1 Motivation

In today's world, where technology is a big part of our daily lives, the ability of machines to understand and respond to human emotions is an important area of development to enhance human-computer interaction. **Facial Emotion Recognition (FER)** is a step in this process, as it gives machines the capability to interpret human emotions through facial expressions. [1]

FER systems are not just about new technology; they are relevant in many areas, such as healthcare, education, and customer service. In healthcare, these systems can help track patients' emotions, which is important for mental health treatments [2]. In education, they can be used to see how engaged and understanding students are, which could lead to teaching methods that adapt to each student's needs [3]. In customer service, understanding the emotions of clients can greatly improve how companies interact with their customers, but also offer potential for upselling and retention strategies. [4]

One of the main challenges of FER is accurately recognizing emotions in a wide variety of people, each showing emotions in their own way. This requires a system that works well for everyone. This project therefore aims to develop a system using a **Convolutional Neural Network (CNN)** that accurately classifies facial expressions into distinct emotions while also being sensitive to the variety of facial expressions. By effectively training a CNN to perform this task, the project aims to contribute to the field of affective computing, enhancing the capability of machines to interact with humans in a more natural and intuitive manner.

2 Learning Task

The project's primary objective is to develop a FER system employing **supervised learning** techniques. In this approach, the model will be trained on a labeled dataset where each image of a face is tagged with a corresponding emotion. The task entails classifying these facial images into predefined emotional states like happiness, sadness, anger, fear, surprise, disgust, and a neutral state. Details on the training experience, the learning task and an applicable performance measure are elaborated in the following.

2.1 Training Experience

To develop a reliable FER system, the training experience will largely depend on the quality and diversity of the dataset used. The complexity of this task arises from the need to accurately interpret a wide range of human facial expressions, which can vary across individuals and cultures. Furthermore, the model must be robust to variations in image quality, lighting, background, and facial obstructions, such as glasses or facial hair. Therefore, for training a robust model, it is important that the images are sourced from a diverse population, have different lightning and background, contain typical facial obstructions, and provide a broad range of facial expressions.

Figure 1 shows three faces of the dataset *FER-2013* with three different facial expressions. In the picture, faces of the categories fear, happy and neutral are represented. Figure 2 displays sample images from the *RAF-DB* dataset with faces sorted into several categories.

Figure 1: Three faces, each labeled by facial expression. Image Source: https://www.kaggle.com/datasets/msambare/fer2013

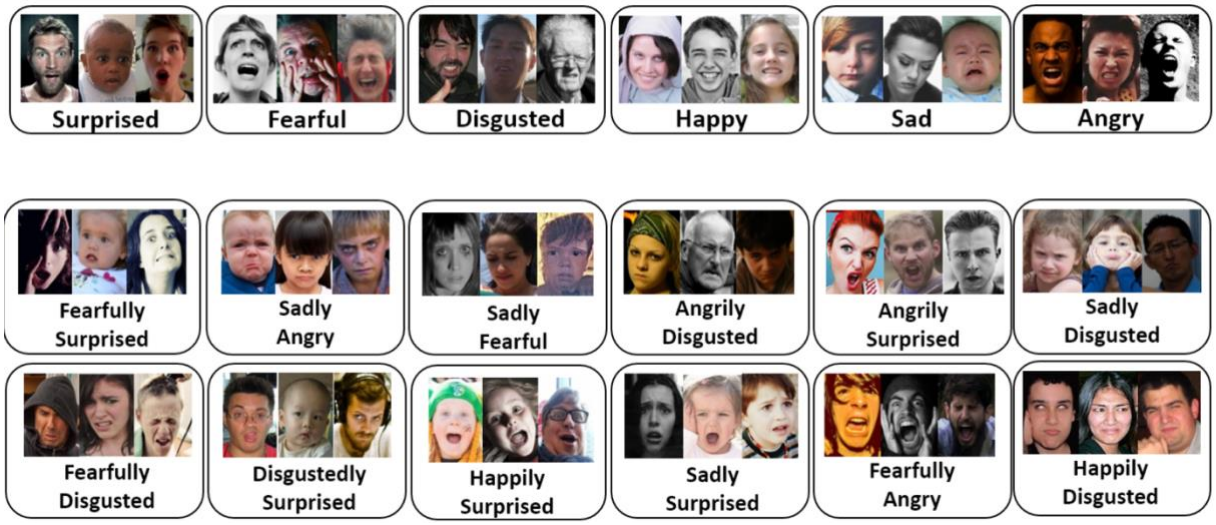

Figure 2: Several images of faces categorized by facial expression from single-label and two-tab subsets. Image source: http://www.whdeng.cn/raf/model1.html

