets 17

2.3.1 FER2013 Dataset 18

2.3.2 Expression in-the-Wild (ExpW) Dataset 18

2.3.3 Aff-Wild2 Dataset 18

Methodology 20

3.1 Introduction 20

3.2 Dataset Selection 20

3.3 Data Preprocessing 21

3.3.1 Face Detection 21

3.3.2 Face Alignment 22

3.4 Evaluation 23

3.5 Web Application Development 23

Implementation and Results 24

4.1 Introduction 24

4.2 Classical Machine Learning Method 24

4.2.1 Local Binary Pattern Features 25
4.2.2 Dense SIFT + CNN 27
4.2.3 Face Landmarks + HOG using SVM classifier 27

4.3 Deep Learning Method 29
4.3.1 Proposal 29
4.3.2 Implementation 32

4.4 Graph Convolution Networks for Emotion Recognition 35
4.4.1 Dataset Preparation 35
4.4.2 Implementation 36

4.5 Discussion and Conclusion 37
4.5.1 Summary of Key findings 37
4.5.2 Implications of the Results 38

Web Application Development 39

5.1 Introduction 39

5.2 Architecture 39
5.2.1 Upload/Record a Video 40
5.2.2 Facial Emotion Recognition Model 40
5.2.3 Dashboard 40

5.3 Practical Applications 42
5.3.1 Customer Service 42
5.3.2 Mental Health Monitoring 42
5.3.3 Education 43
5.3.4 Human-Computer Interaction 43
5.3.5 Security and Surveillance 43
5.3.6 Market Research and Advertising 43
5.3.7 Healthcare 44
5.3.8 Entertainment Industry 44

5.4 Conclusion 44

Conclusions 45

6.1 Conclusion 45

6.2 Future Work 46
6.2.1 User Interface Refinement 46
6.2.2 Enhanced Model Robustness 47
6.2.3 Multimodal Emotion Recognition 47
6.2.4 Usability Studies and User Feedback Analysis 47
6.2.5 Scalability and Deployment Optimization 47

Chapter 1

Introduction

1.1 Background

Facial emotion recognition (FER) stands at the intersection of computer vision, affective computing, and artificial intelligence. It involves the automatic identification of human emotions from facial expressions captured in images or videos. This capability has significant implications for a broad range of applications, including human-computer interaction, mental health assessment, security systems, and entertainment. The ability to discern and interpret emotions accurately is a quintessential aspect of human communication, and replicating this capability in machines has been a longstanding pursuit in the field of artificial intelligence [1].

Human facial expressions are a rich source of emotional information. According to the work of psychologist Paul Ekman, there are six basic emotions universally recognized across different cultures: happiness, sadness, fear, disgust, surprise, and anger [2]. These basic emotions, along with neutral expressions, form the foundation for most facial emotion recognition systems.

1.2 Importance of Facial Emotion Recognition

The importance of FER can be seen across various domains:

1. **Human-Computer Interaction (HCI)**: FER can enhance the interaction between humans and computers by enabling machines to understand and respond to the emotional states of users. This leads to more intuitive and natural interfaces, improving user experience and satisfaction [3].
2. **Mental Health Assessment**: FER can be used to monitor and assess the emotional well-being of individuals. For instance, it can help detect signs of depression, anxiety, and other mental health conditions by analyzing facial expressions over time [3].

3. **Security Systems**: In security and surveillance, FER can be employed to identify suspicious behavior or stress indicators, aiding in the prevention of criminal activities [3].
4. **Education and Training**: In educational settings, FER can be used to gauge student engagement and emotions during learning, allowing educators to tailor their teaching strategies accordingly [3].
5. **Entertainment and Gaming**: FER can enhance the gaming experience by enabling characters to respond to the emotions of players, creating a more immersive and interactive environment [3].
6. **Customer Service**: In customer service, FER can help in understanding customer emotions and improving the quality of service by providing real-time feedback to customer service representatives [3].

1.3 Challenges in Facial Emotion Recognition

Despite its potential, FER presents several challenges:

1. **Variability in Facial Expressions**: Facial expressions can vary significantly between individuals due to differences in age, gender, ethnicity, and personal expressiveness. This variability makes it challenging to develop models that generalize well across diverse populations [4].
2. **Occlusions and Head Poses**: Real-world scenarios often involve occlusions (e.g., glasses, beards, masks) and varying head poses, which can obscure facial features and complicate the recognition process [4].
3. **Lighting Conditions**: Variations in lighting can affect the appearance of facial features, making it difficult for FER systems to maintain consistent performance [4].
4. **Subtlety of Emotions**: Some emotions, such as fear and surprise, can be very subtle and difficult to distinguish from each other. This requires highly sensitive and accurate models [4].
5. **Data Scarcity**: High-quality annotated datasets are crucial for training FER models. However, obtaining large, diverse, and accurately labeled datasets is a challenging and resource-intensive task [5].

1.4 Evolution of Facial Emotion Recognition

The evolution of FER can be broadly categorized into three phases:

1. **Early Approaches**: Initial methods relied on hand-crafted features and classical machine learning techniques. Features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) were commonly used to represent facial textures and gradients. These features were then fed into classifiers like Support Vector Machines (SVM) and Random Forests [6, 7].
2. **Deep Learning Era**: The advent of deep learning revolutionized FER. Deep Convolutional Neural Networks (DCNNs) demonstrated remarkable success in image classification tasks by automatically learning hierarchical feature representations from raw data. Prominent architectures like VGG, ResNet, and Inception have been widely adopted for FER tasks, significantly improving accuracy and robustness [8, 9, 10].
3. **Graph Convolutional Networks and Multimodal Approaches**: Recent advancements include the use of Graph Convolutional Networks (GCNs) to model spatial and temporal dependencies in video data. Additionally, multimodal approaches that integrate facial expressions with other modalities such as voice and body language have been explored to provide a more comprehensive understanding of emotions [11, 12].

1.5 Objectives of This Research

This research's main objective is to develop a robust real-time facial emotion recognition web application leveraging deep learning and graph convolutional networks. The specific goals are as follows:

1. **Dataset Selection and Preprocessing**: Identify suitable datasets for training and testing the FER models. Implement preprocessing techniques to enhance data quality and model performance [5].
2. **Model Development and Benchmarking**: Develop and benchmark various FER models, including classical machine learning methods and state-of-the-art deep learning architectures [5, 6, 8].
3. **Graph Convolutional Networks for Video FER**: Explore the use of GCNs to capture the dynamic nature of facial expressions in video sequences [11].

4. **Web Application Development**: Design and implement a user-friendly web application that allows users to upload or record videos for real-time emotion analysis [5].
5. **Evaluation and Future Work**: Evaluate the performance of the developed models and identify areas for future improvement, including the potential for multimodal emotion recognition [3].

1.6 Significance of the Study

This study contributes to the field of facial emotion recognition in several ways:

1. **Advancement of FER Techniques**: By exploring and benchmarking various FER models, this research advances the understanding of effective techniques for emotion recognition from facial expressions [6].
2. **Real-time Application**: The development of a real-time web application demonstrates the practical applicability of the research, bridging the gap between theoretical models and real-world usage [5].
3. **Exploration of GCNs**: The use of graph convolutional networks for video-based FER represents a novel approach that leverages the spatial and temporal dynamics of facial expressions, potentially leading to more accurate and robust models [11].
4. **Foundation for Future Research**: This research lays the groundwork for future studies on multimodal emotion recognition and the integration of FER systems into various applications [12].

Chapter 2

Literature Review

2.1 Introduction

Facial emotion recognition (FER) has evolved significantly over the past few decades. This section reviews the key developments and approaches in FER, highlighting the transition from early methods to the latest advancements in deep learning and graph convolutional networks.

2.2 Early Approaches to Facial Emotion Recognition

2.2.1 Hand-crafted Features

Early FER systems relied heavily on hand-crafted features to represent facial expressions. These features were designed manually to capture specific aspects of facial appearance and geometry.

2.2.1.1 Local Binary Patterns (LBP)

- **Overview**: LBP is a texture descriptor that labels the pixels of an image by thresholding the neighborhood of each pixel and considering the result as a binary number.
- **Advantages**: LBP is robust to changes in lighting conditions and computationally efficient.
- **Applications**: Widely used in face recognition and FER due to its effectiveness in capturing local texture information [6].

2.2.1.2 Histogram of Oriented Gradients (HOG)

- **Overview**: HOG captures edge and gradient structures in an image by dividing it into small, connected regions (cells) and computing the histogram of gradient directions for the pixels within each cell.
- **Advantages**: HOG is effective in capturing local shape information and is robust to small changes in pose and lighting.

- **Applications**: Used for human detection and FER tasks [15].

2.2.1.3 Geometric Features

- **Overview**: Geometric features involve the extraction of the shapes and locations of facial landmarks, such as the eyes, nose, and mouth.
- **Advantages**: These features are intuitive and closely related to the muscle movements that generate facial expressions.
- **Applications**: Commonly used in combination with appearance features for improved FER accuracy [16].

2.2.2 Classical Machine Learning Methods

Once the features were extracted, classical machine learning methods were employed to classify the emotions.

2.2.2.1 Support Vector Machines (SVM)

- **Overview**: SVM is a supervised learning algorithm that finds the hyperplane that best separates the data into different classes.
- **Advantages**: Effective in high-dimensional spaces and robust to overfitting.
- **Applications**: Widely used for classification tasks in FER [17].

2.2.2.2 Random Forests

- **Overview**: Random Forests is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks.
- **Advantages**: Handles large datasets with higher dimensionality and provides high accuracy.
- **Applications**: Used for various classification and regression tasks, including FER [7].

2.2.3 Deep Learning Methods

The introduction of deep learning has significantly advanced the field of FER. Deep Convolutional Neural Networks (DCNNs) have become the standard approach due to their ability to automatically learn hierarchical feature representations from raw data.

2.2.3.1 Deep Convolutional Neural Networks (DCNNs)

- **Overview**: DCNNs consist of multiple layers of convolutional filters that learn to extract features directly from raw pixel values.
- **Advantages**: Capable of learning complex, high-level features without manual intervention, leading to state-of-the-art performance in image classification tasks.
- **Applications**: Widely used in FER for their superior accuracy and ability to handle large datasets [1].

