# Emotion Detection from Facial Expressions in Pakistani Videos

Ifrah Siddiqui
msds18002@itu.edu.pk

Maham Nasir Khan
msds18041@itu.edu.pk

Mirza Elaaf Shuja
msds18051@itu.edu.pk

**Abstract-- Over the years, emotion detection has amassed high attention in the research community. Recent architectures of deep neural networks have significantly improved emotion detection. Although there are very efficient emotion detection algorithms out there based on various deep learning and machine learning approaches, but the problem of uneven distribution of data in the publicly available datasets has long been overlooked in the research community. These datasets are usually biased, i.e the datasets are race specific. Our goal is to solve this problem of emotion recognition for Pakistani people by collecting our own dataset and training an efficient deep neural network based on Convolutional Neural Networks (CNNs) and and to do a comparative study of the models trained on pakistani faces dataset and western dataset.**

**Keywords--Emotion Detection, Deep Learning, Facial Expression Recognition, Local Binary Pattern (LBP), Facial Action Coding System (FASC)**

## 1. Introduction

Human beings communicate their emotional state through facial expressions. Understanding the emotional state of a person by looking at his face is a fundamental human trait and it plays a vital role in our social interactions. The research community in the field of human-computer interaction, computer vision and deep learning has been actively taking part in emotion detection. In a country like Pakistan that is still in its developing phases, automating emotion detection will have the biggest impact in the public awareness and security fields, since the public would be able to see beyond what's being told to them. Recent research challenge regarding Emotion Detection also depict the global research community's rising interest in the matter. Convolutional Neural Networks (CNNs) have given amazing results for this problem. The convolution, pooling and layered architecture for local and global feature learning make CNN a contender in this emotion recognition based on facial expressions [1].

## 2. Related Works

Fernandez et. al [1] proposed an attention based deep neural network architecture for facial expression recognition. The proposed architecture first uses an encoder-decoder style network and then a convolutional network for feature extraction to obtain the attention map. Finally the classifier classifies emotions based on the representation of the attention map. This model has been trained using traditional datasets like CK+ and BU3DFE. Zheng et. al [2] proposed a novel deep neural architecture for single shot face detection, which is fast and capable of detecting faces with large scale variations (especially tiny faces). Levi et. al[3] has solved the problem of varying photometric transformations for emotion detection by proposing an approach to map the image intensities to 3D-space. In this model, the image intensities are first converted to local binary pattern (LBP) codes and then these coded values are mapped to a 3D-metric to be fed as an input to the CNN. Li et. al [4] has used facial action coding system (FACS) and uniform LBP to extract facial features from the images, these features are then fed to K-nearest neighbors (KNN) classifier to classify the facial expressions.

Francois et. al [5] proposed Xception architecture which is a modification of inception architecture. In the proposed model, the inception module has been replaced by depth-wise separable convolutions to improve classification performance. Christian et. al [6] has used multiple striding and pooling layers in the CNN architecture to extract local features from the whole image efficiently. Shan et. al [7] has done a survey on the existing publicly available datasets and the proposed algorithms for facial expression recognition. This paper can be used by us to compare our results to the results reported by recent papers. Xiaojie et. al [8] has proposed a detector-in-detector network to capture multi-level features from an image. Their model first detects human body from the image and then further pay attention to the body parts in a coarse to fine manner.

## 3. Methodology

In this section we discuss our methodology for the problem at hand:

### 3.2. Dataset collection

A dataset of 500 clips from Samaa News was collected. Each clip was under a minute. Frames were extracted from the video via a python script and 1 in every 50 frames is selected. Figure 1 shows the images captured from our video dataset. These images were annotated by random people with the help of Google forms (50 forms of 20 images each, were made to get a diverse response keeping in mind the time constraints). 8 annotations of each image were received to ensure quality. At the end, the most popular label was picked. Sample results of the annotations from the Google forms are shown in figure 2.


Figure 1. Sample data frames from news clips dataset

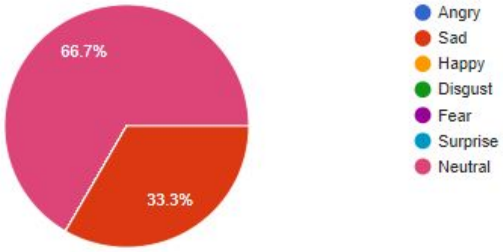



Figure 2. Sample annotations of a frame

These images are to be classified for the emotions as in Table 1:

**Table 1. List of Emotions**

| |
|---|
| **Happy** |
| **Sad** |
| **Neutral** |
| **Angry** |
| **Disgust** |
| **Fear** |
| **Surprise** |

After carefully evaluating the annotated dataset, it was observed that the dataset was heavily biased around "Neutral" emotions. The main reason being that the clips were mostly of Pakistani politicians. Politicians, in general, keep their expressions neutral, thus the bias. In order to cater this, frames were extracted from the "Suno Chanda" TV series, since actors are very expressive. This helped us in overcoming the bias. Figure 3 shows a few samples from the "Suno Chanda" dataset.

