



The impact of using a deep learning-based facial expression recognition system on improving social interaction among visually impaired individuals

Prepared By

Name: Chadi Omar Elminawi

ID No.: 252001554

Submitted in fulfillment of the requirements for the degree Master of Science in
Artificial Intelligence

At

Collage of Informatics
Midocean University
UAE

Supervisor Name: Dr. Amjad Qatish

Collage of Informatics, Midocean University, Comoros

Email: 252001554@st.midocean.edu.km

Acknowledgements

Special thanks to Dr. Amjad Qatish .

Abstract

This project aims to create a deep learning facial emotion recognition system that will help the visually impaired to communicate with others socially. The system is designed to help the visually impaired community communicate with one another by correctly detecting and understanding facial expressions, in many cases, where visual cues are not feasible for a person with visual impairment. The system will be executed on an embedded platform, which is portable and accessible. It will process facial expressions in real-time with high accuracy and low latency, using advanced deep learning algorithms specifically developed for facial expression analysis. The emotions identified will be communicated to the user in audible reports, allowing the user to intuitively understand the social and emotional context during interaction. This pioneering strategy extends the social integration of people with blindness and provides an opportunity for greater inclusivity and empathy in the wider community. This project aims to give the visually impaired individual more independence in social setting and make a significant contribution to improving their quality of life through a technologically advanced solution.

Contents

Acknowledgements	2
Abstract.....	3
1. Chapter 1: Introduction and Overview.....	7
1.1. Problem Definition	7
1.2. Thesis Objectives.....	8
1.3. Thesis Contribution	8
1.4. Thesis Organization	9
2. Chapter 2: Background and Related Work.....	10
2.1. Background.....	10
2.2. Related Work	11
2.2.1. Traditional Machine Learning and Feature-Based Approaches	11
2.2.2. Deep Learning-Based FER Systems	11
2.2.3. FER for Visually Impaired Assistance Systems	12
2.2.4. Research Gap.....	12
3. Chapter 3: Proposed Approach	15
3.1. Introduction.....	15
3.2. User Needs.....	15
3.3. Requirement Specification.....	16
3.4. Mapping User Requirements to Engineering Requiremen.....	16
3.5. Overview of The System and Concept Design.....	19
3.6. Machine/Deep learning:.....	19
3.7. Hardware Platform.....	22
3.8. Camera	25
3.9. Preliminary Work	26
3.9.1. FER2013Dataset:.....	26

3.9.2. CK+ Dataset (Extended Cohn-Kanade Dataset):.....	27
3.10. System Design Overview Design Level 0.....	32
3.11. System Design Overview Design Level 1:.....	34
4. Chapter 4: Experimental Results and Discussion	36
4.1. Introduction:.....	36
4.2. Prototype Implementation	36
4.3. Prototype Verification	41
5. Chapter 5: Conclusion and Future Work.....	2
5.1. Important Features	2
5.2. Future Work.....	4
References.....	6

List of tables

Table 1 matching Marketing requirements to Engineering requirements.....	17
Table 2 Concept overview	19
Table 3 Technique comparison.....	20
Table 4 Hardware comparison.....	22
Table 5 Hardware comparison.....	24
Table 6 Camera (Strengths & Weaknesses).	25
Table 7 Comparisons of models on FER2013.....	28
Table 8 tp,tn,fn and fp of densenet121	28
Table 9 Comparision of tp,tn,fn and fp on VGGnet.....	31
Table 10 Explaining the functionality of the system.....	33
Table 11 Functionality of Camera	34
Table 12 Functionality of Raspberry Pi.....	34
Table 13 Functionality of speaker	35
Table 14 Acceptance Test.....	41
Table 15 Integration Test.....	43

List of figures

Figure 1 Samples from the EFER2013 dataset with the actual and predicted labels.....	13
Figure 2 Samples from the CK+ dataset with the actual and predicted labels.....	14
Figure 3 Deep Learning: Comparative Study.....	21
Figure 4 Camera V1 and Camera V2 respectively	25
Figure 5 sample from FER2013	26
Figure 6 A sample from CK+ dataset.....	27
Figure 7 Confusion matrix of densenet121	30
Figure 8 Confusion matrix of VGGnet.....	32
Figure 9 System design level 0.....	33
Figure 10 Detailed system design.....	34
Figure 11 Component of the system.....	36
Figure 12 MicroSD card and Raspberry Pi Imager.....	37
Figure 13 sample from dataset and the predicted output.....	38
Figure 14 The final prototype	40
Figure 15 Interface of the system	40

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 create and develop a computer vision based real-time face detection and capturing system using a normal camera.
- To build and train a deep learning model (CNN) to classify the facial expressions into predefined emotion classes with accuracy.
- To obtain the important facial features from the input images and process them efficiently for emotion recognition.
- To transform the identified emotions into an intelligible and present-day sound for the blind.
- To make the system work in different environments like lighting, background and orientation of faces.
- To portability and access the system on a low cost embedded platform (such as Raspberry Pi)
- To obtain high accuracy and low latency, required for real-time assistive applications.
- To improve social inclusion and communication effectiveness of people with visual impairment by means of simple user interface.

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:

- Real-time facial emotion recognition system using deep learning (CNN) which can classify facial expressions into the standard emotional categories with a high degree of accuracy.
- Development of computer vision methods and embedding in hardware (Raspberry Pi) for a mobile and usable assistive system.
- End-to-end pipeline design: Face detection, Feature Extraction, Emotion Classification and Generation of Audio Feedback in real-time.
- Accessible auditory output system capable of converting recognized emotions to a clear verbal feedback to the visually impaired.
- Optimization of system for low cost and low power embedded use with acceptable performance and responsiveness.
- Real time evaluation of the system to check its effectiveness to deal with changes in light, expression and user environment.
- Support to social inclusion by strengthening the capacity of visually impaired people to understand and respond to emotional situations in social interaction.

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

Chapter 2: Background and Related Work

2.1. 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.

2.2. 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.

2.2.1. 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.

2.2.2. 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.

2.2.3. 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.

2.2.4. 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.