2.2.3.2 VGGNet

- **Overview**: VGGNet is a deep convolutional network with a simple and uniform architecture, using small 3x3 convolution filters.
- **Advantages**: Easy to implement and modify, with high performance on image classification benchmarks.
- **Applications**: Commonly used as a backbone for feature extraction in FER [8].

2.2.3.3 ResNet

- **Overview**: ResNet introduces residual connections that help mitigate the vanishing gradient problem, allowing the training of very deep networks.
- **Advantages**: Enables the construction of extremely deep networks that maintain high accuracy.
- **Applications**: Widely used in FER for its robustness and high performance [9].

2.2.3.4 Inception

- **Overview**: The Inception architecture uses multiple filter sizes in each layer, allowing the network to capture features at different scales.
- **Advantages**: Efficient in terms of computational resources and effective in capturing multi-scale features.
- **Applications**: Used in FER for its ability to handle diverse facial expressions and occlusions [10].

2.2.3.5 YOLOv7

- **Overview:** YOLOv7 builds on the strengths of its predecessors, aiming to improve accuracy, speed, and efficiency. It incorporates advanced techniques like deeper

networks, better anchor generation, and improved feature pyramids, making it more effective for a wide range of object detection tasks.

- **Advantages:** Maintains the real-time detection capability with minimal latency and enhanced detection performance, especially for small objects. Optimized for better computational efficiency, suitable for deployment on edge devices.
- **Applications:** Used in real-time object detection in autonomous vehicles. It has enhanced monitoring and threat detection in security systems.

2.2.3.6 YOLOv8

- **Overview:** It introduces more sophisticated architectural changes, including better backbone networks, enhanced feature extraction methods, and advanced post-processing techniques. YOLOv8 aims to provide higher accuracy and faster inference times, making it suitable for even more demanding applications.
- **Advantages:** Reduced inference times with more efficient network designs. Incorporates the latest advancements in deep learning and computer vision, such as improved backbone networks and better feature fusion techniques.
- **Applications:** Used in medical imaging analysis, including detecting abnormalities in X-rays and MRIs. It has also been widely used in real-time object detection and manipulation for industrial robots and drones.

2.2.4 Graph Convolutional Networks (GCNs)

To address the temporal dynamics in videos, researchers have turned to Graph Convolutional Networks (GCNs), which can model spatial and temporal dependencies in data.

2.2.4.1 Spatial-Temporal Graph Convolutional Network (ST-GCN)

- **Overview**: ST-GCN models both spatial and temporal dependencies in sequential data by representing the data as a graph.
- **Advantages**: Effectively captures the dynamic nature of facial expressions over time.
- **Applications**: Used in action recognition and FER for video sequences [11].

2.2.4.2 Relational Reasoning in GCNs (RA-GCN)

- **Overview**: RA-GCN extends ST-GCN by incorporating relational reasoning, enhancing the model's ability to understand complex interactions in video data.

- **Advantages**: Improves the accuracy of emotion recognition by modeling the relationships between facial landmarks more effectively.
- **Applications**: Used for advanced FER tasks involving video data [12].

2.2.5 Multimodal Emotion Recognition

Multimodal emotion recognition systems integrate facial expressions with other modalities such as voice, body language, and physiological signals to provide a more comprehensive understanding of emotional states.

- **Overview**: Combining multiple modalities can capture a wider range of emotional cues and improve the robustness of emotion recognition systems.
- **Applications**: Used in scenarios where a single modality might be insufficient to accurately detect emotions [3].

2.2.6 Transfer Learning

Transfer learning leverages pre-trained models on large datasets and fine-tunes them for specific tasks like FER. This approach significantly reduces the need for large, annotated datasets.

- **Overview**: Transfer learning allows models trained on large-scale datasets (e.g., ImageNet) to be adapted for FER by fine-tuning on emotion-specific datasets.
- **Advantages**: Reduces training time and computational resources while improving model performance.
- **Applications**: Widely used in FER to achieve state-of-the-art results with limited data [13].

2.3 Datasets

The availability of high-quality annotated datasets is crucial for training and evaluating facial emotion recognition (FER) models. Several benchmark datasets have been developed over the years, providing a foundation for advancements in the field.

2.3.1 FER2013 Dataset

The FER2013 dataset was introduced by Goodfellow et al. during the ICML 2013 Challenges in Representation Learning. It is one of the most widely used datasets for facial emotion recognition.

- **Overview**: FER2013 contains 35,887 grayscale images of faces, each labeled with one of seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral.
- **Image Specifications**: Each image is of size 48x48 pixels, captured in various environments, including real-world scenarios, making the dataset diverse and challenging.
- **Purpose**: The dataset was designed to encourage research in developing robust FER models capable of handling variations in lighting, head poses, and facial expressions [5].

2.3.2 Expression in-the-Wild (ExpW) Dataset

The Expression in-the-Wild (ExpW) dataset was created to address the limitations of lab-controlled datasets by providing images captured in uncontrolled environments.

- **Overview**: The ExpW dataset contains around 91,793 images of faces annotated with six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) plus neutral.
- **Image Specifications**: The images vary significantly in terms of lighting, occlusion, head pose, and background clutter, reflecting real-world conditions.
- **Purpose**: ExpW was designed to test the robustness of FER systems in real-world applications where faces are not perfectly aligned or uniformly lit [18].

2.3.3 Aff-Wild2 Dataset

The Aff-Wild2 dataset is part of the Affective Behavior Analysis in-the-Wild (ABAW) challenge, designed to provide a comprehensive benchmark for FER systems.

- **Overview**: Aff-Wild2 is the largest dataset for FER, containing over 2.8 million frames from 539 videos. It includes annotations for seven basic emotions, valence-arousal, and action units.
- **Image Specifications**: The dataset captures spontaneous expressions in unconstrained environments, with variations in lighting, head poses, and occlusions.
- **Purpose**: Aff-Wild2 aims to promote the development of FER systems that can accurately recognize emotions in dynamic, real-world scenarios [14].

Chapter 3

Methodology

3.1 Introduction

Deep Convolutional Neural Networks have become a state-of-the-art method for the image classification tasks. People can solve many tedious problems using these networks. The credit for their success goes to the large availability of data and compute power. However, their biggest limitation is it requires large amounts of correctly labelled data.

The given problem of emotion recognition suffers exactly from this problem as emotions are very complex and subjective. The existing facial expression databases are not sufficient to train the well-known neural network with deep architecture that achieved the most promising results in object recognition tasks. High inter-subject variations exist due to different personal attributes, such as age, gender, ethnic backgrounds and level of expressiveness tend to be some of the challenges in this domain. On top of these occlusions, pose variations and wearable accessories also exist in facial images which makes the problem even harder.

The development of an effective facial emotion recognition (FER) system involves several critical steps, including dataset selection, data preprocessing, model development, and evaluation. This section outlines the methodologies employed in this research, detailing each step to ensure a robust and accurate FER system.

3.2 Dataset Selection

The choice of datasets is fundamental to training and evaluating FER models. For this research, three primary datasets were selected: FER2013, Expression in-the-Wild (ExpW), and Aff-Wild2. These datasets provide a comprehensive range of facial expressions captured in both controlled and uncontrolled environments, facilitating the development of models that generalize well across different scenarios. We selected the 7 basic emotion categories: **Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise.**

Dataset Name	Training Samples	Validation Samples	Classes	Details
FER 2013	28K	8K	7 Clasess	Image size – 48*48
Expression in-the-Wild	75K	15K	5 Classes	Image size – 48*48
Aff-Wild2	1200	586	5 Classes	Spontaneous actions, head movements and occlusion Frame Length: 45 frames

Table 3.1 Summary of the datasets shortlisted for further analysis

- **FER2013**: Selected for its balanced and well-annotated collection of ~36K grayscale images across seven emotion categories [5].
- **Expression in-the-Wild (ExpW)**: Chosen for its diverse set of images captured in real-world conditions, providing ~90K images labeled with six basic emotions plus neutral [18].
- **Aff-Wild2**: Included for its extensive video data with over 2.8 million frames, capturing spontaneous expressions in the wild, and annotated for emotions, valence-arousal, and action units [14].

3.3 Data Preprocessing

The accuracy of a model can be increased significantly if the data is preprocessed properly. Background, illumination and head pose variation tend to be challenging in unconstrained scenarios. To mitigate this and learn meaningful features, the following pre-processing techniques were included:

3.3.1 Face Detection

To perform Facial Emotion Recognition, one of the prerequisites is to identify the faces in the image. This is a very crucial step in the pipeline because of the following 3 reasons:

- If the component has a high false positive rate, then every other bounding box will be characterized incorrectly as a face.

- If the component has a high false negative rate, then only partial faces will be identified in the video feed.
- The face detection model should be performed in real time to keep up with the rapidly varying facial expressions.

Keeping these factors in consideration, I benchmarked the following 3 methods for face detection:

3.3.1.1 Multi-Task Cascaded Convolutional Networks (MT-CNN)

This framework is comprised of a Proposal Network, Refine Network and Output Network. The proposal network predicted the candidate windows for the faces followed by a refined network which removed the false positive cases. Finally, the output network used non-maximum suppression algorithm and bounding box regression to predict the bounding box and facial landmarks.

Although the model performs accurately, it is slow even on GPU (around 5-6 FPS) which is not suitable for our real time application.

3.3.1.2 Dlib Implementation

Next, I looked at the dlib python library which supports state of the art Face Detection models. This model regresses over 68 facial landmarks and estimates the bounding box based on that. Although the model performs well in terms of speed (around 15 FPS), it lacks accuracy and is not robust to varying poses and partial occlusions.

3.3.1.3 RetinaFace

To strike a balance between speed and accuracy, I turned to another state-of-the-art method for face detection, i.e. RetinaFace. This method proposes a novel pixel wise face localization method with multi-task learning strategy to simultaneously predict face score, face box, facial landmarks etc. I use a lightweight version of this model 'RetinaNetMobileNet' which helps me achieve around 30 FPS while also maintaining a very good accuracy and tracking.

