Facial Emotion Recognition

Faraaz Ahmed	**Manish Arigala**	**Tej Chaugule**
014351779	*014492712*	*013856089*

Abstract

Facial expression recognition is a topic of great interest in most fields from artificial intelligence and gaming to marketing and healthcare. The goal of this paper is to classify images of human faces into one of seven basic emotions. A number of different models were experimented with, including SVM and neural networks before arriving at a final Convolutional Neural Network (CNN) model. CNNs work better for image recognition tasks since they are able to capture spacial features of the inputs due to their large number of filters. The proposed model consists of six convolutional layers. Upon tuning of the various hyperparameters, this model achieved a final accuracy of 0.56.

TABLE OF CONTENTS

1. Introduction

Facial expressions are manifestations of nonverbal communication. Human beings communicate in the form of speech, gestures, and emotions. We see that the importance of emotion detection lies in various fields like business promotions and security. Business thrives on customer responses to all their products and offers. If an artificial intelligent system can capture and identify real time emotions based on user image or video, they can make decisions based on whether the customer liked or disliked the product or offer. People who use social media to share their experiences or express themselves, primarily use pictures and videos. It can also be seen as a next step of face detection where we may be required to set up a second layer of security, where with face, emotions are also detected. This is useful to verify that the authentic person standing in front of the camera is not just a 2-D representation.

Human emotions are classified as fear, disgust, contempt, surprise, anger, sad, happy, and neutral. These emotions are very elusive. Facial muscle contortions are very minimal to detect the differences and can be very challenging as even a small difference results in different expressions.

The dataset is the FER2013 dataset obtained from Kaggle. The dataset consists of 48x48 grayscale images of faces. All the faces are centrally aligned and occupy around the same amount of space. The objective of this project is to classify human faces into one of the seven emotions. With this in mind, a number of different models were experimented with to get comparable results of the highest accuracy that models can reach.

2. Background

2.1 Decision Tree (DT)

Decision trees are a supervised learning technique that predicts a value based on ”learning” rules based on a set of training data given a set of inputs. It is an enormous tree of if-then-else rules. The decision-making cycle starts at the tree's root and descends by answering to series of yes-no questions. It arrives at a single predicted label at the end of the if-then-else chain. This is the output of a decision tree.

2.2 Convolutional Neural Network (CNN)

A Convolutional neural network is a neural network composed of convolution layers which do the computation by performing convolution. Convolution is a mathematical operation on two functions resulting in a third function. The image is represented as pixel value numbers. The convolution operation takes place on these numbers. We make use of both fully-connected layers as well as convolutional layers. In a fully-connected layer, every node is connected to every other neuron in the network. They are the layers used in standard feed-forward neural networks. In convolutional layers, the connections are made across localized regions. Convolution is effective in classification and image recognition compared to a feed-forward neural network. This is because convolution allows us to take advantage of spatial locality by reducing the number of parameters in a network. Furthermore, convolutional neural networks introduce the concept of pooling that reduces the number of parameters by downsampling. Applications of Convolutional neural networks include robotics, self-driving cars and image recognition.

2.3 Support Vector Machines (SVM)

SVM is one of the most powerful classification algorithms. The idea is to divide the two classes accurately to find an optimal hyperplane. There is also a concept of margin, which is the maximum from both classes avoiding any overlapping between two the class. To achieve better classification results, data that is not linearly separable is mapped into a higher dimension. Kernel functions such as radial basis function (rbf) and polynomial are used for non-linear data. In the case of emotion detection, usually a multi-class SVM is used instead of a binary to detect the emotion expressions. K-fold cross-validation is used to compare different machine learning algorithms and to remove variances in the database. In k-fold cross-validation, the dataset is divided k times into k slices, and prediction results are averaged over all iterations. Principal component analysis (PCA) is used for feature set reduction and then the reduced feature set is fed to SVM. The image feature space is transformed to eigenspace using an eigen matrix in the PCA algorithm. Along with kernel specification, SVM has methods for tuning parameters like C and γ. Here, C is the penalty function for misclassification and gamma helps to optimize the decision boundary. Both these parameters affect the accuracy of the classifiers and can be tuned to get optimal results in both binary and multi-class classification.