Hence, this research seeks to fill the aforementioned gaps by creating a real-time and lightweight deep learning based FER system which is deployed on an embedded platform (Raspberry Pi) for the social interaction scenario, to aid the visually impaired.

FER studies have demonstrated the importance of deep learning methods to overcome classic challenges like variability in facial expressions, contextual or cultural differences, and technical constraints. Compared to traditional feature extraction techniques such as Local Binary Patterns and Histograms of Oriented Gradients, deep learning techniques, especially CNN, have shown better results in both emotion detection and emotion recognition. These studies show that while conventional approaches are valid, the incorporation of deep learning approaches has greatly enhanced the accuracy and efficiency of emotion recognition systems. These applications are found in different industries, such as healthcare, security, education, and customer service, highlighting the real-world applications of deep learning. Moreover, deep learning technologies like facial action units for facial reclassification and deep learning-based applications on edge devices that can enable real-time applications especially for people with visual impairments highlight the transformative power of deep learning in FER. The detailed analysis of deep learning architectures and their comparison with conventional techniques validates the effectiveness and flexibility of deep learning as the key technique for the development of FER technology.

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

Actual Label	Anger	Sad	Anger	Fear	Happy	Disgust	Surprise
Face Image							
Predicted Label	Sad	Sad	Anger	Sad	Happy	Disgust	Surprise

Figure 2 Samples from the CK+ dataset with the actual and predicted labels

Chapter 3: Proposed Approach

3.1. Introduction

The proposed approach adopted for the development of the proposed Facial Emotion Recognition (FER) system which aims to help visually challenged people to understand the facial expression by giving an auditory feedback is presented in this chapter. The chapter outlines the overall methodology adopted for designing and implementing the system starting from the identification of the subsequent requirements of the user and the system, followed by converting the requirements into engineering specifications that provide direction to the development process.

Moreover, this chapter explains the system architecture and the reasons behind the selection of the used hardware and software components in the system implementation. The selection of the embedded platform, camera module, and deep learning approaches used for facial emotion recognition is highlighted. Furthermore, a mapping relation between user requirements and engineering requirements is presented to make sure that the developed system meets the functional and non-functional objectives.

The key points of the proposed approach are very high recognition accuracy, fast response time, cost effectiveness and user-friendliness along with portability and reliability for assistive technologies. Thus, this chapter is the basis for the implementation and experimental evaluation of the subsequent chapters.

3.2. User Needs

User needs in projects are the needs, preferences and expectations of end users that are to be satisfied by the project deliverables or project outputs.

- An off-the-shelf general purpose embedded platform should be used.
- Facial expression recognition and classification with high accuracy should be possible in the system.
- The system should be easy to use: The user must take a snapshot of the person's face and then analyze the facial expression.
- The system should play back the identified emotion (s) into a speaker.
- The system needs to be flexible so that it can accurately identify a large variety of facial expressions.
- The system should be robust and accurate in detecting emotions from various people and under various illumination.
- The system should be cost-effective, both in terms of hardware and software.

- The system should possess high scanning speed that can be used to analyze facial expression quickly.
- The system should be designed to save power consumption.

3.3. Requirement Specification

- The system should be based on an off-the-shelf embedded platform as it performs better with deep learning models than other embedded platforms.
- The system needs to be able to identify and classify the facial expressions with a precision of at least 90%.
- The training data of the system should contain variety of facial expressions for the system to recognize with high accuracy.
- The system needs to scan the facial images within 3 seconds or less.
- The system shall recognize and output emotion in less than 3 seconds.
- The system should have (ON/OFF) to facilitate optimum use of energy.
- The system needs to provide the result of the recognition within 5 seconds in audible form.
- The project budget is restricted to 1000 SAR, and cannot exceed this amount.

3.4. Mapping User Requirements to Engineering Requiremen

Table 1 matching Marketing requirements to Engineering requirements

Marketing Requirements	Engineering Requirements	Justification
2,6	The system should use an off-the-shelf embedded platform, due to its superior performance with deep learning models compared to other options.	Using an off-the-shelf embedded platform simplifies the system's development and integration process while delivering the performance needed for deep learning models.
1	The system's training data should include a diverse range of facial emotion datasets (CK+, FER2013) to ensure high accuracy in recognition.	Diverse facial expression training data reduces bias and ensures high recognition accuracy across different demographics.
8	The system should accurately recognize and classify facial expressions with at least 90% accuracy.	High accuracy ensures reliable recognition, providing visually impaired users with dependable emotional cues, which improves their social interactions.
2,5	The system should perform a quick scan of facial images within 3 seconds or less.	A quick scan minimizes latency, allowing users to receive results promptly and interact with others without delays.
2,3,8	The system shall classify and give an output of recognized emotion within 3 seconds	Fast classification and output ensure that users receive timely results, enabling them to understand emotions in real time.
4	The system should audibly output recognition results within 5 seconds.	Delivering the recognized emotion audibly within seconds ensures users can respond promptly to the emotional cues.
7,9	The system should include (ON/OFF) to optimize energy consumption.	Energy-saving modes optimize power consumption and extend battery life, ensuring reliable performance for mobile or wearable systems.

7	The project budget is limited to 1000 SAR and must not exceed this limit.	Limiting the budget ensures cost-effectiveness and accessibility, while still providing high-quality performance.
---	---	---

Marketing Requirements:

1. It is very important that the system is based on a general off-the-shelf embedded platform.
2. The system should be able to detect and categorize Facial emotions with high accuracy.
3. Easy to operate system, user takes an image of a facial emotion.
4. The system should play back the facial expression recognized verbally.
5. The system must be able to detect a variety of facial expressions.
6. The system should be reliable.
7. The system should be cost-effective.
8. The system must have a high speed in scanning.
9. The system should not use excessive electricity..

3.5. Overview of The System and Concept Design

The general off-the-shelf platforms that are available in the embedded system market with an appropriate programming language should be used in our FER design. It must be able to identify facial expressions accurately, and provide the voice feedback of the recognized facial emotion, so that the visually impaired can get timely, accurate and accessible feedback.