3.3.2 Face Alignment

Facial landmarks extracted from the RetinaFace are further used to align the images in the video feed. This is a crucial data preprocessing technique as it can greatly improve the model's performance in case of any downstream task related to faces.

3.4 Evaluation

The performance of the developed models was evaluated using standard metrics such as accuracy, precision, recall, and F1-score and confusion metrics. Cross-validation was employed to ensure robustness and generalizability of the models.

- **Cross-Validation**: K-fold cross-validation was used to assess the model performance, with k set to 10 for this research [19].
- **Metrics**: The primary metrics for evaluation included accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the model performance [20].

3.5 Web Application Development

To facilitate real-time facial emotion recognition, a web application was developed using modern web technologies.

- **Front-End**: The front-end was built using React.js for a responsive and user-friendly interface.
- **Back-End**: The backend was implemented using Django, providing a robust framework for handling model inference requests and serving the results.
- **Deployment**: The application was deployed on TTO server to ensure security and availability. Later, it will be deployed on a cloud platform to ensure scalability and availability, with Docker containers used for environment consistency.

Chapter 4

Implementation and Results

4.1 Introduction

This chapter details the implementation process of the facial emotion recognition (FER) system and presents the results obtained from various models. The implementation phase involves translating the conceptual methodologies discussed in the previous chapters into a functional system capable of accurately recognizing emotions from facial expressions. This process includes setting up the development environment, implementing the selected models, training these models on the chosen datasets, and optimizing their performance through rigorous testing and evaluation.

The results section provides a comprehensive analysis of the performance of different models, including classical machine learning approaches and advanced deep learning architectures. Each model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, the results from various datasets, including FER2013, Expression in-the-Wild (ExpW), and Aff-Wild2, are compared to determine the models' effectiveness in handling diverse and complex real-world data.

4.2 Classical Machine Learning Method

We investigated the efficacy of classical ML models for the task of Emotion Recognition. For this, we considered the hand-crafted features: Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) using facial landmarks. We used the FER 2013 dataset for benchmarking.

Dataset Name	Training Samples	Validation Samples	Classes	Details
FER 2013	28K	8K	7 Classes	Image size – 48*48

Table 4.1 Details of FER 2013 Dataset

FER 2013 dataset contains 7 emotions: *Anger, Fear, Sad, Neutral, Happy and Surprise*

Figure 4.1 Sample images for FER 2013 dataset emotions

4.2.1 Local Binary Pattern Features

In this approach, a facial image is divided into local patches, and then local binary pattern features represent that local appearance. These features have proven to be a powerful descriptor in expression recognition and face verification. These concatenated features give a global descriptor of a face and represent region and pixel level features. Further, to reduce the size of feature vectors, we perform LDA to remove redundant features. The accuracy of these features on the various ML models for the test set of FER 2013 dataset is given in the below table:

Classifier	***Accuracy***	***Weighted F1 Score***
Decision Tree	31%	0.31
Random Forest	38%	0.33
Linear SVM	38%	0.34
SVM (RBF Kernel)	31%	0.26

Table 4.2 Results on various ML models for Local Binary Pattern Features

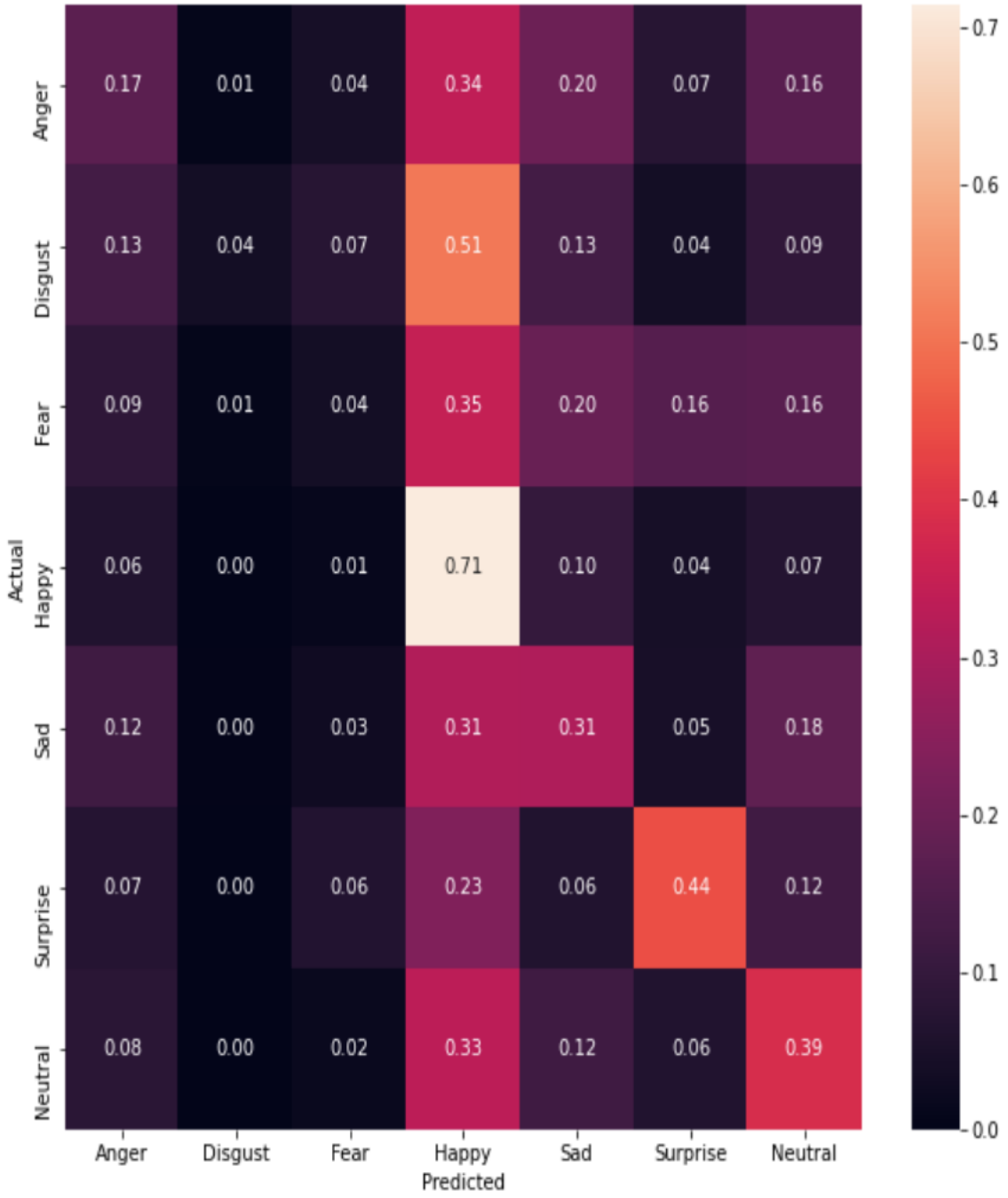

Figure 4.2 Confusion Matrix for Local Binary Pattern Features

4.2.2 Dense SIFT + CNN

The SIFT algorithm is used to identify the key points in an image. For each key point, a key point descriptor is obtained, which is a histogram of local gradients around the key point. We use Dense SIFT in which SIFT descriptors are computed over dense grids in the image and not just at the key points. We use a CNN classifier on top of these features to recognize 5 basic emotions Happy, Neutral, Sad, Anger, Suprise. This model performed slightly better than the previous models with an accuracy of 43% and a weighted F1 score of 0.42 on the FER dataset.

Classifier	**Accuracy**	**Weighted F1 score**
CNN	49.3%	0.42

Table 4.3 Results for Dense SIFT + CNN

4.2.3 Face Landmarks + HOG using SVM classifier

In this approach we use the Facial Landmarks using the Dlib library with the Histogram of Gradient Orientations. These concatenated features are then passed to a multi-class SVM classifier with RBF kernel. We report an accuracy of 51% with a weighted F1 score of 0.48 on the test set.

Classifier	***Accuracy***	***Weighted F1 Score***
SVM (RBF Kernel)	50.76%	0.48

Table 4.4 Results for Face landmarks + HOG using SVM classifier

This is the best performance we achieved using the handcrafted features. However, we can see that these models do not perform very well on classes Disgust and Fear which are very complex. So, to tackle this we plan to train deep learning models on larger datasets and benchmark the transfer learning on the FER 2013 dataset.