Figure 3. Sample data frames from Suno Chanda dataset

### 3.2. Basic Algorithm

After frame extraction from the input image, faces were detected using the Haar Cascade classifier which was imported from OpenCV. After face detection, a mini-Xception model was trained on the FER2013 dataset and then tested for a validation set from the FER2013 itself and a test set that comprised of images from the collected datasets.

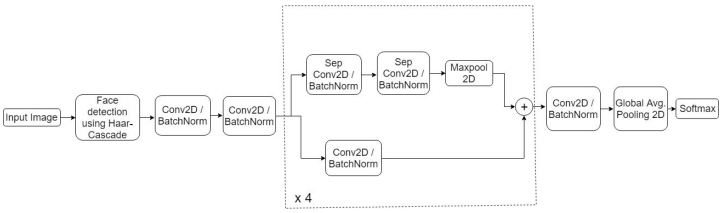



Figure 4. Architecture of the proposed model

### 3.3. Neural Network Training

A pre-trained (on FER2013) mini-Xception model was evaluated to serve as a base model[1]. The FER2013 dataset consists of 35,887 grayscale images of 48x48 dimensions. This is a highly diverse dataset containing images from various age groups and ethnicities, both posed and unposed. Due to its diverse nature, a model trained on FER2013 performs fairly well when tested on our datasets. This model results in a validation accuracy of 68%. The

basic mini-Xception model consists of a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a soft-max activation function to produce a prediction.

Our model is a slight variation of the above mentioned neural network. Instead of all ReLU activations, we alternate between using tanh and ReLU for every alternate layer. This was a combination that gave better results when tested for our dataset (i.e when both of these were trained on FER2013 and tested on our dataset, the results improved)

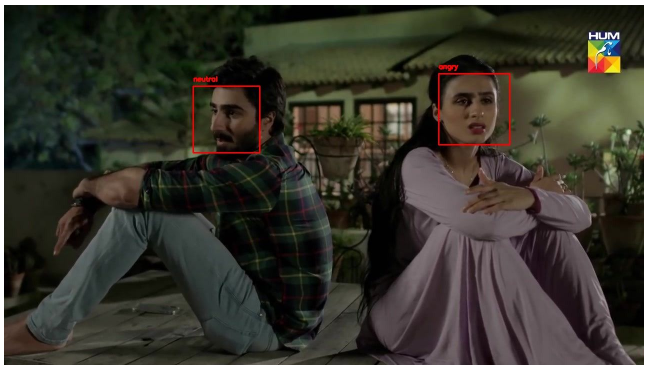
Our deployed model was based on "sequential-CNNs" i.e. no FC layers and depth-wise separable convolutions. Depth-wise separable convolutions are composed of two different layers: depth-wise convolutions and pointwise convolutions. The main purpose of these layers is to separate the spatial cross-correlations from the channel cross correlations [5]. This is inspired by the Xception architecture which is the "Extreme version of Inception" whereas Inception is another name for GoogleNet.

## 4. Results

Sample results of our model for emotion detection are shown in figure 4. The model detects the face in the image and then predict the emotion for that selected part of the input.


[Angry]


[All happy]




[Neutral, Angry]


[Scared]

[Happy]

Fig 5. Sample results of our model

Given below are the correlation matrices for the base model on FER2013, our model on FER2013 and our model when tested for samples from our own dataset.