2.4 Recurrent Neural Networks (RNN)

Recurrent Neural Network is a generalization of a feed-forward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input. Recurrent Neural Networks can also be divided into units called Long Short Term Memory (LTSM) if there are feedback loops present, or delays of time. The vanishing gradient problem of RNN is resolved here. It trains the model by using back-propagation. RNN can model a sequence of data so that each sample can be assumed to be dependent on previous ones. Recurrent neural networks are even used with convolutional layers to extend the effective pixel neighborhood.

2.5 K Nearest Neighbors (kNN)

Traditional kNN classification algorithms decide about the class of an image by searching for the k images of the training set most similar to the image to be classified, and by performing a class weighted frequency analysis. The k closest images are identified relying upon a similarity measure between images. The kNN based classification methods are modified by relying on images represented by means of local features generated over interest points, as for instance SIFT. With the use of local features and interest points, kNN classification algorithms consider the similarity between local features of the images in the training set rather than the similarity between images, opening up new opportunities to investigate more efficient and effective strategies.

3. Methodology

The task of facial emotion recognition can be broadly classified into two approaches. In the first approach features are extracted from the images. Now the images are represented using the features instead of pixels. After feature extraction, machine learning algorithms such as SVM, KNN, Decision tree are used to achieve the task of emotion recognition. This approach is followed by authors of [1]. The second approach would be to use neural networks to achieve the task of emotion recognition. During recent years, deep learning has made great strides in the field of computer vision. The authors of [2] have followed this approach. We will be following both the approaches. We will apply deep learning-based algorithms such as CNN (Convoluted Neural Networks) and RNN (Recurrent Neural Network) to find the solution to our problem along with the Machine Learning algorithms, SVM (Support Vector Machine), KNN (k Nearest Neighbors), and DT (Decision Tree). The following are the two approaches we took

3.1 Approach 1

- In this approach, we started off with the Scale-Invariant Feature Transform algorithm or simply SIFT. The SIFT algorithm is used to extract key points or features of images that are scale and orientation invariant. We extracted the SIFT features from all the images.
- In the next step, we used the bag of visual words methodology to generate feature vectors for each image in the dataset. The bag of visual words algorithm generates feature vectors based on key points and descriptors. All the feature vectors are clustered to form the visual words/codebook. Each image is represented as a histogram of features.
- We then trained and built multiple separate models using KNN, SVM, and Decision Trees.
- In the final step, we evaluated every model on the test data set.

3.2 Approach 2

- The images are normalised using min max normalization.
- A CNN/RNN model is built.
- The model is trained on the training dataset and validated on validation dataset.
- The model is evaluated using the test dataset.

4. Preprocessing

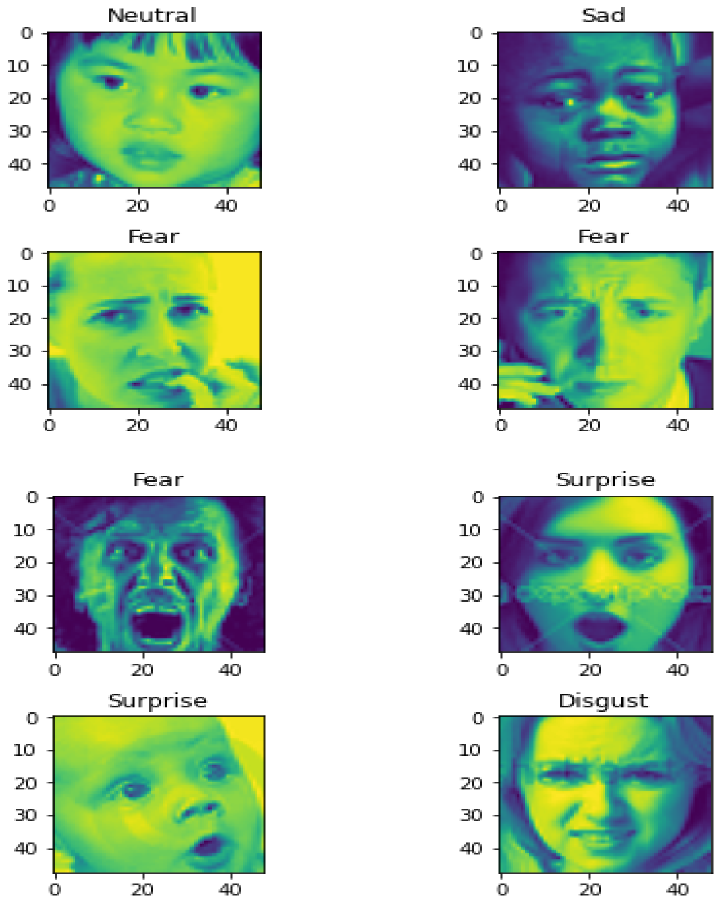
We have implemented two different preprocessing techniques.