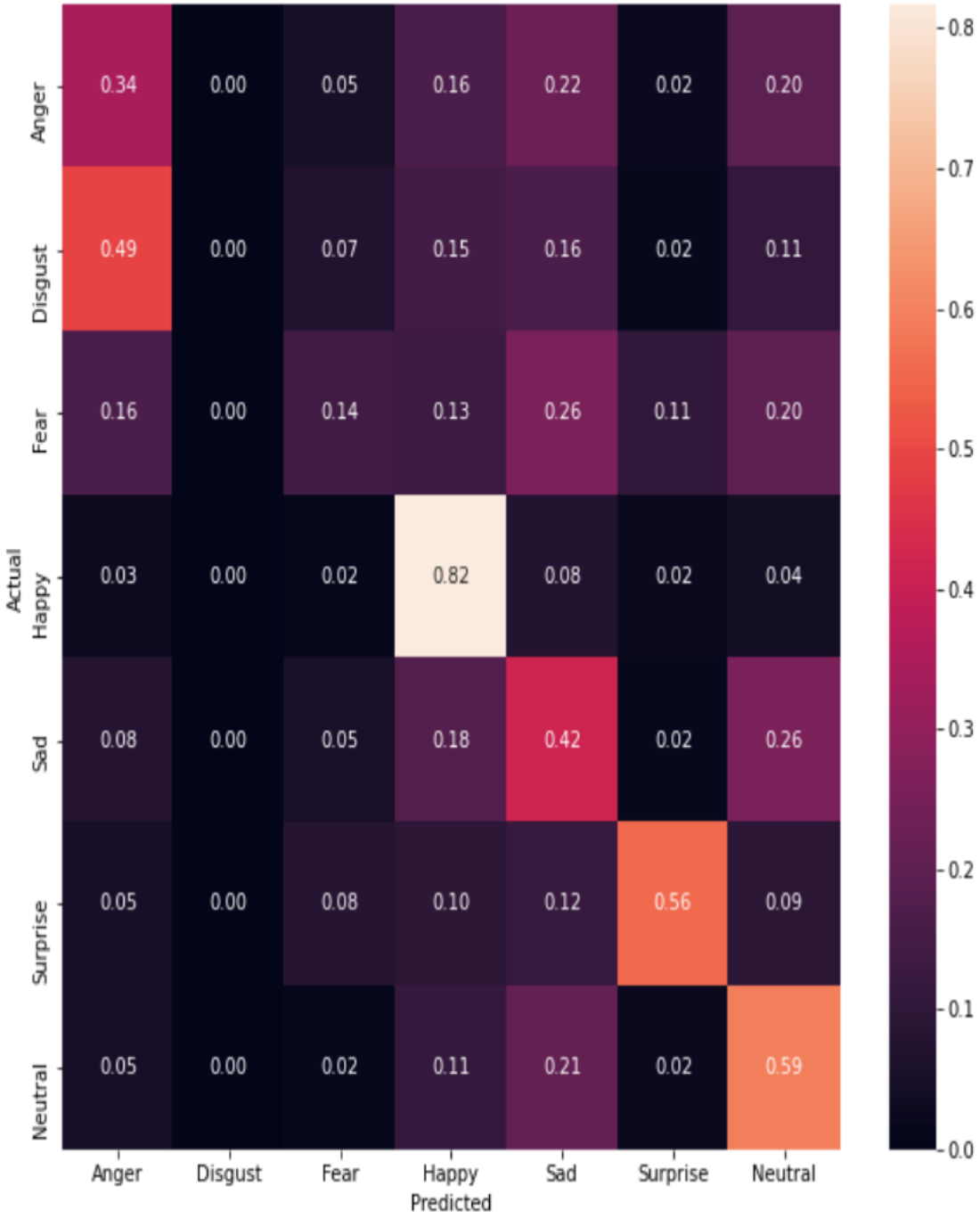

Figure 4.3 Confusion Matrix for Face landmarks + HOG using SVM classifier

4.3 Deep Learning Method

This section is divided into 2 parts: first is what I proposed and second is what I implemented eventually.

4.3.1 Proposal

Based on my literature review I proposed to try the following important ideas discussed in earlier works.

- I planned to follow a 2-stage pipeline where I use a pre-trained deep face net (VGG Face2) and train it on FER 2013 dataset. This would provide feature-level regularization to the convolutional features. Then in the second stage we could train the model on the target emotion recognition dataset by adding new randomly initialized fully connected layers.

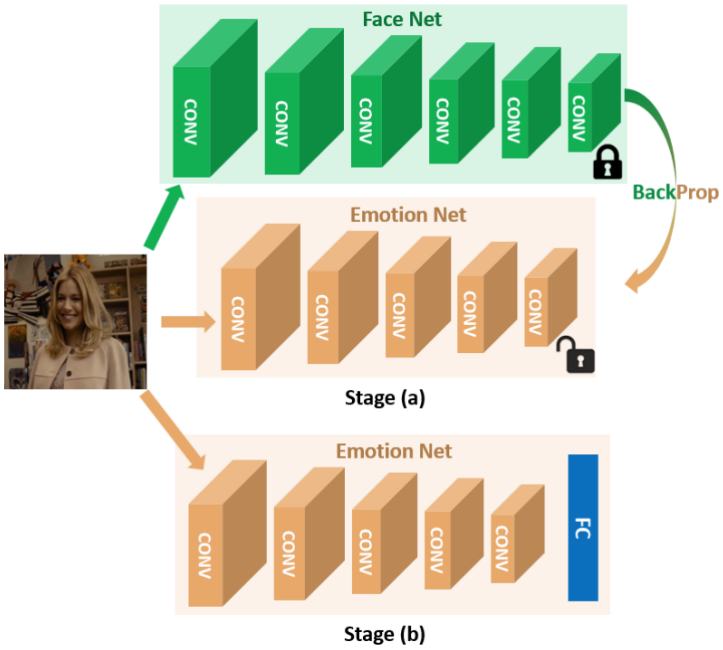

Figure 4.4 Deep Learning Method Proposal Architecture

Stage a) Frozen VGG-Face 2 + Conv layers on FER 2013

Stage b) Add FC layers and train on a new emotion recognition dataset.

- I also planned to concatenate features at intermediate layers in the network and use this ensemble of features to predict the emotion. This idea was derived from the work published as Learning supervised scoring ensemble for emotion recognition in the wild.

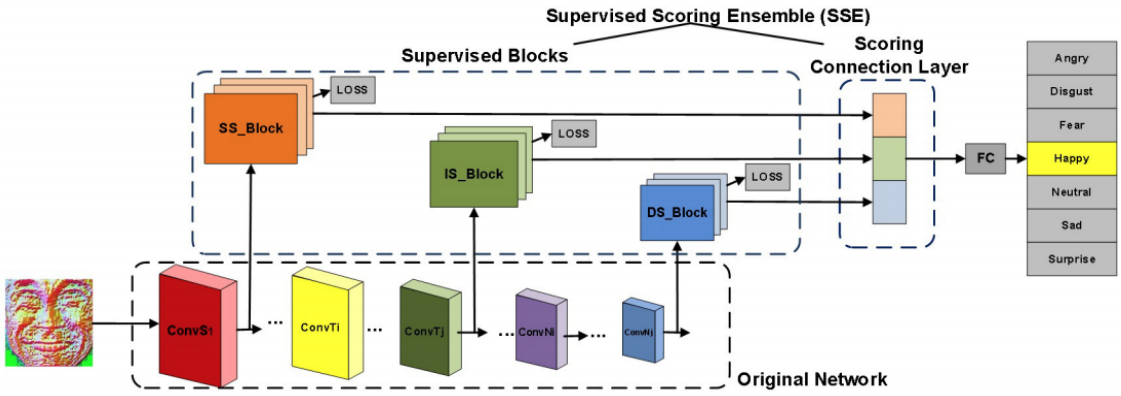

Figure 4.5 Concatenation of features at different layers

- Network Ensemble - We could use an ensemble of multiple networks learning complementary features trained on different datasets. This aggregate of features would be more robust and hence provide better accuracy.

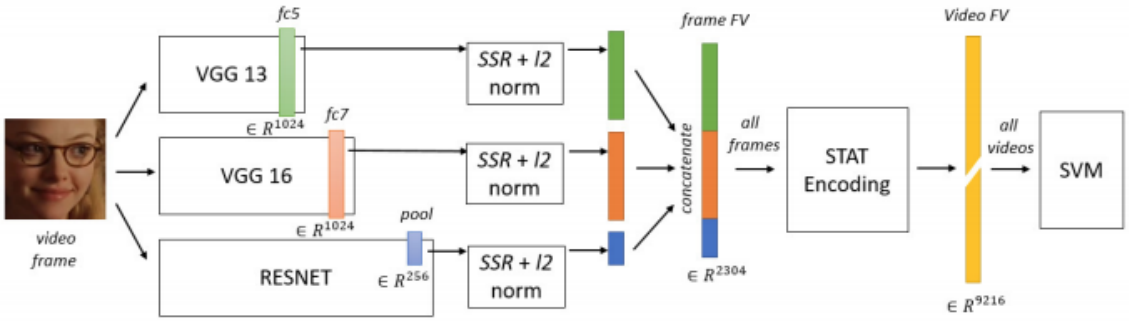

Figure 4.6 Network Ensemble

4.3.1.1 Emotion Recognition in Videos

Most of the above discussed models focus on static images, facial expression recognition can benefit from the temporal correlations of consecutive frames. Hence, I proposed to use frame aggregation techniques to combine deep learnt features from static FER networks.

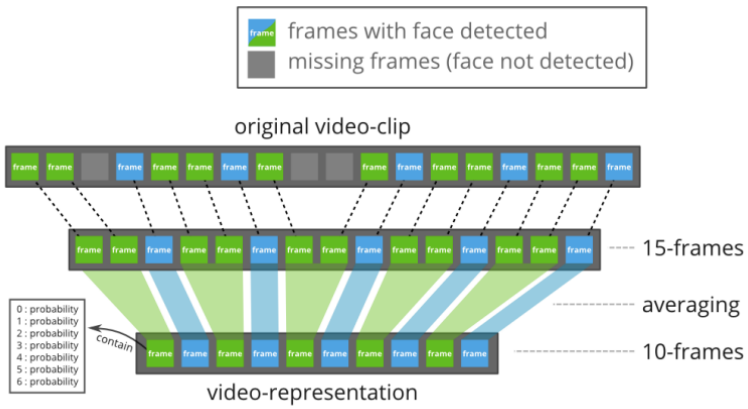

Figure 4.7 Frame Aggregation Technique

In video emotion recognition, the transition between different emotions tends to follow up frames with lower intensity expressions. So, we can train a network that tries to minimize a cross-entropy loss and an auxiliary L2 loss. This will learn better features suited for emotion recognition.

I proposed to benchmark these ideas on the selected datasets and then train an emotion recognition model which is robustly able to infer facial emotions during the state transitions.

