Real-Time Facial Emotion Recognition with ONNX and OpenCV

Niloofar Karimi
Northeastern University
Email: karimi.ni@northeastern.edu

***Abstract*—This paper presents a real-time facial emotion recognition system implemented in C++ using an ONNX-based deep learning model and OpenCV. The system detects human faces from a live video stream using the Haar cascade classifier and standardizes inputs through face alignment based on eye localization. The model classifies emotions into seven categories: angry, disgust, fear, happy, sad, surprise, and neutral. To enhance inference reliability, several techniques are incorporated, including test-time augmentation (TTA) via flipping and rotation, confidence-based filtering (predictions with confidence below 0.2 are labeled as "uncertain"), and temporal smoothing to stabilize predictions. The system is evaluated on static test images for classification accuracy and qualitatively assessed on live input for responsiveness and stability.**

I. Introduction

One of the key channels for human communication is facial expressions, they present different emotions such as anger, happiness, sadness, and more. Accurate recognition of these expressions and emotions is significant in various domains including mental health assessment, computer vision development, education, human-computer interaction, and customer experience enhancement. The project aims to develop a real-time emotion recognition that classifies facial expressions into seven classes Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral after detecting face from a live webcam feed. The classification result (class name and confidence percentage) is overlayed directly on the video stream.

The system combines computer vision and deep learning techniques, and the main pipeline consists of several stages: face detection, face alignment, preprocessing, emotion classification, and visual overlay. The model uses OpenCV's Haar cascade classifier (`haarcascade_frontalface_default.xml`) for face detection which is widely used for face localization in real-time. After faces are detected, they are preprocessed using grayscale conversion, normalization, and using OpenCV haarcascade-eye classifier, faces are aligned based on eye position. Thus, these steps enhanced the overall performance and offered more accurate and consistent classification.

A convolutional neural network (CNN) based on Mini-Xception architecture trained on the FER-2013 dataset which is a benchmark dataset of 35,000 48×48 grayscale facial images across seven emotion classes, then used for classification. The original images 48×48 grayscale, are then resized to 64×64 grayscale because the model requires that input size. This trained CNN is exported to the ONNX (Open Neural Network Exchange) format which resulted in `mini_xception.onnx` and used for optimized inference in real-time using OpenCV's DNN module. This ONNX model is the core of classification and there will be no need of deep learning framework when deploying.

Lastly, to enhance the system robustness and reliability in live streams, we used temporal smoothing, test-time augmentation (TTA) using flipping and rotation, and a confidence-based filtering approach that classifies the low-confidence predictions as "Uncertain." The focus of this project was the accuracy of recognition and consistency of the prediction that can be enhanced further using various tunning methods.

II. Related Works

Facial expression recognition (FER) has been improving in terms of robustness and accuracy due to its significance in the computer vision field and substantial studies had been done including three below studies:

1) **Deep Attentive Center Loss (DACL) for FER:** this method which improves the feature discrimination was introduced by Qi et al., it's an integration of attention mechanism with center loss. Their approach achieved high performance on complex tasks like AffectNet and RAF-DB, they focused on important facial regions and ensuring compactness of classes. Overall, the study highlights the effectiveness of attention-based models in enhancing real-world FER tasks.
2) **Self-Difference Convolutional Neural Network (SD-CNN):** to address variations within class in FER, this framework was introduced by Zhang et al. Using conditional generative adversarial networks, the framework generates synthetic expressions for the same subject. Then the real and generated expressions are compared to each other, the model reduces identity-related mismatches, as a result the recognition accuracy will improve. This method emphasizes the significance of synthetic data to improve accuracy in FER systems.
3) **Hybrid CNN and SVM Approach for FER:** introduced by Kim et al., this approach is a combination of two significant models in facial expression classification, CNN with Support Vector Machine. This system extracts the dynamic facial motion features and geometric landmarks and achieved high accuracy. This study emphasizes on significance of combining the traditional and deep learning for FER enhancements.

These three studies helped us to understand the importance of design choices; our project uses face alignment by eye localization and cropping relates to important facial region mechanism. Also, combining the trained CNN model with temporal smoothing and confidence-based filtering highlights the significance of model's combination for better accuracy.