Table 2 Concept overview

Machine Learning	Hardware Platform	Camera	Speaker
Deep Learning	Raspberry Pi	Raspberry Pi camera V1	Speaker
Image Processing	Arduino	Raspberry Pi camera V2	

3.6. Machine/Deep learning:

Machine/Deep Learning is a subcategory of Artificial Intelligence, where learning and improvement take place through experience and without explicit programming.

Table 3 Technique comparison.

Method	Strengths	Weaknesses
Deep Learning	- Excellent accuracy, especially with large datasets. - Can handle complex data and recognize patterns with high precision. - Capable of improving its accuracy with additional data.	- Requires significant training time. - Requires large datasets. - May require specialized hardware or software.
Image Processing	- Excellent accuracy, especially with large datasets. - Can handle complex data and recognize patterns with high precision. - Capable of improving its accuracy with additional data.	- May not be as accurate as deep learning-based models. - May not handle complex recognition tasks.

There are two ways to go about our project, image processing and deep learning, as it pertains to the visually impaired. In image processing, the features are extracted manually and classifiers like Support Vector Machines or k-nearest Neighbors are used to detect emotions, for example, using Histograms of Oriented Gradients or Local Binary Patterns. These methods are less complex and more readily implemented, but are less accurate due to changes in lighting, facial expression and occlusion. However, deep learning, especially CNN automatically learns features from the data, thus deep learning has greater tolerance to these differences and greatly enhances the recognition accuracy. Such complex patterns in facial expressions are hard to detect from traditional machine learning models, but can be captured by deep learning models, which are able to process large and diverse datasets. Considering the requirements of high accuracy and real-time processing in assistive technologies for visually impaired users, deep learning is the most powerful and flexible approach that can be used for our project.

So, we chose: **Deep Learning.**

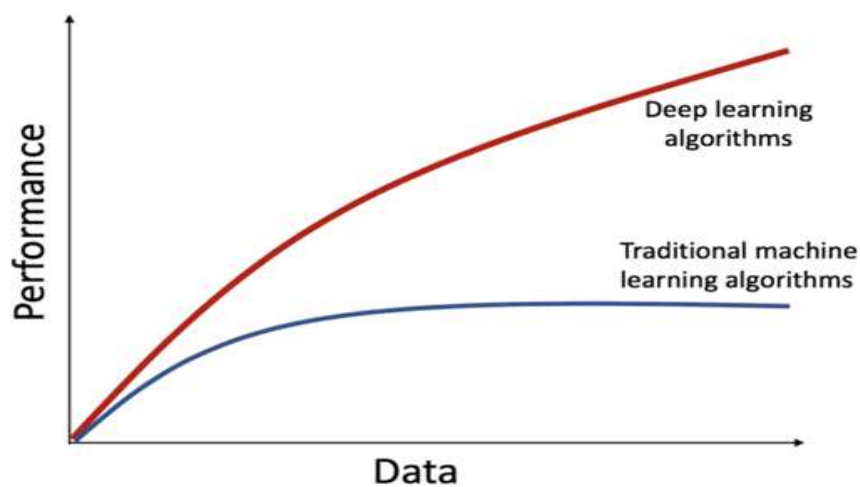


Figure 3 Deep Learning: Comparative Study

3.7. Hardware Platform

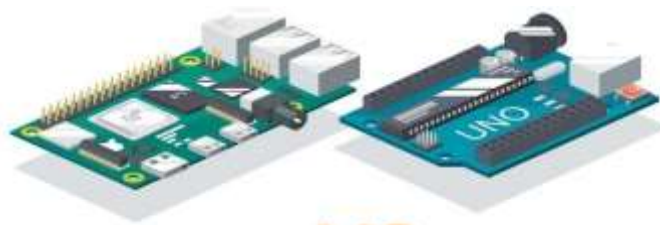
A hardware platform is a collection of hardware components that are compatible with software applications. Each program is written for a particular platform (a defined processor — and processor implementation — and associated hardware), and each platform has a machine language.

Table 4 Hardware comparison

Method	Strengths	Weaknesses
Raspberry Pi	- Low power consumption - Fast memory - great with Python - High Speed - Have his own Linux OS	- High cost - Non-upgradable memory. - Long setup .
Arduino	- Lower cost. - No need for setup. - Lower power.	- Single-tasking. - Lower speed - Doesn't have an OS. - It doesn't deal with Python or deep learning.

The Raspberry Pi and Arduino are two of the most popular hardware platforms among computer engineers, from novice to expert. Raspberry Pi is primarily used for learning computer programming while Arduino is devoted to quicker programming and circuit prototyping. We will examine the similarities and differences between these two boards, and why you should select Raspberry Pi over Arduino.

We selected Raspberry Pi because of its ability to handle multiple tasks, and it has a processor that runs Python. We needed to use a processor that could run Python for our Arabic handwritten character/digit recognition dataset module for Deep Learning.



Raspberry Pi **VS.** Arduino

Table 5 Hardware comparison.

Raspberry Pi		Arduino
- A mini-computer with its own Raspberry OS capable of handling multiple programs at a time.	Definition	- It is a microcontroller which is just a normal part of the computer. It can be used to run a single program code again and again.
- It has 4 USB ports for establishing a connection with different devices.	Number of USB ports	- It has just 1 USB port to connect to the computer.
- Python, Scratch, Ruby, C, and C++ are the primary programming languages for developing applications on the Raspberry Pi.	Languages Supported	- C or C++ programming languages can be used to program Arduino.
- 32 bits	CPU Architecture	- 8 bits
- The Raspberry Pi processor speed- 700MHz- 1.5GHz	Processor speed.	- The average speed of most of the Arduinos is 16 MHz

3.8. Camera

The camera is one of the essential hardware components that acquires the necessary information to be used in the application of the project.

Table 6 Camera (Strengths & Weaknesses).

Camera type	Strengths	Weaknesses
Raspberry Pi Camera V1	- 5 Megapixels. - Easy to install. -Lower cost than the V2.	- Lower Resolution. - requires configuration
Raspberry Pi Camera V2	- 8 Megapixels. - Easy to install.	-Thermal issues -Limited frame rate

We selected the Raspberry Pi Camera Module v2 over the v1 for our project on FER for visually impaired due to its better features. The v2 camera features an 8-megapixel sensor, which provides the higher resolution images than the v1 camera with its 5-megapixel sensor. This is an important resolution as it is vital in recognizing facial expressions accurately. Also, the v2 camera offers improved image quality and colour fidelity which brings better performance to our deep learning models. The v2 camera, although with some drawbacks, is the higher resolution and clearer camera hence better for our project.