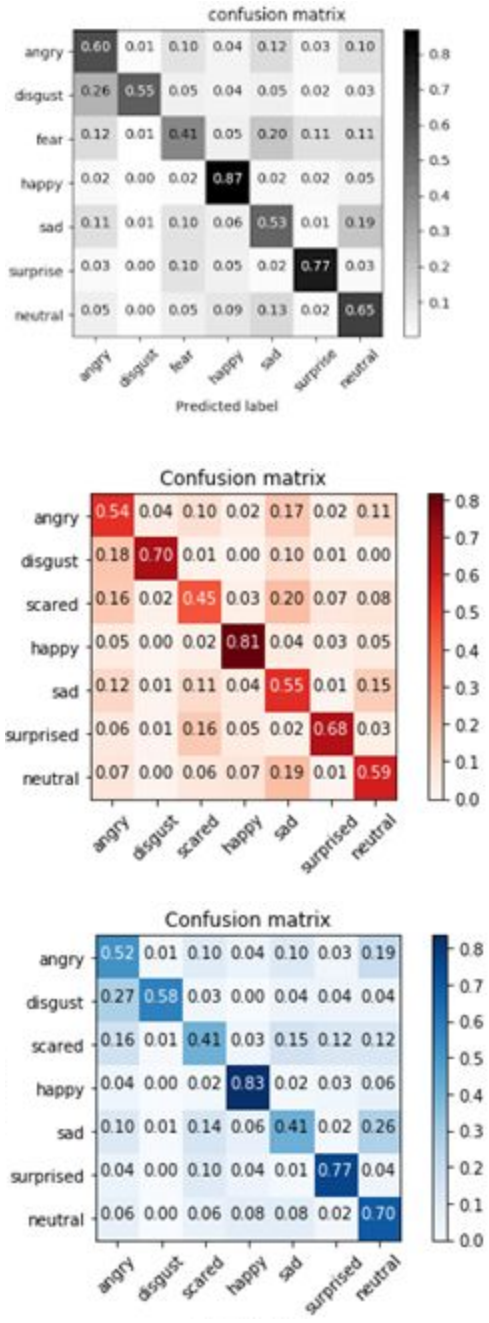



Figure 6. Confusion matrices for classification. The top image shows the matrix for base model (grey image), the middle image shows the matrix of our model tested on FER2013, and the bottom image shows results of our model tested on our dataset.

## 4.1. Model Comparison

The architecture proposed in mini-Xception model, trained on FER2013 was tested on our dataset. But that model failed to detect multiple emotions correctly in the same image. However, the architecture proposed by us performed well in that scenario. The figure below shows an example of this case.

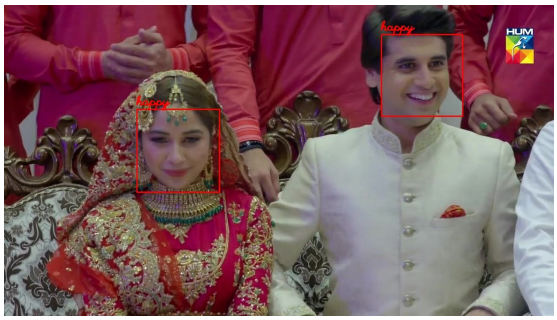



Figure 7. Results of our model (correctly classifies both images)

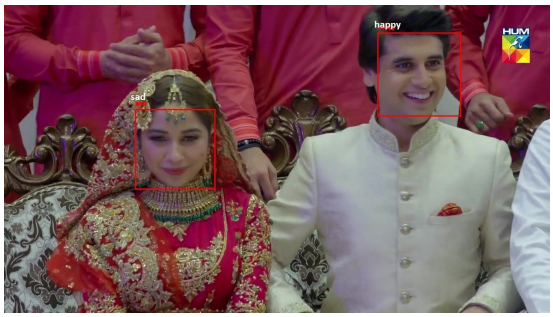



Figure 8. Results of the base mini-Xception model (misclassified the emotions of the female as sad)

## 5. Future Prospects

The proposed model uses Haar-Cascade for face detection in the input image and then apply mini-Xception model for emotion detection. In the future, we can work on emotion detection by applying attention to the significant parts of the input image. This will be done by incorporating an attention module prior to the mini-Xception module in the proposed architecture.

We have done emotion detection on single frames extracted from videos, but we can also make use of video data (i.e data from multiple sequential frames) to predict emotions. In this way we will be able to use contextual information of the most representative frame, whose emotion is to be predicted, by looking at the information in the previous and later frames.

This project can also be extended to do a comparative study of the facial expressions and the corresponding emotions between the people belonging to different geographical and cultural diversities. This will ultimately help to understand human behaviour, belonging to diverse range of ethnicities

in the field of human-computer interaction. Moreover, Social Reward systems are to be installed in the near future (2020 in China). These systems need to be unbiased in their detection. A comparative study of ethnicities can assist in that.

## 6. Conclusion

The proposed model has performed well on pakistani dataset with an accuracy of 64%. The accuracy can be increased further by increasing the size of the training data and training on a dataset collected from a specific ethnicity. However, the architecture can be further improved by incorporating contextual information using video data.

## References


[1] *Fernandez, F. Peña, T. Ren and A. Cunha, "FERAtt: Facial Expression Recognition with Attention Net", 2019.*

[2] *Ranjan, Rajeev, et al. "A fast and accurate system for face detection, identification, and verification." IEEE Transactions on Biometrics, Behavior, and Identity Science 1.2 (2019): 82-96.*

[3] *G. Levi and T. Hassner, "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns", in ACM, 2015.*

[4] *L. Wang, R. Li, K. Wang and J. Chen, "Feature Representation for Facial Expression Recognition Based on FACS and LBP", International Journal of Automation and Computing, vol. 11, no. 5, pp. 459-468, 2014. Available: 10.1007/s11633-014-0835-0 [Accessed 26 April 2019].*

[5] *F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", in CVPR, 2017.*

[6] *C. Bailer, T. Habtegebriel, K. Varanasi and D. Stricker, "Fast Feature Extraction with CNNs with Pooling Layers", in BMVC, 2017.*

[7] *S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey", 2018.*

[8] *X. Li, L. Yang, Q. Song and F. Zhou, "Detector-in-Detector: Multi-Level Analysis for Human-Parts", in ACCV, 2018.*