III. Methods

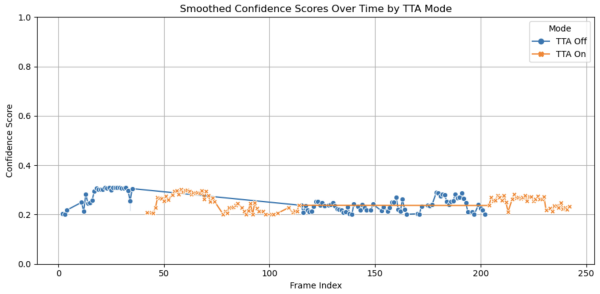
Our project integrates computer vision techniques for face detection and deep learning for emotion classification to propose a real-time facial emotion recognition system. The first task for implementation is to use OpenCV for capturing video frames from webcam, the grayscale conversion is performed on captured frames. The second task begin with passing grayscale frames to a face detection in OpenCV's Haar cascade classifier (`haarcascade_frontalface_default.xml`).

After face detection, face alignments are performed using eye localization, in this task a separate Haar cascade for eyes is used to detect the position of eyes. This alignment can improve accuracy through ensuring consistent face orientation. These aligned face regions are then center-cropped, resized to 64x64 pixels, normalized to the range [-1, 1], and reshaped to the format expected by the neural network.

The second task is emotion classification using the lightweight convolutional neural network CNN based on the Mini-Xception architecture and trained on the FER-2013 dataset. The facial images in the dataset are labeled with one of the seven emotions used for classification. After exporting the model in ONNX format, the model is loaded and executed directly using OpenCV's DNN module, the inference will be efficient and independent of framework in runtime (does not rely on TensorFlow or PyTorch).

We applied the following enhancements to improve robustness of classification:

- **Test Time Augmentation (TTA):** this approach generates the flipped and rotated formats of the input face and calculate the average of SoftMax outputs.



TTA and Standard Comparison in Confidence

- **Temporal Smoothing:** this will help the system to have less jumps from one frame another, the emotion recognition will be more stable and smoother.
- **Confidence Filtering:** the model calculates the confidence with the name of predicted class in video stream, so this approach will classify the prediction with lower confidence below 0.2, this will reduce the false positive rate and week predictions.

The final output is a detected face with green frame around it, with predicted class name and the confidence score. To evaluate the, a toggle key 'T' is used to switch to TTA during execution.

IV. Experiments and Results

This step is mostly the evaluating and experimenting on different setups and testing if various approaches can enhance the accuracy, stability, and overall performance. The system was evaluated on both offline testing data from `FER-2013` dataset and live webcam evaluation:

- **Initial baseline:** pretrained Mini-Xception in ONNX was loaded, converted to grayscale, and resized to 64x64 without no other enhancements or smoothing. The accuracy on test images was 18.7%, and major wrong classifications in live stream. The problem was expectation mismatch with model's trained input.
- **Preprocessing enhancements:** normalization applied and scaled pixel to range [-1,1], applied center-cropping before resizing, accuracy was improved to 25.2%.
- **Face alignment:** applied face alignment using Haar cascade for eye detection, and other adjustments enhanced accuracy to 50.4%.
- **TTA temporal smoothing and confidence filtering:** enhanced the accuracy rate to 61.6% in test images and live stream emotion recognition. The confidence scores improved from 10–20% to above 30%. The live stream is consistent and less jumps from one frame to another and the system can recognize between similar classes like Anger and Disgust better than before.

TABLE I
Accuracy Improvements by Enhancement Step

Configuration	Accuracy (%)
Baseline (Grayscale + Resize)	18.7
+ Normalization + Center Crop	25.2
+ Face Alignment	50.4
+ TTA + Smoothing + Filtering	61.6

```
Per-Class Accuracy:

Anger     : 52.40% (502/958)
Disgust   : 32.43% (36/111)
Fear      : 35.55% (364/1024)
Happiness : 0.00% (0/1774)
Neutral   : 54.01% (666/1233)
Sadness   : 0.00% (0/1247)
Surprise  : 61.97% (515/831)
```

Per-Class Accuracy before any enhancement and alignment

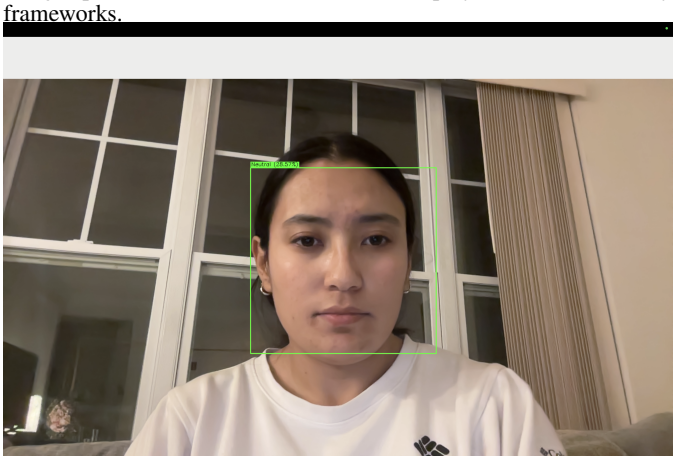
```
Per-Class Accuracy:

Anger     : 52.40% (502/958)
Disgust   : 32.43% (36/111)
Fear      : 35.55% (364/1024)
Happiness : 83.93% (1489/1774)
Neutral   : 54.01% (666/1233)
Sadness   : 56.70% (707/1247)
Surprise  : 61.97% (515/831)
```