So, we chose: **Raspberry Pi Camera V2**

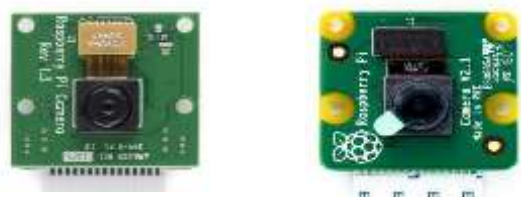


Figure 4 Camera V1 and Camera V2 respectively

3.9. Preliminary Work

The datasets are a key component of any computer vision system. A benchmark dataset enables researchers to easily and fairly compare various machine learning methods and techniques. In this project we have chosen two most popular datasets: FER2013 and CK+ (Extended Cohn-Kanade Dataset).

3.9.1. FER2013Dataset:

FER2013 is an open resource for facial expression recognition problems. It was introduced during the ICML 2013 Challenges in Representation Learning. The database contains 35,887 grayscale, 48x48 images divided into 7 facial expression classes: anger, disgust, fear, happiness, sadness, surprise and neutral. The images were gathered from multiple sources to ensure diversity in age, gender and ethnicity, which makes it a good standard for training and testing facial emotion recognition models.



Figure 5 sample from FER2013

3.9.2. CK+ Dataset (Extended Cohn-Kanade Dataset).

The widely used facial expression data set, CK+, has sequences of images of facial expressions. It contains 593 sequences from 123 participants, and has a focus on seven common expressions:

anger, contempt, disgust, fear, happiness, sadness, and surprise. The images have been marked with Action Units (AUs) and emotion labels, which makes it an excellent dataset for training and evaluation of facial expression recognition models. Data have been gathered in laboratory environment, which is a controlled condition to make the data consistent and high quality annotated data.



Figure 6 A sample from CK+ dataset

The reasons for using a Convolutional Neural Network (CNNs)

We used CNNs as a primary tool of deep learning, which are especially effective for images related tasks, for our preliminary work. CNNs are particularly well suited to learning spatial hierarchy of features from images, which can be done automatically and adaptively, and are well suited for tasks such as facial emotion recognition.

In this project we have been working on a facial emotion recognition system for visually impaired, hence we have tailored our CNN model to perform this task. This enabled us to work with the architecture and parameter optimization of the model to fit the FER2013 database, and to make changes specific to the context of facial emotion recognition.

It was decided not to use transfer learning to optimize the model for the current data set and task, as there was no prior feature set built from other unrelated data sets. Our system was further enhanced with a trained CNN model to analyse facial

expressions and give aural feedback to the users who are visually impaired.

Once they have decided on their use of Convolutional Neural Networks (CNNs), which can learn spatial hierarchy of features, and the fact that they proved to be effective in emotion recognition tasks, the next step was to choose the most appropriate CNN architecture for their project. For optimal performance, two popular models, VGGNet and DenseNet121, were compared. This comparison enabled us to analyze their advantages and disadvantages in terms of accuracy, effectiveness, and applicability for real-time applications in facial emotion recognition. We tried to determine the best model according to our system requirements, by analyzing the performance.

Table 7 Comparisons of models on FER2013

Model	Dataset	Precision	Recall	F1-Score	Accuracy
VGGnet	FER2013	0.683	0.651	0.633	0.701
Densenet121		0.579	0.576	0.575	0.576

The table below presents the result of several evaluation metrics (precision, recall, F1-score, and accuracy) of VGGnet and ResNet152 on FER2013 dataset.

True positive (TP): cases that the system has correctly identified an emotion,

True negatives (TN): indicates if the system makes the correct decision when no face or no emotion is present.

False positives (FP): Number of times the system misclassifies a no emotion as an emotion.

False negative (FN): A case in which the emotion is not detected by the system.

Table 8 tp,tn,fn and fp of densenet121

Class	Emotions	TP	TN	FP	FN
0	Angry	403	4494	438	407
1	Fear	330	4486	443	483

2	Happy	1192	3919	360	271
3	Sad	467	4162	657	456
4	Surprise	436	4947	155	204
5	Neutral	448	4390	349	555

		Confusion Matrix					
True Label	Angry	403	80	54	150	23	87
	Fear	112	330	56	178	65	66
	Happy	61	55	1192	68	30	52
	Sad	115	130	80	467	13	114
	Surprise	19	94	40	22	436	28
	Neutral	108	75	126	222	22	448
		Angry	Fear	Happy	Sad	Surprise	Neutral

Figure 7 Confusion matrix of densenet121

In terms of emotion recognition in faces, DenseNet121 shows good results, such as with the emotion "Happy", where it achieves a high true positive rate of 1,192. It is also very connected, enabling efficient re-use of features and better gradient flow, thereby boosting training speed and overall performance. Furthermore, the efficiency of the parameters in the model decreases the computational cost as compared to other deep architectures, which makes it suitable for jobs with limited memory. It does have trouble differentiating between similar emotions, however, as evidenced by higher false positive rates and confusion in the matrix, such as Sad vs. Fear. The complexity and connectivity of the model can introduce computational challenges, especially when it comes to real-time applications in the case of devices with lower processing capabilities. In addition, the quality and balance of the training set has a significant impact on its performance; for instance, classes such as "Fear" (TP = 330) result in lower accuracy. Lastly, DenseNet121 is a complex network that can suffer from overfitting on small and imbalanced datasets, which requires careful consideration of dataset preparation and model optimization to achieve the best results from DenseNet121.

The Final Model (VGGnet):

After a detailed analysis and comparison, it was clear that the accuracy and performance of VGGNet was better than that of ResNet152. This model had more satisfactory outcomes and was the most appropriate for us. Given the power of Kaggle cloud to train the algorithm and specially choosing the VGGNet model, we were able to get a very accurate and efficient solution to our needs.