There are several extensive datasets, containing images with different facial expressions available. Examples are the following:

The *FER-2013* dataset seems to be widely recognized and a comprehensive dataset in this field. It contains 48x48 pixel grayscale images of faces, each labeled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset contains 28709 examples for the training set and 2589 examples for the public test set. [5]

The *AffectNet* dataset contains around 440 thousand images, labeled with one of eight expressions. Those expressions are the same as in the *FER-2013* dataset plus one more category: contempt. Additionally, the images are also labeled by intensity of the facial expression. [6]

The *RAF-DB* dataset contains about 30 thousand images with large diversity. It contains two subsets, one “single-label subset, including 7 classes of basic emotions [and one] two-tab subset, including 12 classes of compound emotions”. [7]

The *Aff-Wild* and *Aff-Wild2* datasets are known for their real-world focus, featuring over 1 million frames from videos, annotated for valence and arousal in the *Aff-Wild*, and expanded annotations like action units in *Aff-Wild2*. These datasets are well suited for training FER systems under naturalistic conditions. [8] [9]

The *SFEW* dataset, derived from movie clips, includes static images categorized into seven basic emotions. It contains more than 1700 images and was used for realistic scenario testing in the *EmotiW* challenge. [10]

In addition to utilizing a pre-labeled dataset, the data can be augmented to improve the model's generalizability. This augmentation could include variations in lighting, orientation, and background noise, simulating real-world conditions where facial expressions are not always presented in ideal environments. This would ensure that the training experience encompasses a wide array of scenarios, reinforcing the model's ability to adapt and perform accurately in diverse settings.

2.2 Learning Task

This project falls under the category of supervised learning in deep learning. The reason for choosing CNN is their ability in image recognition tasks. CNNs are adept at automatically and efficiently detecting features in images, such as facial features and expressions, without the need for manual feature extraction. This capability makes them suitable for this task.

The core objective of this project is therefore to design and train a CNN-based model to classify input images of human faces into basic emotion categories. This task requires the model to accurately interpret facial features and expressions and distinguishing subtle differences between emotions. The input will either be a grayscale or RGB image and the output will be the corresponding determined emotion category. The project will involve several steps:

1. Preprocessing the images to ensure they are in a format suitable for CNN input.
2. Designing a CNN architecture that can effectively learn features from facial images.
3. Training the model on a dataset, with careful monitoring of its learning progress.
4. Validating the model's performance on a separate subset of the dataset to ensure it generalizes well to new data.

It must be noted that the system can only work reliably on images and for the emotion categories it was (sufficiently) trained on. The more the images differ from the ones the model was trained on the less accurate it can be.

2.3 Performance Measure

To evaluate the effectiveness of the FER system, various key metrics can be employed, each graded from 0 to 1, where 0 is the lowest (worst) and 1 the highest (best) score. The main metric, accuracy, measures the percentage of the model's predictions that are correct, indicating its ability to interpret facial emotions correctly. Alongside accuracy, precision and

recall provide additional insight into the system's performance. Precision assesses how often the system correctly identifies positive instances, helping to minimize false positives. Recall measures the system's capability to capture true emotional expressions, ensuring no significant expressions are missed.

The F1 Score, which harmonizes precision and recall, offers an overview of the system's overall performance. This metric is useful in cases where the representation of emotions in the dataset is not uniform, helping to maintain a balanced approach in recognizing different emotions. A well-performing FER system demonstrates consistent performance across various emotional expressions, indicating its adaptability.

Literature has reported high benchmarks in FER system accuracy, with one study achieving 99.43% accuracy. This level of accuracy was obtained by integrating several models and approaches, combining multiple databases, and implementing data augmentation techniques. These results were achieved under specific conditions that included extensive resources and sophisticated methodologies, which might not be feasible in the planned project [11]. Therefore, a realistic approximation of performance might aim for an accuracy in the range of 80-85%. This target considers the complexity inherent in emotion recognition and the challenges of working with varied datasets. Achieving this level of accuracy could be considered as successful.

3 Plan

The execution of the FER project will follow a structured approach, including several phases from initial data preparation to the final evaluation of the model.

Data Preparation and Preprocessing

The initial phase involves gathering and augmenting various data sources. This preprocessing step includes normalizing the images and resizing them to a uniform size suitable for CNN input. The dataset will likely be augmented to include variations in image orientation, lighting, and background noise, and then split into training, validation, and test sets to separate data for training the model and evaluating its performance.

