Emotion Recognition using Deep Learning Techniques

Jakob Tormalm
HKUST
Clear Water Bay
hjtormalm@connect.ust.hk

Abstract

Facial Emotion Recognition (FER) aims to classify human affective states from facial expressions using computer vision and deep learning. This project develops a ResNet18-based FER pipeline trained on the FER2013 dataset and deployed in a real-time webcam system. FER2013's low-resolution grayscale images make the task challenging, especially for subtle and visually overlapping emotions such as fear, sad, and angry. To address these difficulties, the project incorporates preprocessing, augmentation, fine-tuning, ablation studies, and a full deployment pipeline using OpenCV. Experiments show that the model provides competitive accuracy while maintaining real-time inference speed on a MacBook M1. The final system demonstrates how to translate deep-learning methodology into a practical interactive application.

1. Introduction

FER is an established task in computer vision with applications in assistive technology, affective computing, education platforms, safety monitoring, and interactive entertainment. Despite advances in deep learning, FER remains challenging due to occlusions, identity differences, pose variation, cultural diversity in expression, and subtle differences between certain emotion classes. Low-resolution datasets such as FER2013 intensify these difficulties.

This project builds a complete FER system focusing on efficient architectures suitable for real-time settings. It investigates the performance trade-offs of a compact ResNet18 backbone, explores augmentation strategies for noisy data, examines common misclassification patterns, and deploys a functional, responsive webcam system. The emphasis is not only on model accuracy but also on use-case practicality and interpretability.

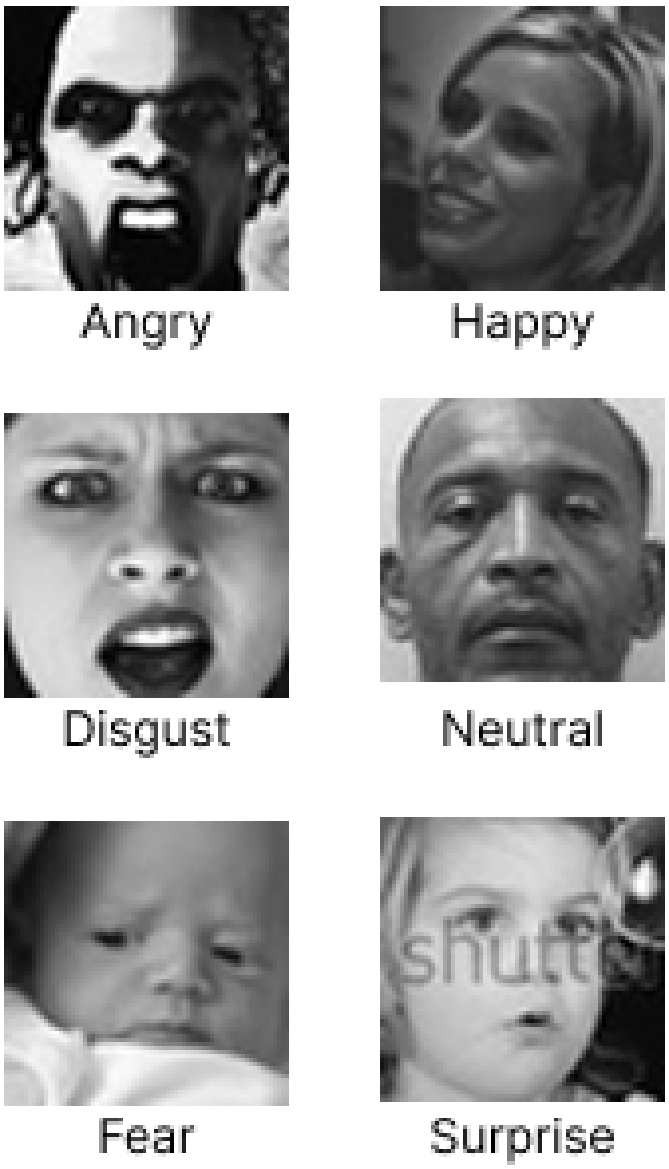

Figure 1. Challenging FER2013 Samples

2. Related Work

Traditional FER solutions employed handcrafted descriptors such as Gabor filters, HOG, or LBP combined with SVM-based classification. These methods struggled with illumination changes, intra-class variability, and non-frontal faces. The shift to deep learning enabled more effective feature extraction and robustness through end-to-end training.