Table 9 Comparision of tp,tn,fn and fp on VGGnet

Class	Emotions	TP	TN	FP	FN
0	Angry	407	4740	253	342
1	Fear	323	4681	234	504
2	Happy	1251	4048	221	222
3	Sad	510	4378	432	422
4	Surprise	479	4995	85	183
5	Neutral	792	4024	706	220

Its performance in this project is evident in the confusion matrix, which exhibits high levels of True Positive rates for discriminating between different facial expressions such as "Happy" and "Surprise," showcasing its reliability in recognizing unique facial expressions. The model performs well overall, but is room for improvement in its recognition of more nuanced emotions to further improve its performance.

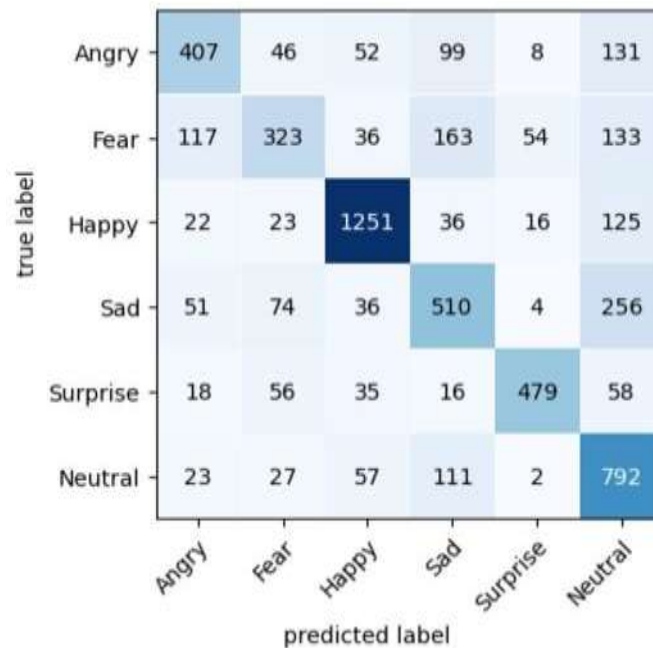


Figure 8 Confusion matrix of VGGnet

Furthermore, the structured nature of VGGNet facilitates it to be stable in recognizing and categorizing complex visual data, ensuring reliability for most emotion categories. Although results indicate minor

Overall, the model's performance highlights its strength and adaptability, making it a promising foundation for future research in creating assistive technologies for supporting social interaction for the visually impaired.

3.10. System Design Overview Design Level 0

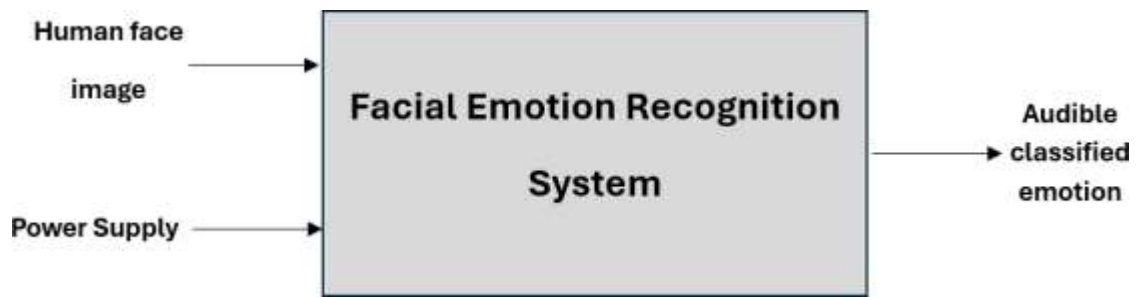


Figure 9 System design level 0.

Table 10 Explaining the functionality of the system

Module	FER System
Input	Human face image: 8Mpixels Power Supply: 220 volts AC rms, 60Hz.
Output	Audio output signal
Functionality	Capturing an image of a person's face using a camera and detecting the face, the system extracts facial features and classifies the emotion using a trained deep learning model. Finally, it outputs the detected emotion through an audio message.

3.11. System Design Overview Design Level 1:

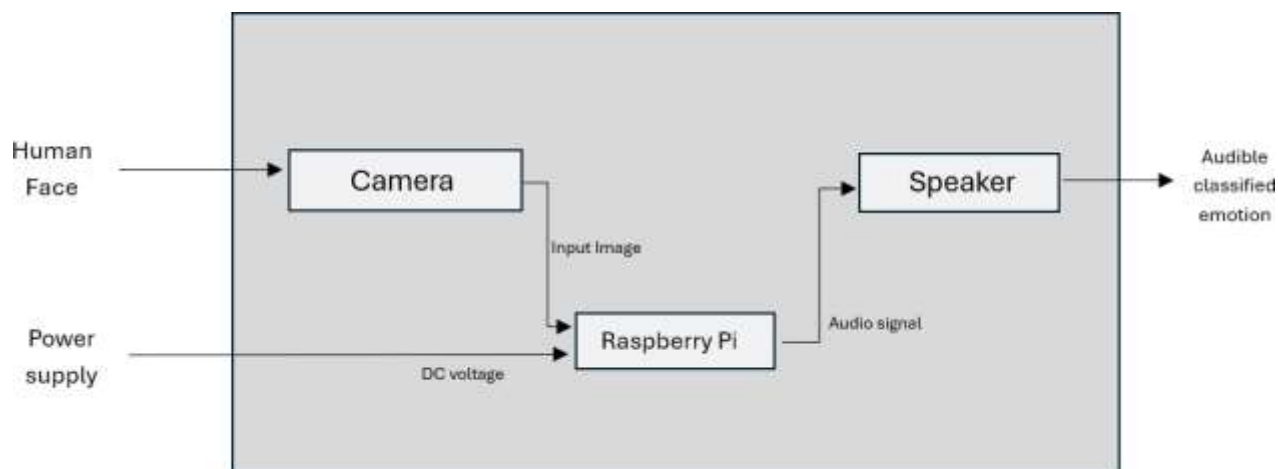


Figure 10 Detailed system design

Table 11 Functionality of Camera

Module	Raspberry Pi Camera V2
Input	Human Face
Output	Image of the human face
Functionality	The camera captures an image of a human face and then outputs it to Raspberry Pi.

