

Chapter 2: Background and Related Work

1.1. Background

Facial expression recognition (FER) has become an important research area within computer vision and artificial intelligence due to its role in understanding human emotions and improving human–computer interaction. The use of facial expressions as a means of communication is one of the most natural, universal and spontaneous ways by which man communicates emotions, many of which can be conveyed only through facial expression. Therefore, automatic recognition of such expressions has recently become a great interest in recent years, especially in the context of deep learning techniques (Sari et al., 2019).

Initial studies of FER used classic image processing techniques and patterns-based machine learning algorithms such as Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) to detect facial features. Traditional image processing techniques and machine learning algorithms like Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) were initially used to detect facial features. While these methods worked well in a controlled set-up, they failed to perform well in the presence of changes in light conditions, head pose, occlusion, and inter-personal variability, reducing their potential practical deployment (Kusuma et al., 2020).

Since the advent of deep learning, and specifically Convolutional Neural Networks (CNNs), FER systems have made great strides in terms of accuracy and robustness. CNN-based models can learn the hierarchical feature representations directly from raw images without manual feature extraction. They have proven to be very useful in tasks like facial emotion classification (Hussain et al., 2023), which are more complex and rely heavily on image data. At the same time, assistive technology has turned its interest to helping people with disabilities, such as people who are blind. This group is very challenged to understand the facial expression in social situations because they cannot “see” non-verbal communication. This restriction can impact communication and social inclusion (World Health Organization, 2019).

The recent studies have thus focused on the incorporation of a deep learning approach to FER into assistive devices that can transform the emotional information from the visual domain to other modalities like speech or audio

signals. These systems are designed to improve social inclusion by providing the visually impaired person with the ability to understand in real-time what is being said in a conversation. Real-time lightweight and cost-effective solutions are still under development and are not able to handle the embedded hardware platform efficiently (Shehada et al., 2023).

In this regard, the current study is a continuation of the previous work in which a deep learning-based facial emotion recognition system is proposed specifically for visually impaired users, while considering real-time performance, portability and auditory feedback.

1.2. Related Work

Facial Emotion Recognition (FER) is a field of research in recent years that has become of great importance since it addresses the understanding of human affective states and for the intelligent human–computer interaction systems. Early works have used traditional machine learning methods, and recent works have shifted to deep learning based methods, especially Convolutional Neural Networks (CNN) because they generated better results in image based tasks.

1.2.1. Traditional Machine Learning and Feature-Based Approaches

Early FER systems employed handcrafted feature extraction techniques like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) together with classical classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Random Forest. Widely studied approaches were those which were simple and computationally inexpensive.

Sikri et al. (2018) have introduced a framework for emotion detection with the help of Python and OpenCV, where the facial image is preprocessed and emotion is detected using conventional machine learning pipelines. Likewise, Mitra et al. (2018) pointed out the utility of the image processing techniques in FER system, especially coupled with structured dataset and classical classifiers.

Lee et al. (2013) showed that the use of feature extraction from the area of the face (like eyes, eyebrows, and mouth) could enhance the recognition accuracy. But they are still susceptible to changes in illumination, pose and occlusion, making them difficult to use in the real world.

To enhance the robustness, Jung (2019) tried to combine Facial Action Units (AUs) with retrieval-based classification by Multi-Layer Perceptrons. This hybrid method helped to boost performance, but it introduced a huge dependency on hand-crafted features and was not scalable to complex environments.

1.2.2. Deep Learning-Based FER Systems

The accuracy and robustness of FER systems have been greatly improved with the development of deep learning, especially CNN architectures. Instead of manually extracting features from images, deep learning methods automatically learn hierarchical representations from the raw images.

Mellouka and Handouzia (2020) gave a detailed review of deep learning based FER systems and pointed out that CNN based architecture consistently outperforms with traditional systems on the benchmark datasets. Likewise, Patil (2023) showed that CNN and hybrid CNN-SVM models have higher accuracy in more difficult scenarios like occlusion and light variation.

Onyema et al. (2021) demonstrated that the ConvNet-based models have the capability to perform well in emotion classification tasks, especially in healthcare-related applications. Their findings demonstrate the efficiency of deep learning models in boosting classification accuracy with computational feasibility.

In their work of 2022, Kirkvik did a comparison of various architectures, notably sequential CNNs, pre-trained models like ResNet50 and VGG16. The final results showed that transfer learning can boost the accuracy and decrease the training time and can be applied in resource constrained applications.

1.2.3. FER for Visually Impaired Assistance Systems

In recent years, research has focused on assistive technologies based on the FER systems for social interaction situations involving people who are visually impaired.

Kusuma et al. (2020) introduced a deep learning based FER system with the help of wearable devices, in which the facial expression is transformed to audio feedback with CNN model. The system showed some advantages in real time and achieved the accuracy of about 78-80% in real-world conditions.

Shehada et al. (2023) have created a lightweight FER system with optimized partial transfer learning for edge devices like Raspberry Pi. They attained with high accuracy

(up to 82.1%) and low computation complexity in their system, which makes it suitable for real-time assistive applications.

The studies are indeed successful to demonstrate that FER can be implemented on embedded platforms to help visually impaired users. But, the current solutions still suffer from problems in terms of accuracy, real-time performance and hardware constraints.

1.2.4. Research Gap

Although much progress has been made with the FER systems, there are still some drawbacks:

- Many deep learning models are compute intensive and cannot be deployed on low-cost devices.
- The current assistive systems are usually not capable of real-time performance in uncontrolled environments.
- Limited research has been done on end-to-end integrated systems that include detection, classification, and auditory feedback optimized for embedded platforms.

This thesis fills these gaps by proposing an optimized, two-phase trained CNN pipeline integrated with OpenCV face detection, class-weight balancing, and automated gTTS speech synthesis within an accessible interface.

Actual Label	Angry	Disgust	Sad	Surprise	Happy
Face Image					
Predicted Label	Sad	Disgust	Sad	Sad	Happy

Figure 1 Samples from the EFER2013 dataset with the actual and predicted labels