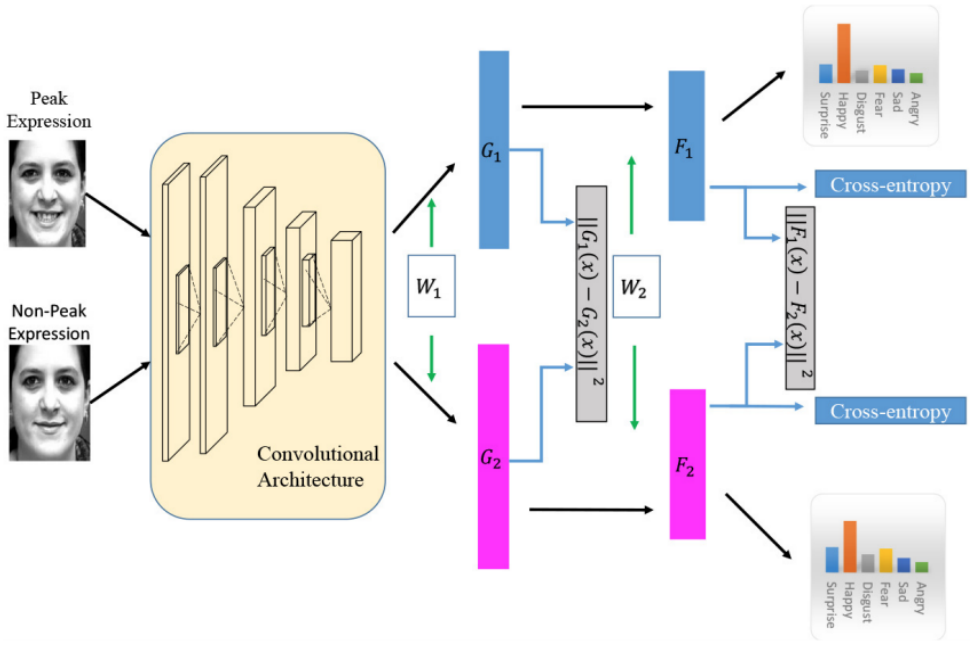

Figure 4.8 Cross-Entropy Architecture

4.3.2 Implementation

In this section, I will discuss the CNN based models that I tried on the different datasets for Facial Emotion Recognition and the milestones achieved. To evaluate the transfer learning ability of these models I trained the models on Expression in the Wild (EXPW) dataset and evaluated on the FER 2013 dataset. We omitted the class 'disgust' because it is very difficult to express it only through facial expressions which is the reason for very few samples for these classes in the training dataset. So finally, we ended up with 45k training samples distributed over 6 classes.

4.3.2.1 Facial Landmarks + HOG using CNN

The CNN model uses a concatenated feature vector comprising of Histogram of Gradient Orientation using a sliding window and the Facial Landmarks of the detected face. These also turned out to be the best performing features during our benchmarking of the traditional methods (Refer to this post). On validation set we found an accuracy of 55.0% on EXPW dataset and 60.8% on FER 2013 dataset. This was considerable progress considering that the model was not trained on FER dataset.

4.3.2.2 VGG19 Feature using CNN

Our next approach was to obtain more robust and discriminative features for Facial Emotion Recognition. Hence, we used a pretrained VGG19 on the facial dataset as the backbone for the features followed by a CNN. This resulted in a marginal boost of 3.8% on EXPW dataset but an increase of 11.5% on the validation set of FER 2013 dataset. This stacks our model close to some of the best performing models specifically trained on the FER 2013 dataset.

Features	***Classifier***	***Accuracy on Expr dataset***	***Accuracy on FER dataset***
HOG + Facial Landmarks	CNN	55.0%	60.8%
VGG 19 Features	CNN	58.8%	72.3%

Table 4.5 Results for Deep Learning Methods

4.3.2.3 YOLOv7

YOLO models use multiple convolutional layers to extract features from images. These layers learn to detect edges, textures, patterns, and more complex structures as the network depth increases. We trained YOLOv7 on the FER2013 to extract facial regions and then fed it

into a CNN. This resulted in an accuracy of 63% which placed it close to some of the best performing models specifically trained on the FER 2013 dataset.

Features	***Classifier***	***Accuracy on FER dataset***
YOLOv7	CNN	60.8%

Table 4.6 Results for YOLOv7 model

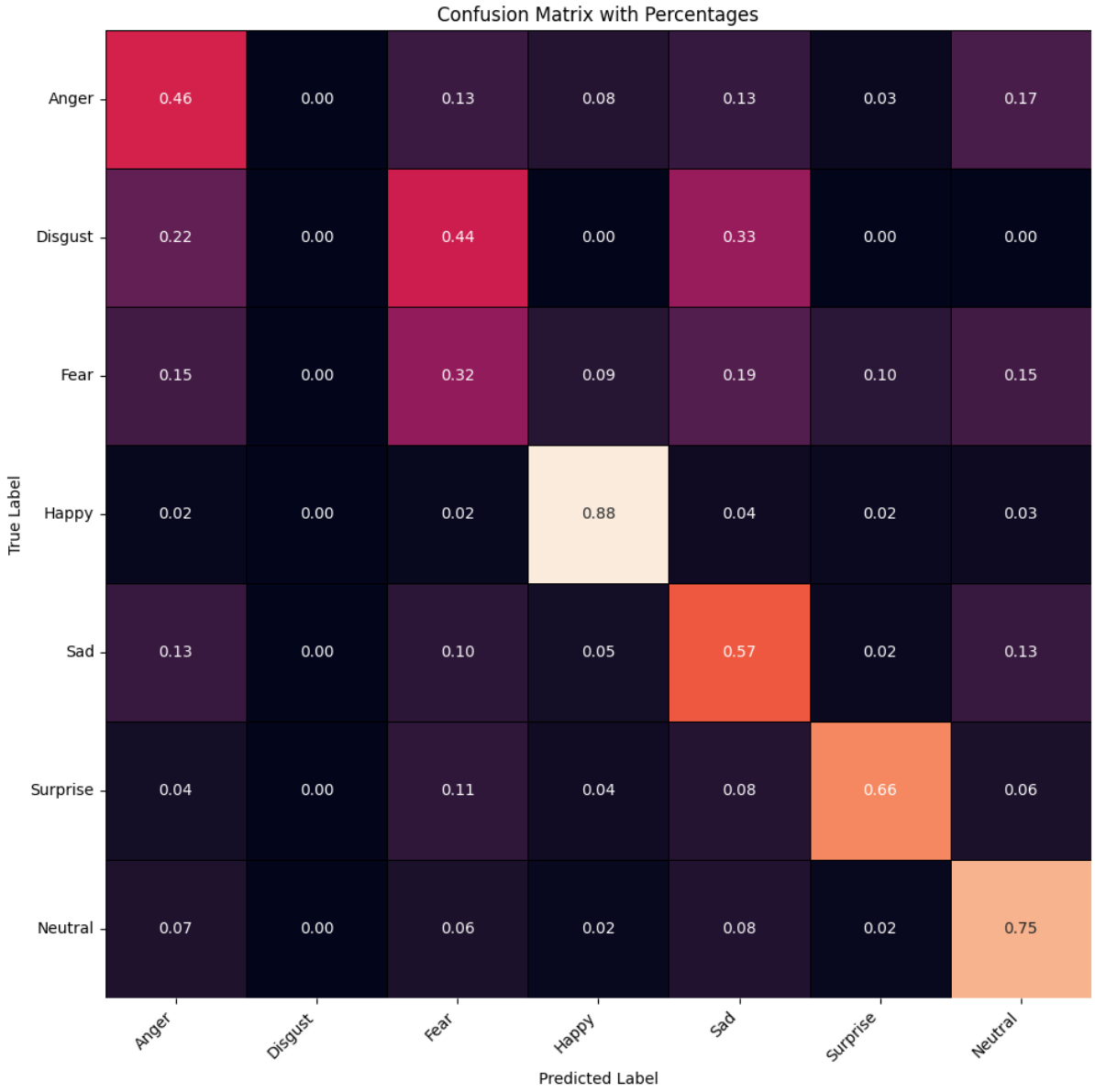

Figure 4.9 Confusion matrix for YOLOv7 model

4.3.2.4 YOLOv8

Our next approach was to try the next model in YOLO series i.e. YOLOv8. We found further improvements in detection accuracy, outperforming YOLOv7. This model gave us accuracy of 55% but it could be improved further by running more epochs.