Table 12 Functionality of Raspberry Pi

Module	Raspberry Pi 4 model b
Input	Image of the human face DC voltage
Output	Audio signal
Functionality	The Raspberry Pi processes the input image then performs tasks like extracting facial features, identifying emotions expressed on the human face and emotion classification then outputs the classified emotion as an audio signal.

Table 13 Functionality of speaker

Module	Speaker
Input	Audio signal
Output	Audible classified emotions
Functionality	The speaker receives this audio signal and produces an audible output, allowing the user to hear the classified emotion

Chapter 4: Experimental Results and Discussion

4.1. Introduction:

The implementation phase is the stage in which all project elements are combined, and we give details to explain the project idea. This section is intended to give a clear idea of the feasibility and applicability of the project to the desired goals. We will also be talking about our challenges that we had when implementing it, so that we can get a whole picture of our journey.



Figure 11 Component of the system

4.2. Prototype Implementation

The first step of implementing this process was to install the Raspbian operating system on a MicroSD card using the Raspberry Pi Imager program. But some Raspbian versions had compatibility problems, due to the fact that not all supported Tensorflow library. To overcome this, we've found and used a particular version of the Raspberry Pi's operating system which was compatible with Tensor flow, 64-bit.



Figure 12 MicroSD card and Raspberry Pi Imager

After the operating system was installed successfully, we installed the necessary environment and the required dependencies, such as Tensorflow, PiCamera2 and OpenCV libraries. We chose to implement a CNN using the framework Tensorflow rather than using Fastai and transfer learning because the original intention was to use the former but it proved to be too slow. The Raspberry Pi camera module was interacted with using PiCamera2. These tools all improved the functionality of the system, allowing for a strong deep learning application to be implemented.

We then used the python IDE (Thonny) to write code to test the model with images from the desktop. We initially tested the model, which was built based on a VGGNet-inspired architecture, with the FER2013 dataset and specifically took a test image of a happy face. This is when the camera and speaker had not yet been connected to the Raspberry Pi yet. When we ran the code, the output was displayed on the screen of the Thonny shell, and we could compare the model's predictions with its performance. The decision to do this was important because the model was built to take advantage of the structure of the VGGNet, so it had to be working properly and the test image was being correctly classified with respect to the emotions that it has.



Figure 13 sample from dataset and the predicted output

Once we had all the components of the project together, and the system was set up,

we tested it with real time images and noted the critical need for a stable setup to take images. To achieve accurate emotion recognition, it was crucial to ensure the alignment and centering of images. For the first couple of attempts, we tried a case we built using a 3D-printed Raspberry Pi 4 case which would securely hold all of the components except for the audio output as shown in figure 4.4. This arrangement offered a solid platform for the camera as well as other crucial parts, while keeping unwanted movement while shooting disruptions to the image. The case design enabled the exact placement and seamless integration of the Raspberry Pi, camera module and other components. This made it possible to test and continually refine the emotion recognition model in a reliable and consistent manner that was required for our project.



Figure 14 The final prototype

The user interface of the system has been designed as shown in figure 4.5 which is simple and easy to use for visually impaired people. It has been developed with Python and Tkinter library which is a basic and reliable library for creating user-friendly interfaces. The interface has basic components: it can take photographs of the people, analyze their facial expression, and give them voice feedback. It has clear buttons and audible prompts to help users follow the steps. This design emphasizes ease of use, rapidity and effectiveness, promoting the user to understand and act on feelings in real time.

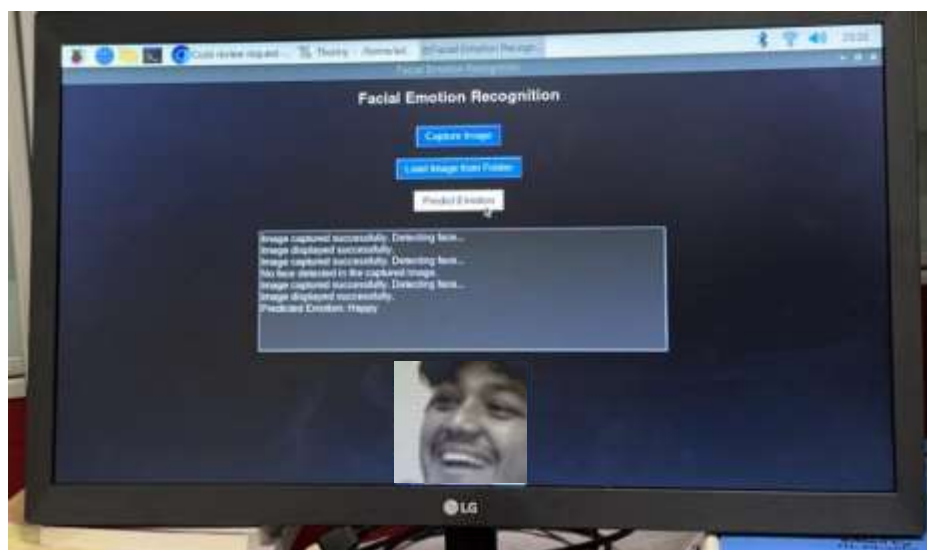


Figure 15 Interface of the system

4.3. Prototype Verification

The results of prototype testing are included in this section for the purpose of verifying the design requirements. Testing process consisted of acceptance tests and integration tests.