Per-Class Accuracy after any enhancement and alignment

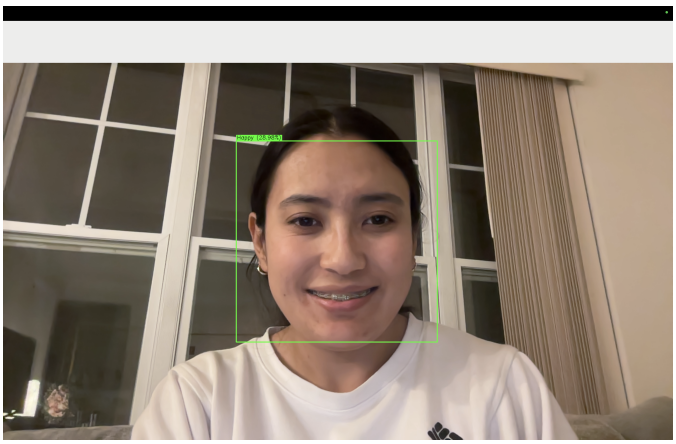
V. Discussion and Summary

The proposed system for real-time emotion recognition for both live video and test images achieved a reliable result; after applying enhancements and approaches for better prediction, the accuracy improved from 18% to 61%. Approaches such as Test Time Augmentation TTA and Temporal smoothing were applied to address issues such as noise, lighting variation, and inconsistent face orientation. Overall, TTA was helpful by averaging the predictions across images/frames which are flipped or rotated. Temporal smoothing reduced the jumps and flickers from one frame/class to another.

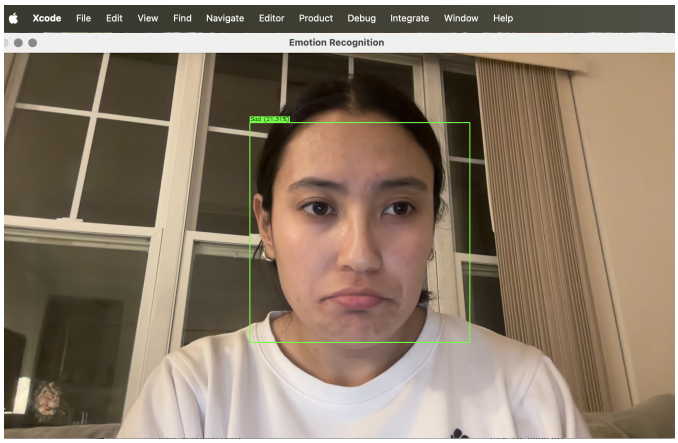
The final system achieved all the goals specified in the proposal. The system provides a modular real-time emotion recognition pipeline that combines classical computer vision with deep learning using OpenCV and ONNX for efficient deployment without heavy frameworks.



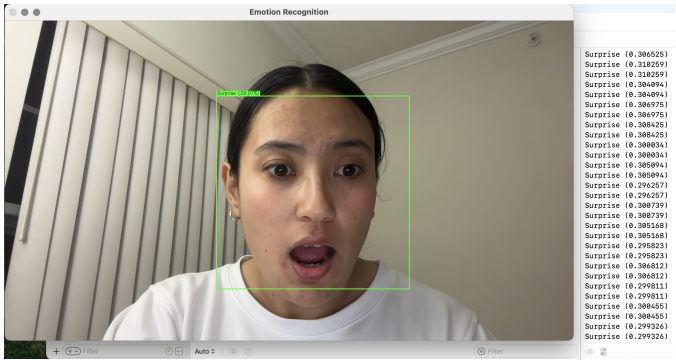
(a) Neutral



(b) Happy



(c) Sad



(d) Surprise

References

[1] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv:1710.07557*, 2017.

[2] E. Barsoum *et al.*, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ICMI*, 2016.

[3] FER-2013 Dataset. [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[5] A. Mollahosseini *et al.*, "AffectNet," *IEEE Trans. Affective Computing*, 2016.

[6] Y. Tang, "Deep learning using support vector machines," in *ICML Workshop*, 2013.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.