Features	***Classifier***	***Accuracy on FER dataset***
YOLOv8	CNN	55%

Table 4.7 Results for YOLOv8 model

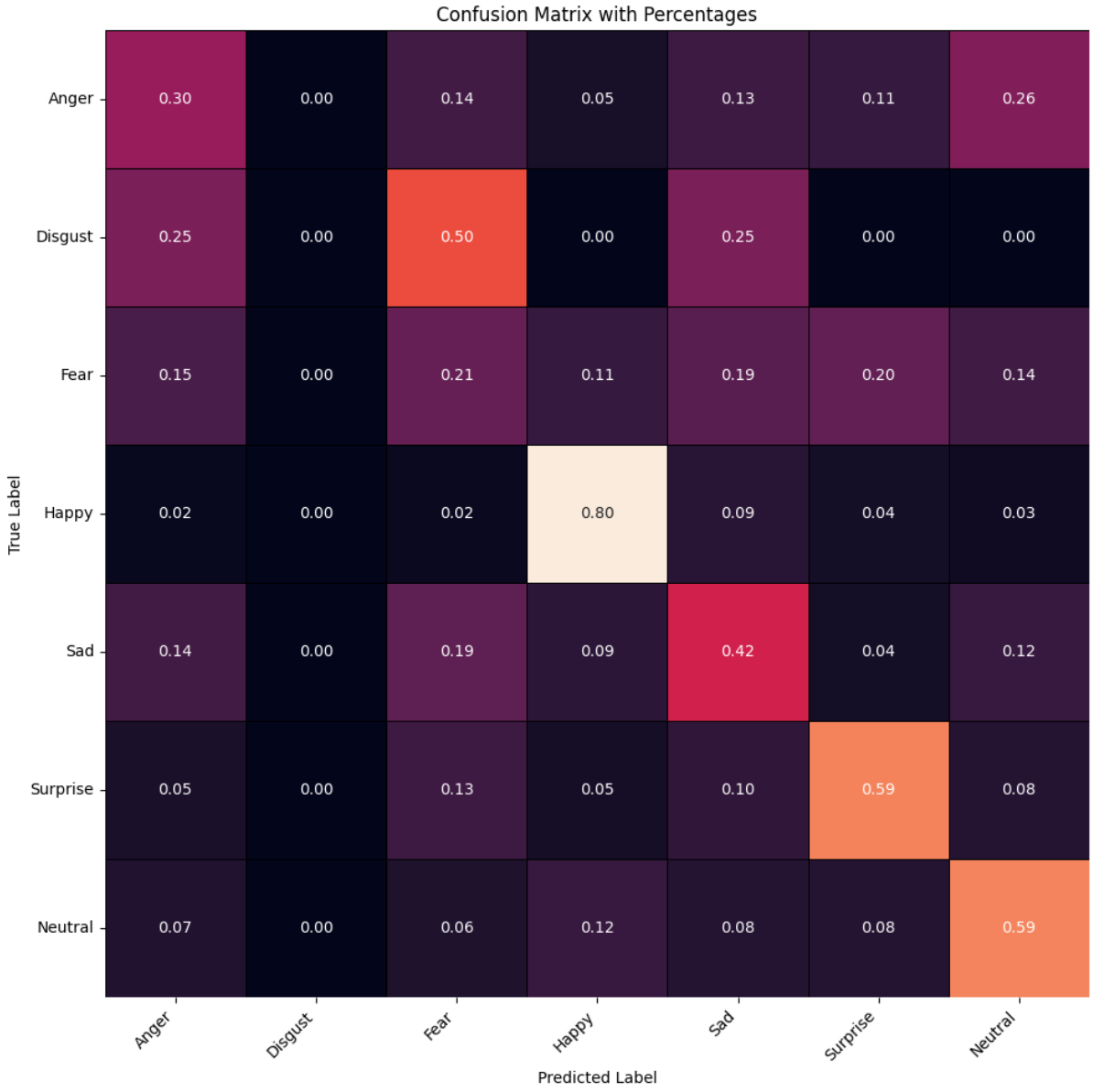

Figure 4.10 Confusion matrix for YOLOv8 model

4.4 Graph Convolution Networks for Emotion Recognition

Facial expressions for emotions are not consistent for each image in a video. There are intermediate frames during a transition between emotions. Such facial expressions are very difficult to classify as they don't belong to any particular class. This is also one major reason why models trained on image datasets for facial emotion recognition perform poorly on videos. To circumvent this obstacle, I used a Spatio-Temporal Graph Convolutional Neural Networks for our task.

4.4.1 Dataset Preparation

For these networks I used a video dataset Aff-Wild2. I used the 6 basic emotion categories: **Happy, Sad, Neutral, Anger, Surprise, Fear**. The cleaned dataset consisted of 1200 training samples and 586 validation samples. Each sample consisted of 45 frames representing the same emotion class with varying intensity and transition frames. To create this dataset, for each video sample, I computed the 68-dimensional facial landmarks for the 45 frames. This represents our feature vector of the shape 45 X 68 X 2 as our input to the GCN model. Such a feature vector allows us to capture both the spatial and temporal changes over the video.

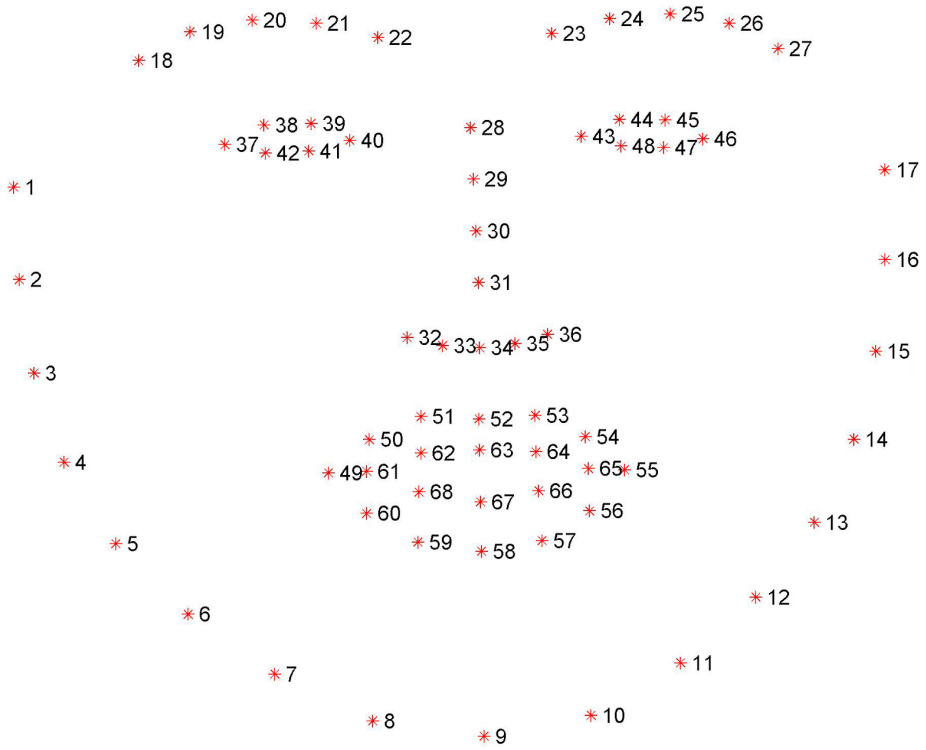

Figure 4.11 68-dimensional facial landmarks

4.4.2 Implementation

4.4.2.1 ST-GCN

Next, we define the graph where each 2D facial landmark represents a graph node and natural connectivity in the facial structure represents the graph edges. To capture the temporal dynamics of the facial expressions we also define edges between same joints across the consecutive frames. The input to the ST-GCN is therefore the joint coordinate vectors on the graph nodes. This can be considered as an analog to image based CNNs where the input is formed by pixel intensity vectors residing on the 2D image grid. Then we pass this through the multiple layers of spatial-temporal graph convolution as defined in ST-GCN paper followed by a SoftMax layer for classification. With this model we got an accuracy of 58.27% on the validation set.

4.4.2.2 RA-GCN

Further inspired by RA-GCN we also tried a multi-stream ST-GCN with same dataset and graph structure. However, this model overfitted to the training set due to a smaller dataset with many parameters. We achieved a low accuracy of 44.93% on the Aff-Wild2 dataset.

Features	*Classifier*	*Accuracy on Aff-Wild2 dataset*
Facial Landmarks + Temporally associated	ST-GCN	58.27%
Facial Landmarks + Temporally associated	RA-GCN	44.93%

Table 4.8 Results for Graph Convolution Networks

4.5 Discussion and Conclusion

The implementation and results chapter has provided a comprehensive overview of the development, training, and evaluation of various facial emotion recognition (FER) models. This section synthesizes the key findings and discusses the implications of the results.

4.5.1 Summary of Key findings

- **Model Performance**: The deep learning models, particularly VGG 19 Feature using CNN demonstrated superior performance compared to classical machine learning models such as SVM and Random Forests. This was evident across multiple datasets, including FER2013, ExpW, and Aff-Wild2. The advanced architectures of these models allowed for better feature extraction and handling of complex variations in facial expressions.
- **Dataset Comparisons**: Each dataset presented unique challenges. FER2013, with its balanced and well-annotated images, served as an excellent benchmark for model training. ExpW provided a diverse set of real-world images, testing the models' robustness in uncontrolled environments. Aff-Wild2, with its extensive video data, enabled the evaluation of models in dynamic scenarios, highlighting the importance of temporal information for accurate emotion recognition.
- **Graph Convolutional Networks (GCNs)**: The implementation of ST-GCN and RA-GCN models showed promising results in capturing spatial and temporal dependencies in video data. These models performed particularly well on the Aff-Wild2 dataset, demonstrating their potential for real-time emotion recognition in video sequences.

4.5.2 Implications of the Results

- **Advancement of FER Technology:** The successful implementation and evaluation of advanced deep learning models and GCNs represents a significant step forward in FER. These models offer higher accuracy and robustness, making them suitable for various real-world applications, from mental health monitoring to interactive entertainment systems.
- **Dataset Diversity**: The results underscore the importance of using diverse datasets for training and evaluation. Models trained on a single dataset may not generalize well to different environments. Therefore, combining multiple datasets can lead to more robust and versatile FER systems.

Chapter 5

Web Application Development

5.1 Introduction

The development of a real-time web application for facial emotion recognition (FER) marks a crucial step in translating research into practical, user-friendly solutions. This chapter details the design and implementation of the web application, which leverages advanced deep learning models to analyze and classify facial emotions from user-uploaded or recorded videos. The application aims to provide a seamless and interactive user experience, demonstrating the real-world applicability of the FER system. Key components of the development process include front-end design using React.js, back-end integration with Django, and deployment on a TTO lab server. Through this application, users can effortlessly access cutting-edge FER technology, highlighting the potential for widespread adoption in various domains such as customer service, education, and mental health monitoring.

5.2 Architecture

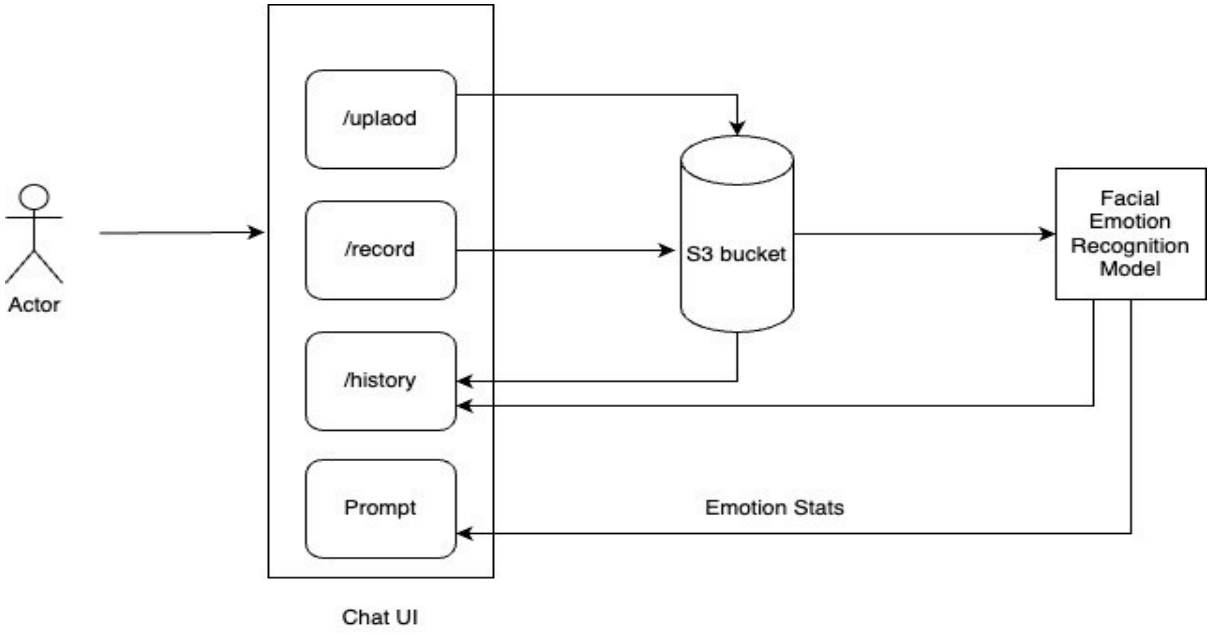

Figure 5.1 FER Web Application Architecture

5.2.1 Upload/Record a Video

The "Upload a Video" and "Record a Video" sections serve as gateways for users to input their video data into the facial emotion recognition model. The simplicity of the user interface is intentional, ensuring accessibility for a wide range of users.

