

Chapter 1: Introduction and Overview

1.1. Problem Definition

The lack of interpretation of non-verbal cues in social interaction is an aspect of major challenges that visually impaired people may encounter, especially in regards to facial expressions, which are another form of expression where emotions and intentions are most visibly communicated in human communication. Failure to recognize the cues can lead to some of the following problems: lack of understanding, poor communication and social isolation which can have a negative impact on the psychological health of a person and their social participation (World Health Organization, 2019).

The recognition of facial emotion is a critical issue for the designing of assistive technologies for human–computer interaction and to assist vulnerable groups, such as visually impaired individuals (Sari et al., 2019). In spite of the development of computer vision and deep learning technologies, most current assistive systems have not yet been capable of emotionally interpreting real-time data, processing in a portable device, and at a affordable price in everyday social settings.

The majority of traditional assistive technologies mainly address navigation support, object detection, or TTS (text-to-speech) conversion, whereas there are relatively fewer systems that address the understanding of the emotions in real time in social situations (Kusuma et al., 2020). In addition, despite the performance in terms of accuracy of deep learning models like Convolutional Neural Networks (CNNs) in facial expression recognition (FER) applications, the computational requirements of such models are still a concern when deployed in embedded systems with limited resources (Shehada et al., 2023).

Recent studies indicate that optimized deep learning architectures can significantly improve the accuracy and efficiency of facial emotion recognition systems while enabling real-time performance on edge devices (Hussain et al., 2023). However, there is a lack of research in

creating systems that can be placed on embedded platforms to help VIVs interpret emotions in social scenarios for which they are lightweight, accurate and affordable.

Thus, an intelligent facial emotion recognition system real-time and low cost which can accurately classify facial emotion and transform it to easy-to-interpret audio feedback is needed. This kind of system would be beneficial to social inclusion, communication effectiveness and independent interaction for visually impaired persons.

1.2. Thesis Objectives

Main Objective

The primary goal of this thesis is to create an intelligent real-time emotion recognition system, that relies on deep learning techniques, able to recognize the emotions of facial expressions and provide auditory feedback during social interactions with a visually impaired user.

Sub-Objectives

- To develop an automated computer vision pipeline using OpenCV for real-time face detection and region-of-interest (ROI) cropping from camera inputs.
- To design and train a custom Convolutional Neural Network (CNN) architecture capable of classifying facial expressions into 7 standard categories (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral).
- To mitigate dataset class imbalance by integrating dynamic class weighting techniques during model training.
- To improve model performance on difficult samples through a two-phase training strategy (hard-example fine-tuning).
- To convert recognized facial emotions into natural spoken verbal phrases using Text-to-Speech (TTS) technology.
- To build an accessible interactive web-based interface (using Gradio) that supports real-time camera evaluation and immediate audio feedback.
- To optimize the deep learning pipeline to ensure high accuracy and low latency suitable for lightweight and portable deployment.

1.3. Thesis Contribution

This thesis aims to make a contribution to the field of assistive technologies and facial emotion recognition by proposing and implementing a real-time, low-cost facial emotion recognition system that is specifically tailored to the task of assisting the visually impaired in social communication. The main accomplishments of this research are:

- End to end Integration: A unified system encompassing real-time face detection (OpenCV Haar Cascade), ROI extraction, CNN classification, and emotional voice feedback generation (gTTS).
- Advanced training strategy: Implementation of class weight balancing alongside a novel two-phase fine-tuning approach that isolates and re-trains the model on hard/misclassified facial images to boost accuracy.
- End-to-end pipeline design: Face detection, Feature Extraction, Emotion Classification and Generation of Audio Feedback in real-time.
- Accessible Auditory Output Integration: Mapping emotional classifications directly into context-aware spoken sentences (e.g., "I am feeling very happy today!").
- Web & Live Camera Deployment: Deployment of an interactive user interface using Gradio and Browser-based camera capture using JavaScript MediaDevices API.

1.4. Thesis Organization

This thesis consists of five chapters that deal with each of the stages of the research process, ranging from problem identification, system building, evaluation of the system, and finally conclusions.

Chapter 1: Introduction and Overview

Chapter 2: Background and Related Work

Chapter 3: Proposed Approach

Chapter 4: Experimental Results and Discussion

Chapter 5: Conclusion and Future Work