CNN-based architectures such as VGG, Inception, and ResNet have become the standard foundation for modern FER systems. ResNet-like backbones in particular have shown strong performance due to their ability to mitigate vanishing gradients and capture hierarchical features efficiently. Several works integrate attention mechanisms or multi-branch modules to address subtle facial cues. For example, Residual Attention Networks [**?**] and Spatial Transformer [**?**] modules have been used to emphasize critical facial regions such as the eyes, eyebrows, and mouth.

More recent systems explore Vision Transformers (ViTs) [**?**], which model global attention across the entire face. Although these models often outperform classical CNNs on datasets such as RAF-DB or AffectNet, they are significantly more computationally heavy, making them unsuitable for real-time applications. Temporal models combining CNN backbones with LSTMs or 3D CNNs have also been proposed for video-based FER.

In contrast to these large architectures, this project favors an efficient ResNet18 model to maintain real-time inference on consumer hardware while achieving competitive accuracy on FER2013.

3. Data

FER2013 provides 35k labeled images of size 48×48, each assigned to one of seven emotion categories. The dataset was originally introduced as part of the 2013 Kaggle Facial Expression Recognition Challenge. Images were gathered by querying emotion-related terms on Google Image Search and other public image sources, followed by an automated screening pipeline and human annotation. This collection pipeline introduced substantial noise: many images are low resolution, incorrectly cropped, or contain ambiguous expressions.

The dataset is divided into training, validation, and test splits, but contains known issues. Several samples exhibit extreme lighting, motion blur, or partial occlusions. Additionally, annotator disagreement is common for subtle emotions, contributing to label noise. Another key challenge is class imbalance: *disgust* is severely underrepresented, while *happy* appears far more frequently. This imbalance impacts training stability and leads to biased predictions unless mitigated.

A more detailed analysis reveals that facial expression intensity varies widely. Some images present exaggerated, prototypical expressions, while others contain mild or culturally atypical expressions that are harder to classify. Since the images are grayscale, color-based cues are unavailable, forcing the network to rely only on texture and shape features.

Preprocessing includes:

- Normalization to zero mean and unit variance.
- Data augmentation: horizontal flips, small rotations, random crops, and slight brightness variations.
- Resizing to 224×224 when using pretrained CNN backbones.

Augmentation plays an essential role due to the dataset's noise and small image size. It improves robustness to real-world conditions encountered during webcam deployment.

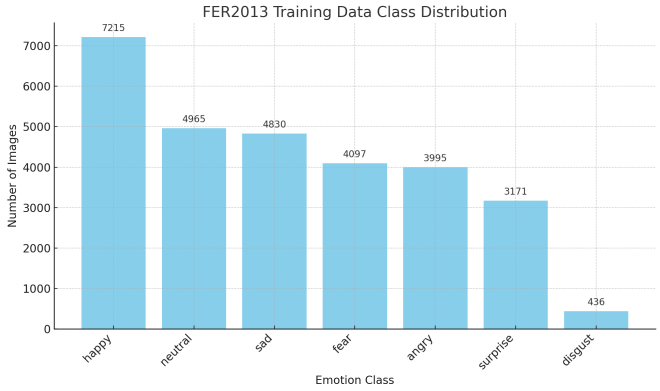

Figure 2. FER2013 Class Distribution

4. Methods

4.1. Model Architecture

The project uses ResNet18, pretrained on ImageNet. Only the final fully connected layer is modified for seven output classes. This architecture balances depth, feature richness, and speed. Earlier layers capture edges and textures, while deeper layers specialize in higher-level emotional cues. Fine-tuning allows the model to adapt to facial-expression patterns despite the low resolution of FER2013.

