

Emotion Recognition using Deep Learning Techniques

Jakob Tormalm

HKUST

Clear Water Bay

hj.tormalm@connect.ust.hk

Abstract

This project explores facial emotion recognition (FER) using deep learning techniques. A ResNet18 model is trained on the FER2013 dataset to classify facial expressions into seven emotions: angry, disgust, fear, happy, sad, surprise, and neutral. The system is designed for efficient, real-time performance on Apple M1 hardware. Preliminary experiments show the model reaching around 68% accuracy on the test set after 10 epochs, with strong results for the happy and surprise classes. A live webcam demo using OpenCV has also been implemented, capable of detecting faces and predicting emotions in real time. Future work will focus on improving class balance, tuning hyperparameters, and enhancing robustness under different lighting and pose conditions.

1. Introduction

Facial Emotion Recognition (FER) is a rapidly developing area of computer vision that aims to automatically detect and classify human emotions from facial expressions. It has broad applications across domains such as human-computer interaction, mental health monitoring, education, and entertainment. The ability for machines to understand human emotions can significantly improve the user experience by enabling more empathetic and adaptive systems.

Despite the progress made with deep learning methods, FER remains a challenging task. Variations in facial structure, occlusion, illumination, pose, and cultural differences can cause significant performance degradation. Moreover, emotion categories such as “fear” and “surprise” or “sad” and “neutral” can appear visually similar, further complicating classification. Therefore, FER serves as a compelling problem that balances theoretical depth with practical impact.

In this project, I will build a deep learning model capable of recognizing facial emotions from static images. The system is trained using the FER2013 dataset and later applied in a live webcam setting, allowing real-time emotion detection.

The primary objective is to explore how convolutional neural networks (CNNs), specifically a ResNet18 architecture, can effectively learn discriminative features from facial data, achieving strong accuracy while maintaining computational efficiency suitable for real-time deployment on a MacBook M1.

This project also emphasizes the practical application of deep learning models. Beyond training and testing, I will deploy the model in a simple interactive setting using OpenCV, where the system captures frames from the webcam, detects the user’s face, resizes it to 48×48 pixels, and predicts the emotion in real-time. This bridges the gap between theoretical coursework and a real-world, user-facing application.

2. Problem Statement

The goal of this project is to design and implement a deep learning-based facial emotion recognition system capable of classifying images into seven emotion categories: *angry, disgust, fear, happy, sad, surprise, and neutral*. The model should generalize well to unseen data and be able to operate in a real-time webcam application.

The dataset used for training and evaluation is the FER2013 dataset, which contains 35,000 grayscale facial images of size 48×48 pixels. Each image is labeled with one of the seven emotion classes. The data is split into training and test sets following the original dataset’s structure. Although the dataset provides a strong foundation, it is known to contain noisy labels and unbalanced class distributions, which make training and generalization more difficult.

To mitigate overfitting and improve robustness, standard preprocessing techniques such as normalization and data augmentation (random horizontal flips, small rotations, and random cropping) will be applied. The expected outcome is a classification model achieving at least 70% accuracy on the FER2013 test split, which is competitive for lightweight CNNs. Evaluation will be quantitative, using metrics such as accuracy, F1-score, and confusion matrices.

This project’s final deliverable includes a trained model

and a live demo system using the webcam feed, where the detected face is automatically cropped, resized, and passed through the trained ResNet18 model for emotion prediction.

3. Technical Approach

The technical pipeline consists of four main stages: data preprocessing, model design, training and evaluation, and deployment.

Data Preprocessing: The FER2013 dataset is organized into `data/train` and `data/test` directories, each containing subfolders for the seven emotion classes. Images are loaded using the PyTorch `ImageFolder` class. Each image is normalized to have zero mean and unit variance. The training set is used exclusively for training, and the test set is reserved for final evaluation.

Model Architecture: A ResNet18 model pretrained on ImageNet is used as the backbone, chosen for its strong feature extraction capabilities and efficient size, which make it suitable for real-time applications. The final fully connected layer is modified to output seven emotion classes. This lightweight yet powerful model structure helps achieve high accuracy without requiring extensive computational resources.

Training Procedure: The model is trained using the cross-entropy loss function and optimized with the Adam optimizer (learning rate = 0.001, weight decay = 1×10^{-4}). Training is conducted for 10 epochs with a batch size of 32 using the MPS backend for GPU acceleration on the MacBook M1.

Evaluation Metrics: Model performance is assessed on the test set using overall accuracy, class-wise F1-scores, and confusion matrices. Additionally, qualitative visualizations of the model's predictions are generated to understand failure cases, such as confusion between visually similar emotions.

Deployment: For real-time testing, OpenCV's `VideoCapture` is used to read webcam frames. Each frame is converted to grayscale, and a Haar Cascade classifier detects the face region. The detected face is resized to 48×48 pixels and normalized before being passed through the trained model to predict emotion probabilities. The top predicted emotion is then displayed on the video feed in real-time.

4. Intermediate / Preliminary Results

At this milestone stage, we have completed data preprocessing, model setup, and initial training runs. The ResNet18 model achieves promising results on the test set:

- Training accuracy: 78%
- Test accuracy (preliminary): 68.3%

The detailed classification report on the test set is as follows:

Table 1. Classification Report on the Test Set

Class	Precision	Recall	F1-score	Support
Angry	0.655	0.564	0.606	958
Disgust	0.624	0.658	0.640	111
Fear	0.579	0.474	0.521	1024
Happy	0.862	0.876	0.869	1774
Neutral	0.600	0.681	0.638	1233
Sad	0.544	0.574	0.559	1247
Surprise	0.787	0.833	0.809	831
Accuracy			0.683	7178
Macro Avg	0.664	0.666	0.663	7178
Weighted Avg	0.681	0.683	0.680	7178

From these results, the model shows strong performance particularly in recognizing *happy* and *surprise* emotions, with F1-scores above 0.8. Emotions such as *fear*, *sad*, and *angry* remain more challenging, reflected in their lower recall and precision. The *disgust* class is the smallest in support, which likely affects its performance.

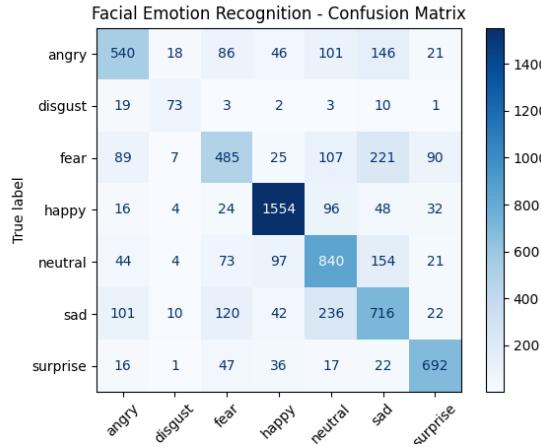


Figure 1. Confusion matrix on the test set illustrating per-class prediction performance.

The confusion matrix further highlights common misclassifications between similar emotions such as *fear* and *sad*, which will be a focus for improvement in subsequent work.

We have also successfully integrated a functional real-time webcam demo using OpenCV. The system detects faces from the live video stream, preprocesses them to 224×224 , and predicts the emotion label live on the MacBook M1 hardware.