5.2.1.1 Upload a Video

Users have the option to upload pre-recorded videos for emotion analysis. Upon selection, the video file is securely stored in an Amazon S3 bucket, ensuring data integrity and accessibility for subsequent processing.

5.2.1.2 Record a Video

For those preferring real-time interaction, the "Record a Video" section facilitates direct recording through the user's device camera. Similar to the "Upload a Video" section, the recorded video undergoes the same process—first saved in the S3 bucket and subsequently routed to the facial emotion recognition model.

5.2.2 Facial Emotion Recognition Model

Once the video data is stored in the S3 bucket, it is seamlessly fed into the trained facial emotion recognition model. The model processes the video, leveraging convolutional neural networks and transfer learning techniques developed in earlier sections of this thesis. Importantly, the model responds at various timestamps, segmenting the video based on moments of silence to provide nuanced insights into the user's emotional expressions.

5.2.3 Dashboard

The "Dashboard" section serves as a comprehensive overview, presenting a history of the processed videos along with the model's responses. Users can review the emotional analysis results, segmented by timestamps, providing a detailed account of their emotional expressions throughout the video.

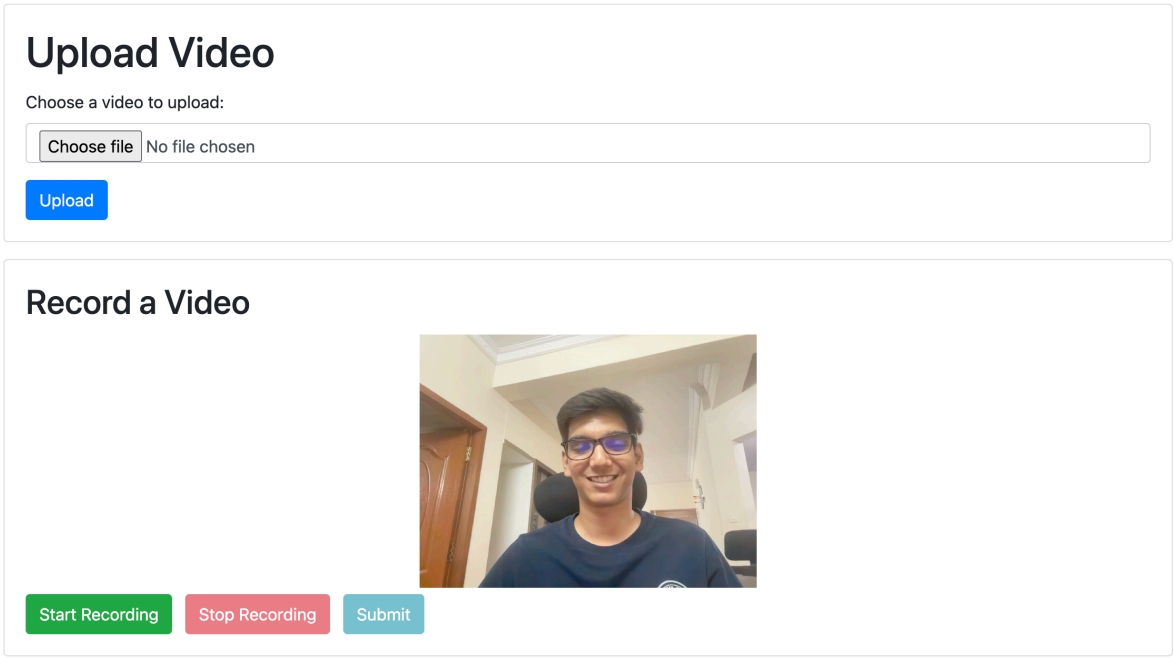

Figure 5.2 Dashboard Image 1

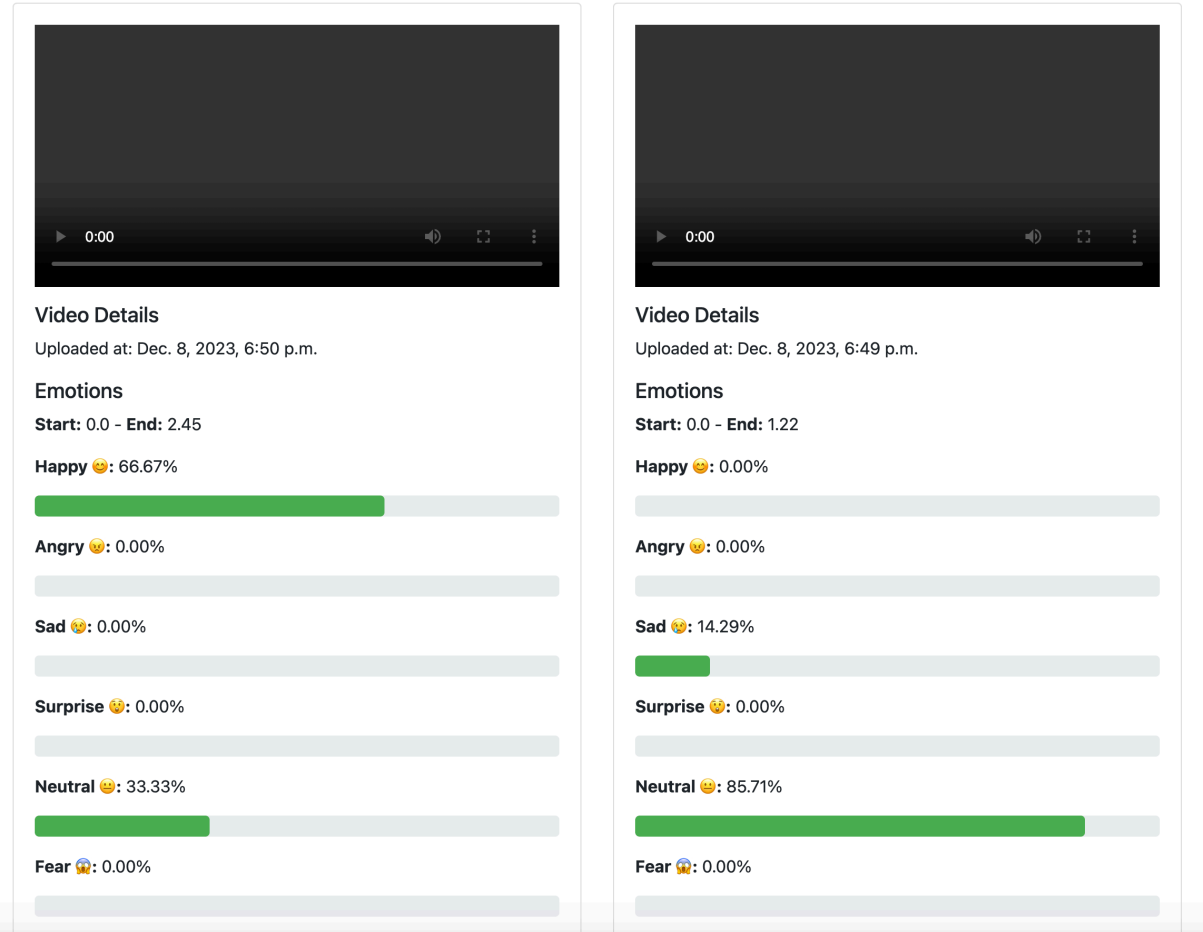

Figure 5.3 Dashboard Image 2

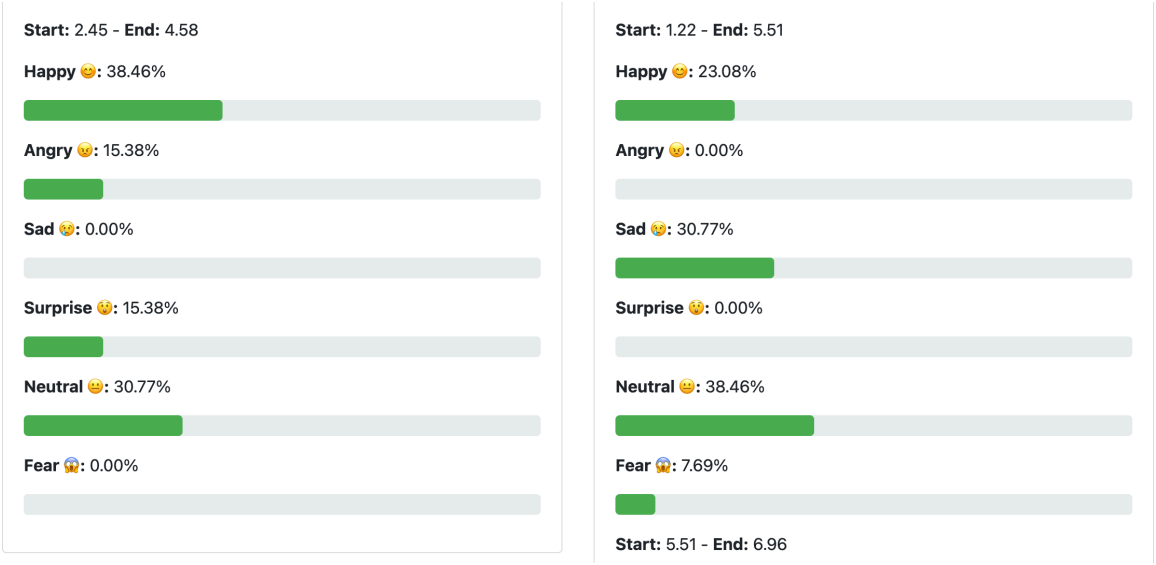

Figure 5.4 Dashboard Image 3

5.3 Practical Applications

The practical application of the facial emotion recognition (FER) web application extends across various domains, showcasing its versatility and potential to impact real-world scenarios. This section explores several key areas where the web application can be effectively utilized, demonstrating its value in enhancing user experiences, improving operational efficiency, and providing valuable insights into human emotions.