4.2. ResNet18 and Transfer Learning in Detail

ResNet18 contains 18 layers with residual connections [**?**], which address the vanishing gradient problem by enabling gradient flow through identity shortcuts. Each block learns a residual function, allowing the model to refine features without overwriting previously learned representations. This property is particularly important for FER, where subtle differences between emotions may require fine-grained adjustments.

Pretraining on ImageNet provides strong general-purpose filters in early layers (e.g., edge detectors, texture patterns) that transfer effectively to FER. Although

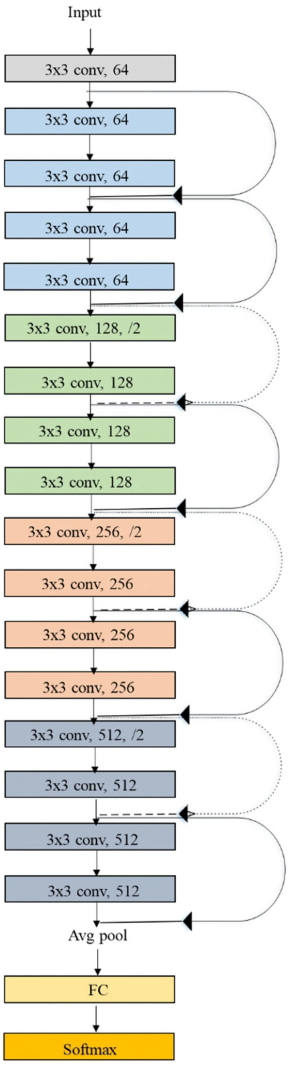

Figure 3. ResNet18 Architecture Diagram [**?**]

FER2013 images are grayscale, replicating the single channel into three channels allows the pretrained filters to remain usable. During fine-tuning, deeper layers adapt to emotion-specific cues such as eyebrow curvature, eye aperture, or mouth symmetry.

Initially, all convolutional layers are frozen, and only the classifier head is trained. Subsequently, the last two residual blocks are unfrozen for fine-tuning. This staged training strategy stabilizes optimization and prevents catastrophic forgetting. ResNet18 was chosen over deeper variants (ResNet34/50) due to the risk of overfitting on FER2013 and the real-time constraints imposed by webcam deployment.

4.3. Training Protocol

The model is trained for 10 epochs using:

- Adam optimizer (lr $= 0.001$)
- Weight decay $= 1 \times 10^{-4}$
- Batch size 32
- MPS acceleration on MacBook M1

Augmentation reduces overfitting and helps the model generalize to real webcam inputs.

4.4. Evaluation Metrics

Evaluation uses accuracy, macro-F1, weighted-F1, and confusion matrices. Macro metrics are emphasized due to class imbalance. Qualitative analyses examine common misclassifications and model attention through heatmaps.

4.5. Deployment Pipeline

The live system uses OpenCV:

1. Capture frames from webcam.
2. Use Haar Cascade to detect face bounding boxes.
3. Crop, resize, normalize, and feed to the ResNet18 model.
4. Display predicted emotion on-screen.

Figure 4. Real-Time FER System Screenshot

5. Discussion and Future Work

5.1. Limitations and Future Directions

While the compact ResNet18 backbone achieves a favorable balance between accuracy and efficiency, several limitations remain. Primarily, the model lacks temporal awareness, which restricts its ability to leverage motion cues inherent in facial expressions. Integrating temporal models such as LSTMs or 3D CNNs could capture dynamic emotional transitions, potentially improving recognition robustness, especially in video-based settings. Moreover, the exclusive reliance on grayscale images limits the utilization of color cues that might enhance discrimination of subtle expressions. Future work might explore multimodal inputs, including audio or physiological signals, to complement visual information. Additionally, emerging architectures incorporating lightweight attention mechanisms or transformer-based modules offer promising directions for better feature focus without compromising real-time performance.