1) Min-Max Normalization
2) Principal Component analysis

We have used min-max normalization with approach 2 and we have used PCA to reduce the number of dimensions of the feature vector with approach 1. Min-max normalization increased the accuracy of the model significantly whereas the PCA did not have any improvement. In min-max normalization we divide every pixel value by 255. In PCA we set the number of dimensions to 50.

5. Preliminary Results

Sample images from the dataset



The dataset is split into training, testing and validation sets. The number of images belonging to each class in these sets is shown in the figure below.

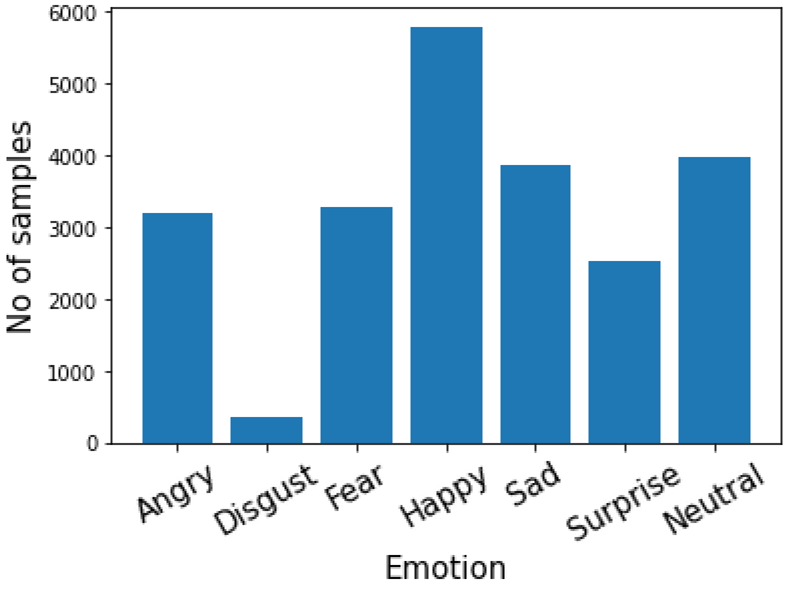


Fig:Training set

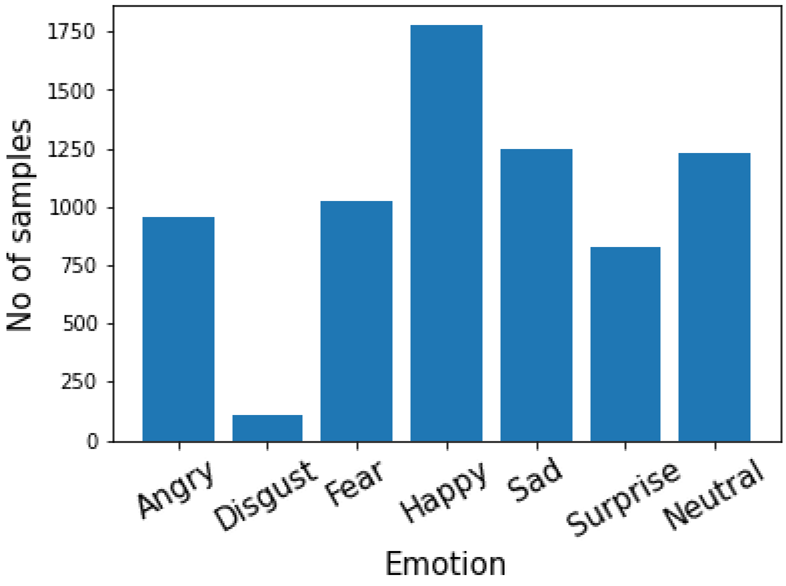


Fig:Testing set

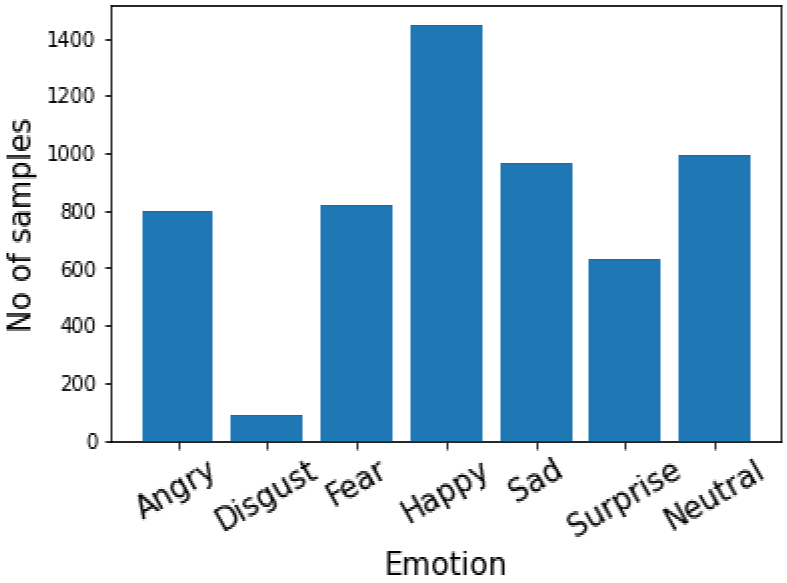
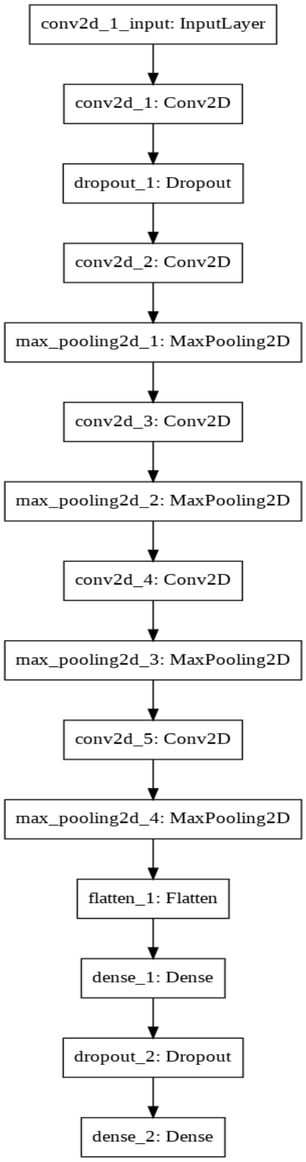


Fig: Validation set

Number	Algorithm	Accuracy
1	Convolution Neural Networks	55.84 %
2	Recurrent Neural Networks	41.84%
3	Support Vector Machine	33.18%