Table 14 Acceptance Test

Test Case Name:		Facial emotion recognition System – Camera module test	Test ID #:		FERS-01	
Description:		capture the image of emotion and return the exact value of the emotion using a speaker.	Type:		■ white box	
Setup:		Connecting camera V2, and integrating space key and with speaker to the raspberry pi4 with good lighting and clear image				
Step	Action	Expected Results	Pass	Fail	N/A	Comments
1	Connect a speaker or headphone	Sound is clear	✓			
2	Capture the image	When capturing there should be a sound of camera click	✓			
3	Listening to value note	Listening to the value of the note captured by the camera.	✓			

Overall test result:

✓

Table 15 Integration Test

Test Case Name:		Facial Emotion Recognition system – Camera with deep learning models	Test ID #:	FERS-02		
Description:		The system's main requirement is for a Deep Learning model to recognize the emotions of an individual from the camera image.	Type:	■ black box		
Setup :		Select an accurate Deep Learning model for recognizing Emotions from camera images under optimal conditions				
Step	Action	Expected Results	Pass	Fail	N/A	Comments
1	Load the deep learning model using python code	The VGGNet model should load effortlessly using Python and TensorFlow libraries on the Raspberry Pi system.	✓			Try again if its not recognized
2	Take a picture of happy	The model should accurately detect each emotion, particularly the emotion “happy” with high precision.	✓			Try again if its not recognized

3	Take a picture of sad	The model should accurately detect each emotion, particularly the emotion "sad" with high precision.	✓			Try again if its not recognized
---	-----------------------	--	---	--	--	---------------------------------

4	Take a picture of angry	The model should accurately detect each emotion, particularly the emotion "angry" with high precision.	✓			Try again if its not recognized
5	Take a picture of neutral	The model should accurately detect each emotion, particularly the emotion "neutral" with high precision.	✓			Try again if its not recognized
6	Take a picture of fear	The model should accurately detect each emotion, particularly the emotion "fear" with high precision.	✓			Try again if its not recognized
7	Take a picture of surprised	The model should accurately detect each emotion, particularly the emotion "surprised" with high precision.	✓			Try again if its not recognized
Overall test result:			✓			

Chapter 5: Conclusion and Future Work

5.1. Important Features

Each of the features proposed in the facial emotion recognition system is carefully designed and finely-tuned to ensure that the system is accurate, efficient, and practical for use in assistive technology. The features have been designed to meet the functional needs of facial emotion recognition systems as well as significant constraints of embedded systems such as computational resource, power consumption, and processing time. The overall design is based on simplicity, accessibility and reliability in uncontrolled social settings.

Real-time facial emotion recognition using deep learning methods in particular with Convolutional Neural Networks (CNNs) is one of the basic features of the system. This module allows live video stream captured using a camera module to be continuously processed, and face classification (face detection) to be performed to extract features of the face relevant to the system automatically, without requiring to manually engineer the features. The CNN model is trained to recognize facial expressions into the standard emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. In assistive scenarios, the real-time aspect of the system becomes especially significant, because emotional feedback is provided in real-time in the midst of a social interaction, thereby enabling a visually impaired user to make sense of the emotional cues and respond appropriately in a conversation.

A notable advantage of the system is the use of an embedded low-cost hardware platform – Raspberry Pi. This design consideration is crucial in maintaining the portability, compactness, and affordability of the system for various users. Using edge computing concepts, all of the image acquisition, image preprocess, inference and audio generation can be executed locally on the device. This means that, unlike cloud-based computation, the system doesn't need to be connected to the Internet at all times and that latency is significantly reduced, as is the need for system responsiveness, while at the same time providing data privacy and security. Moreover, the embedded system is suitable for use in everyday life in various environments such as social life in public and private areas.

There is also an audio based emotion feedback mechanism which converts the classified emotional states into output speech by means of a text-to-speech (TTS) engine. This is crucial for making the information available to people who may not be able to see it, as it translates what is normally seen into something that can be heard. The audio output should be concise, meaningful and clear, and should be able to let the user understand the emotions of the other person. This directly leads to the improvement of the quality of communication, the prevention of misunderstandings, the improvement of the user's social confidence and independence in interpersonal communication.

Moreover, the system is built with a high degree of robustness towards different environmental factors and generalization. In practical use, the conditions are often different: illumination will differ, the angle of view will change, there will be partial overlays (like glasses, masks or hand movements), complex or dynamic backgrounds, etc. To combat these problems, it features pre-processing methods and a deep learning framework for learning invariant feature representations. This allows the model to perform consistently and accurately in the real-world, critical for its assistive application.

Furthermore, the system emphasizes on computational efficiency and cost-effectiveness, allowing it to operate effectively within the constraints of embedded hardware. This is accomplished by adopting a proper model design, optimizing the network architecture and removing unnecessary computational complexity without sacrificing the level of accuracy. A balance is carefully struck between performance and efficiency, to ensure that emotion recognition and audio output takes place within a reasonable amount of time to respond to the real-time requirement, and is not too long to be used in interactive situations. This optimization will make the system technically feasible and economically scaleable to be widely adopted.

These features are integrated in a seamless and logical manner, and the technical infrastructure is sound, while the practical and user-friendliness is outstanding. The proposed system is shown to be an effective assistive technology solution, due to its real-time processing, embedded deployment, audio accessibility, environmental robustness, and computational efficiency. It has a profound impact to increase the understanding of emotional signals in social situations,

and to promote social inclusion, communication, and independence among people with visual impairments.

5.2. Future Work

The proposed facial emotion recognition system shows good performance in the real time emotion classification and assistive audio feedback for the visually impaired, however, there are some important directions for future improvement. The enhancements are designed to make the system more accurate, robust, adaptable and easy to use with greater complexity and real-world scenarios, and to extend the functionality of the system.

An important direction for future work is the use of more sophisticated deep learning architectures. Convolutional Neural Networks (CNNs) are effective in facial expression recognition, but recent deep learning developments like attention-based models, Vision Transformers (ViT), and hybrid CNN-Transformer networks have demonstrated greater potential to capture local facial characteristics and to capture global context relations. These models are especially suited to identify nuanced or vague emotions like fear versus surprise, sadness versus neutral face etc. The use of such structures could greatly improve the accuracy of facial expression recognition, particularly in practical applications where facial expressions are incomplete, not ideal, or noisy due to environmental factors.

Transfer learning and large pre-trained models like ResNet, EfficientNet, MobileNetV3, or lightweight and powerful models for edge devices are also promising avenues to explore. Transferring features from pre-trained networks trained on big and diverse datasets can result in better quality of features, less training time and better generalization over different populations, ethnicities, ages and lighting conditions. This method is especially useful for assistive systems in that it would not require large sets of domain-specific data with the same performance. Furthermore, adapting these models to domain specific facial emotion datasets can further enhance adaptability and robustness.