5.2. Importance of Data Augmentation

Data augmentation proved crucial in mitigating overfitting and enhancing generalization to real-world conditions. Beyond standard augmentations such as flips, rotations, and brightness adjustments, more advanced techniques like CutMix [**?**] or MixUp [**?**] could be employed to synthetically enrich the dataset and address class imbalance, particularly for underrepresented emotions like *disgust*. Augmentation strategies that simulate occlusions or pose variations would further bolster the model's resilience to challenging input scenarios encountered during deployment.

5.3. Real-Time Deployment Considerations

Real-time deployment presents unique challenges beyond pure model accuracy. Latency constraints necessitate lightweight architectures and efficient preprocessing pipelines, such as the Haar Cascade face detector chosen for its speed despite lower cropping precision. Environmental factors including variable lighting, background clutter, and diverse camera quality also impact system robustness. To address these, future iterations could incorporate model compression techniques like pruning or quantization [**?** **?**] and explore optimized inference on embedded or mobile hardware platforms. Such advancements would enable broader adoption of FER systems in consumer devices and assistive technologies where computational resources are limited.

6. Experiments

6.1. Quantitative Performance

Initial training yields:

- Training accuracy: 78%
- Test accuracy: 68.3%

The confusion matrix shows strong performance on *happy* and *surprise*, while *fear*, *sad*, and *angry* often overlap. The *disgust* class suffers due to small sample size.

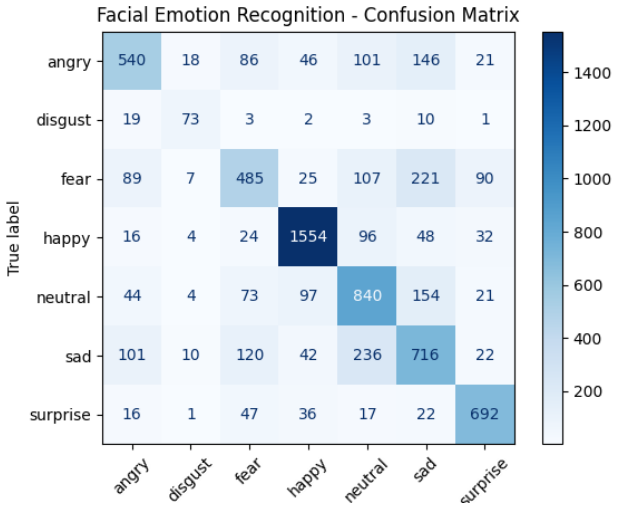

Figure 5. FER2013 Confusion Matrix

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

6.2. Ablation Studies

Three ablations were performed:

(1) With vs. without augmentation: Removing augmentation reduces test accuracy by 3–4%.

(2) Haar Cascade vs. DNN face detector: The DNN detector improves cropping quality but reduces real-time FPS. Haar Cascade remains optimal for interactive use.

(3) Input resolution: Feeding 224×224 resized images improves accuracy slightly but slows down inference.

6.3. Failure Case Analysis

Misclassifications often arise from:

- Occlusions (hands, hair, masks)
- Side-profile faces
- Dim lighting
- Ambiguous expressions

Grad-CAM reveals that the model often fixates on the mouth region, which may explain confusion between similar classes when mouth shape lacks detail.

Figure 6. Grad-CAM Visualization of Model Attention

7. Conclusion

This project demonstrates that a compact ResNet18 architecture can achieve respectable FER performance on FER2013 while supporting real-time deployment on consumer hardware. The model handles expressive emotions well but struggles with subtle or low-resolution categories. Improvements could include more robust face detectors, larger and cleaner datasets, attention mechanisms, transformer-based architectures, or temporal modeling using video sequences.