Facial Emotion Recognition:Comparison Analysis

Syed Rayyan Adil* Sameen Jawaid* Hafiza Amna Sadiq*

k173904@nu.edu.pk k173868@nu.edu.pk k173679@nu.edu.pk

Ijlal Shiekh* Ahmed Shakeel*

k172398@nu.edu.pk k173642@nu.edu.pk

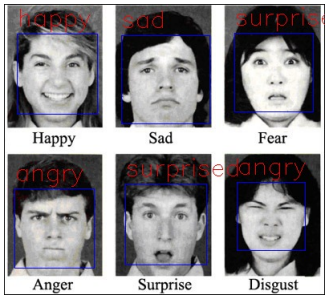

Fig. 1: Emotions to be Detected in FERC

***Abstract*—The growing world of technology demands Facial Emotion Detection is applied in several areas such as safety, education and health. In this study we identified different algorithms of FER using pre-trained Deep Convolution Neural Networks which will trained on data set. Features are extracted and transferred to a for classification. We performed different algorithms of FER to get the best finding with maximum accuracy. The results**

I. Introduction

Face detection is popular among numerous application whether in smartphones or gadgets. The change in expression on a human's face which highlights emotional states or a person's intent. An upturn in the world of neural networks has improved numerous object detection algorithms, especially the facial emotion detection and face detection.[2] Artificial Intelligence has played a major role here as emotional AI is a technology that responds to human facial expression as well as emotions. The states which will be identified using facial emotion recognition (eg; sad, fear, disgust, happy, neutral, anger, contempt) are based on particular facial images.[3] Our main challenge is to get the automatic facial emotion state with high accuracy which makes it more complex to categorize them. As an example, some attributes like skin, color, age, and the environment vary the emotions of a person.

II. Methodology

Convolutional neural network (CNN) is the most popular way of analyzing images. CNN is different from a multi-layer perceptron (MLP) as they have hidden layers, called convolutional layers. The proposed method is based on a two-level CNN framework. The first level recommended is background removal [1], used to extract emotions from an image. Here, the conventional CNN network module is used to extract primary expressional vector (EV). The expressional vector (EV) is generated by tracking down relevant facial points of importance. EV is directly related to changes in expression. The EV is obtained using a basic perceptron unit applied on a background-removed face image. In the proposed FERC model, we also have a non-convolutional perceptron layer as the last stage. Each of the convolutional layers receives the input data (or image), transforms it, and then outputs it to the next level. This transformation is convolution operation, as shown in 2. All the convolutional layers used are capable of pattern detection. Within each convolutional layer, four filters were used. The input image fed to the first-part CNN (used for background removal) generally consists of shapes, edges, textures, and objects along with the face. The edge detector, circle detector, and corner detector filters are used at the start of the convolutional layer 1. Once the face has been detected, the second-part CNN filter catches facial features, such as eyes, ears, lips, nose, and cheeks. The second-part CNN consists of layers with 3 × 3 kernel matrix, e.g., [0.25, 0.17, 0.9; 0.89, 0.36, 0.63; 0.7, 0.24, 0.82]. These numbers are selected between 0 and 1 initially. These numbers are optimized for EV detection, based on the ground truth we had, in the supervisory training dataset.[4] Here, we used minimum error decoding to optimize filter values. Once the filter is tuned by supervisory learning, it is then applied to the background-removed face (i.e., on the output image of the first-part CNN), for detection of different facial parts (e.g., eye, lips. nose, ears, etc.) To generate the EV matrix, in all 24 various facial features are extracted.[5] The EV feature vector is nothing but values of normalized Euclidian distance between each face part.

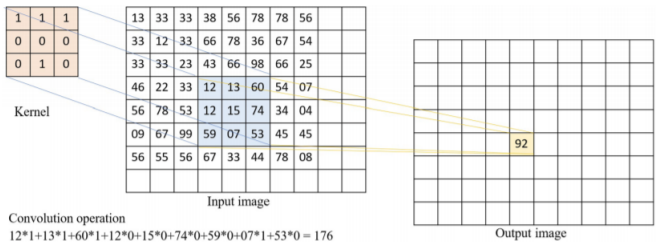

Fig. 2: Convolution filter operation with the 3 × 3 kernel. Each pixel from the input image and its eight neighboring pixels are multiplied with the corresponding value in the kernel matrix, and finally, all multiplied values are added together to achieve the final output value