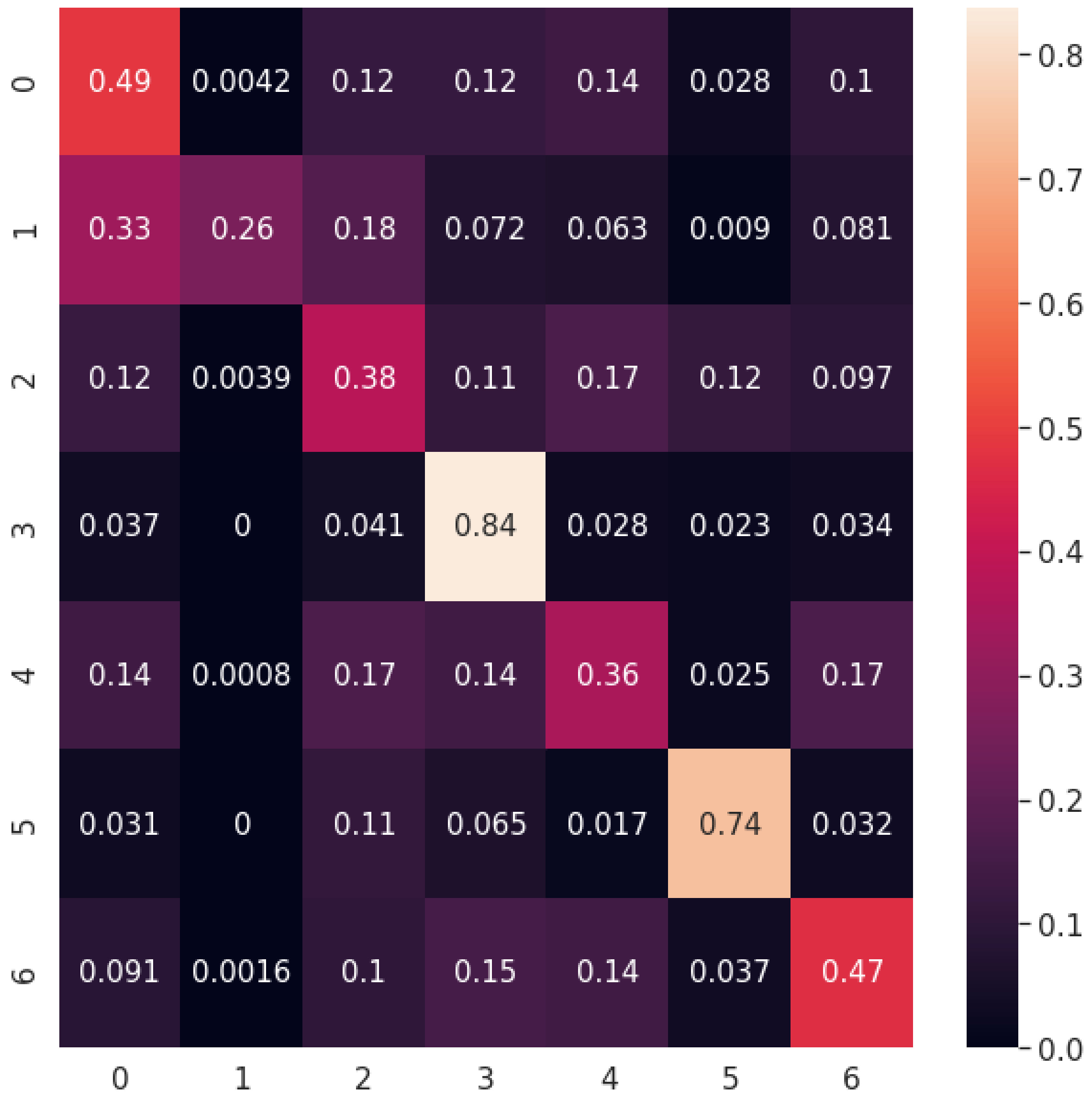
We have achieved the highest accuracy using CNN. The CNN we have built has 6 hidden layers.

The CNN architecture is given below.



The confusion matrix is plotted.

0-angry
1-disgust
2-fear
3-happy
4-sad
5-surprise
6-neutral



6. Comments

Since this is a dataset that a lot of people have worked on this dataset. Can you please answer the following:

1. Summarize what methods others have applied.

    We have listed the methods followed by others in the Methodology section.

2. How does your approach differ?

    Our approach would not be remarkably distinct from what others have followed. We plan to use image augmentation which is something most of the papers do not talk about.

3. Did you find anything new?

    Just the simple act of normalization has a huge impact on accuracy.

Try other methods as well such as logistic regression or using pre-processing the data via PCA (or other dimensionality reduction methods). How does this change the results?

We did train a model using SVM. We achieved average performance. We repeated the same thing by applying PCA to reduce the dimensions of feature vectors. But the performance did not improve. We are currently tuning the hyper parameters to see if we can achieve improvements in accuracy.

Would you be able to apply your method to 3D data? I have a set of 1000 3D scanned faces that have similar expressions. You could always just take a picture of the model, but I think it would be interesting to see a model applied to the actual 3D data.

After reading a few papers we feel that we can build a model that directly works with 3D data.

Curious, do you think that color images would not add any value to your model?

Most of the papers talk about converting the color images to grayscale for feature extraction. Hence we believe color would not significantly improve the model performance.

Finally, what will be the workload distribution among your teammates? Write down the tasks that each person will do.

We decided that every one of us would be building the complete emotion recognition pipeline. Hence we have divided the different algorithms that we would be using for this task between us.

Faraaz Ahmed - CNN
Manish Arigala - SVM
Tej Chaugule - RNN

This is what we have built so far. We still plan to make use of KNN and decision trees.
Furthermore, the report was done together by all three of us.

7. References

[1] R. T. Ionescu, M. Popescu, and C. Grozea. Local Learning to
Improve Bag of Visual Words Model for Facial Expression Recognition. In *Proceedings of ICML Workshop on Chal- lenges in Representation Learning*, 2013

[2] M. Georgescu, R. T. Ionescu and M. Popescu, "Local Learning With Deep and Handcrafted Features for Facial Expression Recognition," in IEEE Access, vol. 7, pp. 64827-64836, 2019.