Model Design and Selection

In the model design and selection phase, a CNN architecture will be designed for facial emotion classification. This involves selecting the number and types of layers, activation functions, and other hyperparameters. Popular deep learning libraries such as *TensorFlow* or *PyTorch* may be utilized for model implementation.

Model Training and Validation

During model training and validation, the training data will be fed into the model, with weights adjusted through backpropagation and optimization algorithms used to minimize loss. Hyperparameter tuning will involve experimenting with various learning rates, batch sizes, and other parameters to find the most effective combination. The model's performance will be

periodically evaluated on the validation set to monitor progress and make necessary adjustments.

Performance Evaluation

Finally, in the performance evaluation phase, the completed model will be assessed on a test set using metrics such as accuracy, precision, recall, and F1 score. An error analysis will be conducted to understand areas of difficulty and to identify potential improvements.

4 Related Work

In the field of FER using CNNs, their already exist several studies, papers, and projects. The approach mentioned in [12] proposes a novel FER model using a four-layer CNN, focusing on detecting seven specific emotions. This model's use of Local Binary Pattern (LBP) and Oriented FAST and rotated BRIEF (ORB) for feature extraction, along with its high accuracy on various datasets like *FER-2013*, *JAFFE*, and *CK+* [13]. Adapting its approach, especially its multi-layered CNN architecture and feature extraction techniques, could inform the model design for static image analysis.

Another project discusses the use of the *FER-2013* dataset for emotion recognition. This dataset, consisting of grayscale images of faces labeled with emotions, could be relevant to the planned project. The described project's focus on accurately classifying facial expressions into distinct emotions aligns closely with the planned project's objectives. The methodologies employed in this research, especially in handling the *FER-2013* dataset, can provide a reference point for the model development. [14]

Additionally, the paper "Real Time Emotion Recognition from Facial Expressions Using CNN Architecture" demonstrates a LeNet-based CNN model, merging three datasets to achieve high accuracy in classifying seven emotions. An accuracy of 96,43% was achieved for real time emotion recognition [15]. This approach could also provide a foundation for the planned emotion recognition system.

The review of existing projects and studies in FER using CNNs shows the potential of these technologies in emotion recognition. Adapting the multi-layered CNN architecture and sophisticated feature extraction methods from these studies could be used to implement a competent FER system. While comprehensive replication of these advanced models might not be feasible, integrating specific elements, such as simplified versions of their CNN structures or feature extraction techniques, might be applicable. The project aims to build a model that is both effective in its classification accuracy and feasible within its scope, using these studies as a reference point.

5 Risk Management

This section outlines the main identified risks associated with the project and the contingency plans to address them.

Data Quality and Diversity: One identified risk involves the quality and representativeness of the dataset. If the used dataset lacks diversity or contains biases, the model may perform poorly in real-world scenarios. To mitigate this, the dataset will be evaluated for diversity and balance. If needed, additional datasets will be used to enhance the representativeness of the training data. Data augmentation techniques may also be employed to enhance the variability of the training data. Since extensive databases with quality images are available this seems like a low risk.

Model Overfitting and Underfitting: Overfitting is a common risk in deep learning, where the model performs well on training data but poorly on unseen data. On the other hand, underfitting occurs when a model is too simple to capture the underlying patterns, affecting performance on both training and unseen data. To address this, techniques like cross-validation, regularization, and dropout layers in the CNN architecture may be employed. The model's performance on the validation set will also be closely monitored and the complexity of the model adjusted as needed.

Computational Constraints: Given the resource-intensive nature of training deep CNNs, there's a risk of computational limitations, especially if the project requires more extensive data or more complex models than anticipated. As a contingency, the CNN can be designed to be scalable, allowing for adjustments in size and complexity based on available computational resources. Cloud-based computing resources will also be considered if local resources are insufficient. Furthermore, the potential use of transfer learning, where a pre-trained model on a similar task is adapted for this specific project, could be explored.

Performance Metrics Not Meeting Expectations: If the model fails to achieve the desired performance in terms of accuracy, precision, recall, and F1 score, the model architecture and training process will be re-evaluated. This might involve experimenting with different CNN architectures, adding more layers or filters, or adjusting hyperparameters like the learning rate.

If the project progresses quickly and demonstrates high performance, further enhancements could be explored. These may include implementing real-time processing capabilities for dynamic emotion recognition or extending the model to recognize emotions of multiple faces in a single image.