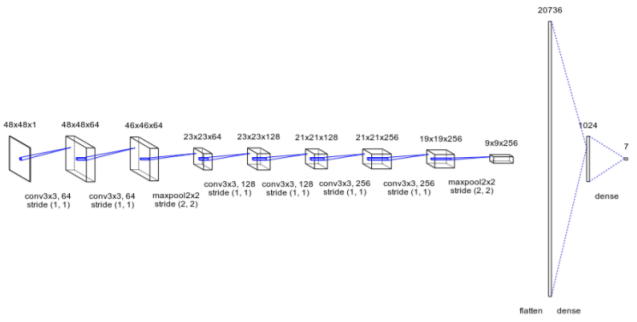

Fig. 3: Final Model Architecture

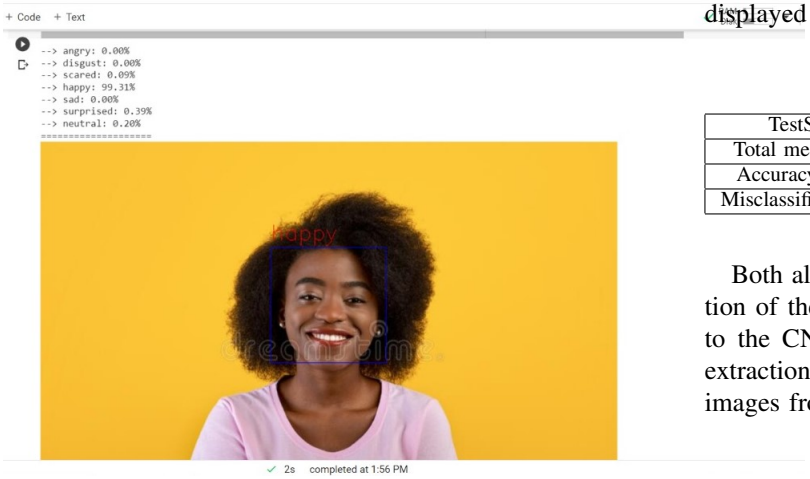

Fig. 4: Detected emotions on a face

III. Results

In the results derived from the detection algorithms, two main properties were considered, speed and, detection rate. Speed was tested on a Lenovo IdeaPad 320s laptop using the built-in web camera (as mentioned in the methods and materials section) for the haar-cascade based Viola-Jones algorithm and the deep-learning based algorithm pre-trained with ResNet. The cascade features and the ResNet weights were provided from OpenCV and GitHub.[2] The same laptop web camera was used in section 6.4 for real-time detection. When using the classifiers for live detection, there was no noticeable slowdown to the eye when using the DL approach compared to the haar-cascade based approach, as they only differed by about 2 fps, however, the DL approach appears to be faster in this case. One must consider that the Viola-Jones algorithm only performs detection on grayscale images, therefore, the image converting step is included in the time taken to perform the detection.

The final CNN model reached an accuracy of 81.67after running for 80 epochs. The training process took about 8-9 hours where each step took 359 ms and approximately 1100 steps (one step = one image) were taken in each epoch. On the test sets, the model performed strongest on the AffectNet

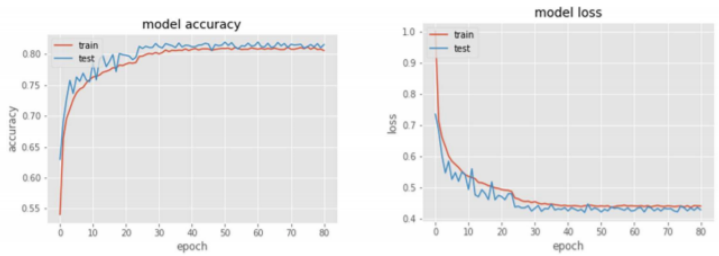

Fig. 5: Accuracy and Loss Graph

test set (1410 images) and performed below training/validation accuracy on the FER2013 test sets (2155 and 2099 images). The average from 5 runs on each of the test accuracy are displayed in I.

TABLE I: Test Accuracy of CNN

TestSet	AffectNet	FER2013(Public)	FER2013(Private)
Total mean acc	0.8132	0.7204	0.7414
Accuracy Best	0.8216	0.7560	0.8031
Misclassified Best	250	640	51

Both algorithms failed some detections when in the extraction of the ROI process when preparing the images for input to the CNN. The DL algorithm that was used for the final extraction of the ROIs failed about detection in about 3000 images from the 75 000 total images given from the dataset.

IV. Future Work

FER is a problem formed from a union of many other problems that together form a FER system pipeline, including but not limited to, image preprocessing, computer vision and AI. In this work, the goal was to achieve solutions and/or partial solutions to the subproblems described back in the problem formulation in section 4. Improvement in any of these subtasks would benefit the study of FER, as all the parts impact the overall practice in the art. In the pursuit of this, some of the results prove promising, giving some new insight in not that well-studied areas of interest, while other subtasks failed. The CNN accuracy did not achieve what is considered state-of-the-art accuracy (see related works) and was performed on a smaller set of classes than most of the state-of-the-art algorithms. Due to the time constraint and the lack of manpower and hardware, more experimentation with hyperparameters, architectures and different datasets were limited and would have been beneficial. For future studies, more data would always be beneficial, not only to the classification process but also the detection stage. As mentioned previously, all deep learning models benefit from a higher quantity of quality data. To use more GPUs and more powerful hardware is also a great way to improve the training time of algorithms, hence giving more time for the trial and error algorithms, getting more feedback from the training sessions on what needs tuning. All machine learning models evaluated in this thesis work benefits from higher accuracy of CNN's when performing the feature extraction, some more than others. Without a doubt, better

feature extraction would mean better classification capabilities for all the tested machine learning models. Finally, given the constant increase in computational power, the great work of data gathering from the computer vision community and the recent increase in popularity and accuracy of CNN's, FER will very likely have many real-time application areas in the future.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[2] Prudhvi Raj Dachapally. Facial emotion detection using convolutional neural networks and representational autoencoder units. *arXiv preprint arXiv:1706.01509*, 2017.

[3] Ninad Mehendale. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(3):1–8, 2020.

[4] Jinesh Mehta, Eshaan Ramnani, and Sanjay Singh. Face detection and tagging using deep learning. pages 1–6, 02 2018.

[5] Vedat Tümen, Ömer Faruk Söylemez, and Burhan Ergen. Facial emotion recognition on a dataset using convolutional neural network. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5, 2017.