5.3.1 Customer Service

In customer service, understanding and responding to customer emotions is crucial for delivering exceptional experiences. The FER web application can be integrated into customer service platforms to analyze customer emotions during interactions. By detecting emotions such as frustration, satisfaction, or confusion in real-time, customer service representatives can adjust their responses, accordingly, ensuring more empathetic and effective communication. This capability can lead to increased customer satisfaction, loyalty, and overall service quality.

5.3.2 Mental Health Monitoring

The FER web application has significant potential in mental health monitoring. By analyzing facial expressions over time, the application can help identify emotional patterns and potential signs of mental health issues such as depression, anxiety, or stress. This non-

invasive monitoring tool can provide valuable insights to mental health professionals, enabling them to offer timely and personalized interventions. Additionally, the application can be used in telehealth platforms to enhance remote consultations and support.

5.3.3 Education

In educational settings, the FER web application can be used to gauge student engagement and emotional responses during lessons. By analyzing students' facial expressions, educators can gain insights into their emotional states and adapt their teaching methods to maintain engagement and address any issues. For example, if a significant number of students appear confused or disinterested, the teacher can modify their approach to better suit the students' needs. This application can enhance the learning experience and improve educational outcomes.

5.3.4 Human-Computer Interaction

The integration of FER technology into human-computer interaction (HCI) systems can lead to more intuitive and responsive interfaces. The web application can be used in various HCI applications, such as virtual assistants, gaming, and interactive media, to detect and respond to user emotions in real-time. For instance, in gaming, characters can react dynamically to players' emotions, creating a more immersive and engaging experience. In virtual assistants, understanding user emotions can lead to more personalized and effective interactions.

5.3.5 Security and Surveillance

In security and surveillance, the FER web application can enhance threat detection and situational awareness. By analyzing the emotions of individuals in real-time, security systems can identify potential threats or unusual behavior, such as stress or anger, which may indicate a security concern. This capability can be particularly useful in high-risk environments such as airports, public events, and critical infrastructure. The application can provide security personnel with valuable information to take proactive measures and ensure safety.

5.3.6 Market Research and Advertising

Understanding consumer emotions is vital for market research and advertising. The FER web application can be employed to analyze the emotional responses of individuals to products, advertisements, or marketing campaigns. By capturing real-time emotional feedback, businesses can gain insights into consumer preferences, enhance product development, and

optimize marketing strategies. This application can help companies create more engaging and effective advertisements that resonate with their target audience.

5.3.7 Healthcare

In healthcare, the FER web application can be used to monitor patients' emotional well-being during medical consultations and treatments. By analyzing facial expressions, healthcare providers can gain insights into patients' emotional states, which can be indicative of pain, anxiety, or discomfort. This information can help healthcare professionals tailor their approach to better address patients' needs and improve the overall patient experience.

5.3.8 Entertainment Industry

In the entertainment industry, the FER web application can revolutionize interactive gaming, VR/AR experiences, and film/TV by providing real-time emotional feedback, enabling personalized and dynamic content adjustments. This technology enhances audience engagement in live performances and allows streaming services to offer tailored content recommendations based on viewers' emotions, leading to a more immersive and satisfying entertainment experience.

5.4 Conclusion

The practical applications of the FER web application are diverse and far-reaching, spanning customer service, mental health monitoring, education, human-computer interaction, security, market research, and healthcare. By leveraging advanced FER technology, this application offers valuable insights into human emotions, enhancing interactions, improving outcomes, and providing actionable data across various domains. The potential for widespread adoption underscores the importance of continued development and refinement of the FER system to meet the evolving needs of these applications.

Chapter 6

Conclusions

6.1 Conclusion

In the pursuit of advancing real-time Facial Emotion Recognition (FER), this research has explored Convolutional Neural Network (CNN)-based models across different datasets. The endeavor aimed to harness the transfer learning capabilities of these models, with a focus on robust feature extraction for accurate emotion classification. Through experimentation and analysis, several milestones and insights have been achieved.

The initial approach incorporated Facial Landmarks and Histogram of Gradient Orientation (HOG) using a CNN, yielding commendable results. The concatenated feature vector showcased notable performance, achieving 55.0% accuracy on the Expression in the Wild (EXPW) dataset and 60.8% on the FER 2013 dataset. This not only underscored the importance of facial landmarks and HOG but also demonstrated the adaptability of the model to datasets on which it was not explicitly trained.

Subsequently, the integration of VGG19 features as the backbone for CNN further elevated the model's discriminative capabilities. The pretrained VGG19, when combined with CNN, resulted in a 3.8% improvement on the EXPW dataset and an impressive 11.5% boost on the FER 2013 dataset. This performance positioned our model close to some of the best-performing models trained on the FER 2013 dataset.

Recognizing the challenges posed by the variability in facial expressions within video sequences, we extended our exploration to Graph Convolutional Networks (GCN) for emotion recognition. Leveraging the Aff-Wild2 dataset, a spatio-temporal GCN (ST-GCN) was implemented, capturing both spatial and temporal changes over video frames. This approach achieved a notable accuracy of 58.27% on the validation set, showcasing the efficacy of GCNs in handling the dynamic nature of facial expressions in videos.

Despite these achievements, the attempt to further enhance performance through a multi-stream ST-GCN inspired by RA-GCN encountered challenges related to overfitting. The

model, while demonstrating the potential of this approach, faced limitations stemming from the dataset size and complexity.

In conclusion, this thesis has contributed valuable insights to the field of real-time Facial Emotion Recognition by examining diverse CNN-based models and delving into the realm of Graph Convolutional Networks for video-based emotion analysis. The amalgamation of traditional image-based approaches and innovative GCN techniques has not only advanced our understanding of facial emotion recognition but has also laid the groundwork for future research avenues. As technology continues to evolve, the findings presented here provide a stepping stone for the development of more sophisticated and adaptable models capable of deciphering the intricacies of human emotions in real-time applications.

6.2 Future Work

The present research has laid a foundation for advancements in real-time Facial Emotion Recognition (FER) and its practical application through a user-friendly web application. As we look to the future, several avenues for further exploration and improvement present themselves, offering opportunities to enhance both the technical capabilities and user experience of the developed system.

6.2.1 User Interface Refinement

While the current iteration of the web application provides a streamlined interface for users to interact with the facial emotion recognition model, there exists room for improvement in terms of aesthetics and user experience. Future work will involve refining the user interface to ensure a more intuitive and visually appealing experience. This includes optimizing the layout, incorporating user feedback mechanisms, and exploring interactive visualizations for the emotional analysis results.

6.2.2 Enhanced Model Robustness

Continuous efforts to bolster the robustness and adaptability of the facial emotion recognition model will be a focal point of future research. This involves expanding the training datasets to encompass a broader range of demographic and contextual factors, thereby minimizing biases and improving the model's generalization across diverse user groups. Additionally, exploring advanced techniques in model interpretability and explainability will contribute to a more transparent and trustworthy system.

6.2.3 Multimodal Emotion Recognition

Expanding the scope of emotion recognition beyond facial expressions to include other modalities, such as voice and body language, presents an exciting direction for future research. Integrating multimodal inputs can offer a more comprehensive understanding of emotional states, fostering a holistic approach to emotion recognition and analysis.

6.2.4 Usability Studies and User Feedback Analysis

Conducting usability studies and collecting user feedback will be essential for gauging the effectiveness and acceptance of the web application in real-world scenarios. Insights gathered from user interactions and experiences will inform iterative improvements, ensuring that the system aligns with user expectations and requirements.

6.2.5 Scalability and Deployment Optimization

As the user base of the web application grows, ensuring scalability and optimizing deployment will become crucial. Future work will focus on refining the architecture to accommodate increased user traffic, exploring cloud-based solutions for efficient resource utilization, and implementing mechanisms for secure and seamless scalability.

Bibliography

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
2. Ekman, P. (1999). Basic Emotions. *Handbook of Cognition and Emotion*, 45-60.
3. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.
4. Cohn, J. F., Zlochower, A. J., Lien, J. J., & Kanade, T. (1999). Feature-point tracking by optical flow discriminates subtle differences in facial expression. *Proceedings of the IEEE International Conference on Face and Gesture Recognition*.
5. Goodfellow, I., Bulatov, Y., Ibarz, J., Arnoud, S., & Shet, V. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE.
6. Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 915-928.
7. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
8. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
10. Szegedy, C., et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
11. Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*.
12. Li, J., & Chen, Y. (2019). Rethinking the implementation of the RAGCN: Revisiting the spatio-temporal reasoning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

13. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 3320-3328.
14. Zafeiriou, S., Zhang, Z., Kotsia, I., Zhao, G. (2017). Aff-Wild: Valence and arousal 'in-the-wild' challenge. *IEEE International Conference on Computer Vision Workshops (ICCVW)*.
15. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 886-893.
16. Cohn, J. F., Zlochower, A. J., Lien, J. J., & Kanade, T. (1999). Feature-point tracking by optical flow discriminates subtle differences in facial expression. *Proceedings of the IEEE International Conference on Face and Gesture Recognition*, 396-401.
17. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
18. Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2017). From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5), 550-569.
19. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1137-1145.
20. Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.