The system can be further extended to multimodal emotion recognition, a great step forward over single-modality emotion recognition systems. In this method, facial expression analysis can be integrated with other emotional cues, including speech tone and prosody, body posture and even heart rate variability. Combining several modalities can substantially enhance emotion detection accuracy, particularly when emotion is not described in the face, or ambiguous, or purposefully obscured.

Multimodal systems are also better able to withstand environmental factors and give a more holistic representation of human emotion in social interaction settings.

On the hardware side, future efforts can include deployment on more powerful and specialized edge AI hardware. The Raspberry Pi is an affordable and widely available option, however for more demanding applications and an enhanced computational efficiency, other more powerful hardware solutions are available like NVIDIA Jetson Nano, Jetson Xavier, and Google Coral TPU which allow the use of more advanced deep learning models without sacrificing real-time performance. These platforms can be optimized for hardware acceleration of neural network inference, enabling faster processing, lower latency and efficiency. These enhancements would signal its readiness for deployment in real-world dynamic conditions, as is required for continuous operations.

Moreover, there is room for an adaptive and user-friendly interface design is also possible. In future new versions of the system may have customisable options to help users change the speed, tone, language and level of detail of feedback provided by sound. Additional language support would enhance accessibility for users with varying language backgrounds. Furthermore, the integration of user preferences and adaptive learning mechanisms could allow the system to adjust its feedback over time, as it learns from the user's behavior and interaction patterns, leading to a more personalized and user-friendly experience.

Lastly, there's a need for a thorough real-life testing and long-term studies of users who have visual impairment in real social settings. Research of this kind would give valuable information on the practical feasibility, psychological effects and social impact of the system. They would also be able to uncover limitations that can't be seen in the lab, such as user fatigue, environmental changes, and reliability over time. This helps to enhance system performance, the usability of the interfaces, and the system's design based on end-user feedback. Also, user privacy, data security, and informed consent should be integrated to the future system development and evaluation process.

In general, these future improvements target the proposed system to evolve from a functional prototype to a fully scale, intelligent, and socially relevant assistive technology system, that will be able to substantially enhance the quality of life and social inclusion of the visually impaired people.

References

- Alarcos, C., Keel, S., & Kocur, I. (2019). *World report on vision*. World Health Organization.
- Alarcos, C., Keel, S., & Kocur, I. (2019). *World report on vision*. World Health Organization.
- Andrada-Livia, C., Popescu, D., & Iordache, D. (2023). New trends in emotion recognition using image analysis by neural networks: A systematic review. *Sensors*. <https://doi.org/10.3390/s23167092>
- Hawton, A., Green, C., Dickens, A. P., Richards, S. H., Taylor, R. S., Edwards, R., Greaves, C. J., & Campbell, J. L. (2011). The impact of social isolation on the health status and health-related quality of life of older people. *SpringerLink*. <https://doi.org/10.1007/s11136-010-9717-2>
- Jung, S.-W. (2019). Facial action units-based image retrieval for facial expression recognition. *IEEE*.
- Kirkvik, E. B. (2022). *FER using deep learning*. Oslo Metropolitan University.
- Kusuma, H., Attamimi, M., & Fahrudin, H. (2020). Deep learning-based facial expressions recognition system for assisting visually impaired persons. *Bulletin of Electrical Engineering and Informatics*, 9(2).
- Lee, H.-C., Wu, C.-Y., & Lin, T.-M. (2013). Facial expression recognition using image processing techniques and neural networks. *National Quemoy University*.
- Mellouka, W., & Handouzia, W. (2020). FER using deep learning: Review and insights. *ScienceDirect*. <https://www.sciencedirect.com/>
- Mitra, A., Choudhury, S., & Moitra, S. (2018). Image processing facial expression recognition. *RCC Institute of Information Technology*.

Onyema, E. M., Shukla, P. K., Dalal, S., Mathur, M. N., Zakariah, M., & Tiwari, B. (2021). Enhancement of patient facial recognition through deep learning algorithm: ConvNet. *Journal of Healthcare Engineering*.

Patil, D. D. (2023). *FER using deep learning*. California State University, Northridge.

Sikri, M., Gupta, A., Puri, R., Goel, S., Pathak, N., & Tiwari, M. (2018). Emotion detection using image processing in Python. In *Computing for Sustainable Global Development*.

Shehada, D., Turkey, A., Khan, W., Khan, B., & Hussain, A. (2023). A lightweight FER system using partial transfer learning for visually impaired people. *IEEE Access*, 11.

Hussain, A., Khan, B., & Turkey, A. (2023). A lightweight facial expression recognition system using partial transfer learning for visually impaired people. *IEEE Access*, 11, 1–9.

Kusuma, H., Attamimi, M., & Fahrudin, H. (2020). Deep learning-based facial expressions recognition system for assisting visually impaired persons. *Bulletin of Electrical Engineering and Informatics*, 9(2), 1–12.

Sari, M., Moussaoui, A., & Hadid, A. (2019). Automated facial expression recognition using deep learning techniques: An overview. *International Journal of Informatics and Applied Mathematics*, 15(2), 1–15.

Shehada, D., Turkey, A., Khan, W., Khan, B., & Hussain, A. (2023). A lightweight FER system using partial transfer learning for visually impaired people. *IEEE Access*, 11, 1–9.

Hussain, A., Khan, B., & Turkey, A. (2023). *A lightweight facial expression recognition system using partial transfer learning for visually impaired people*. *IEEE Access*, 11, 1–9.

Kusuma, H., Attamimi, M., & Fahrudin, H. (2020). *Deep learning-based facial expressions recognition system for assisting visually impaired persons*. Bulletin of Electrical Engineering and Informatics, 9(2), 1–12.

Sari, M., Moussaoui, A., & Hadid, A. (2019). *Automated facial expression recognition using deep learning techniques: An overview*. International Journal of Informatics and Applied Mathematics, 15(2), 1–15.

Shehada, D., Turkey, A., Khan, W., Khan, B., & Hussain, A. (2023). *A lightweight FER system using partial transfer learning for visually impaired people*. IEEE Access, 11, 1–9.

World Health Organization. (2019). *World report on vision*. WHO Press.