2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-7281-5791-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECT49130.2021.9392584

Real Time Emotion Detection Using Deep Learning

Noel Jaymon, Sushma Nagdeote, Aayush Yadav and Ryan Rodrigues

Department of Electronics Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra

***Abstract*-Facial expressions depict emotions and produce information on the personalities and thoughts of people. The machine performs different tasks constantly in order to increase its use in public. Machines that are able to understand emotions can be used to execute a wide range of tasks. Machine perception demands machines to grasp information about their environment. The understanding of human facial expressions is a key component to understand emotions and also to find broad applications in the field of human-computer interaction. Tensorflow framework, Keras library and the Xception Architecture of CNN are used to train the model on the Fer2013 dataset.**

Index Terms—Facial expressions, Xception, model, InceptionV3, Convolutional Neural Network

I. Introduction

Emotions play a very important role in building relationships and having effective communication between people. A variety of emotions are experienced by people on daily basis. People do not realise or understand most of them because they cannot be noticed by the bare eye. Hence, with the help of technologies like machine learning and deep learning, we can detect and recognise the emotions that are generally difficult to be captured with just the naked eye. In the past few years there has been a rapid growth of technology in building smart solutions which has increased the need to detect a persons emotions. Some of the fields in which human emotion recognition can be leveraged are human-computer interface, animation, medicine and security. Convolutional neural networks can extract and learn important features from images which can be used to create a Real Time Facial Emotion Detection System. For facial expressions, most of the valuable information is obtained from the mouth, eyes, eyebrows etc. Some applications of automatic analysis of emotions from facial expressions can be seen in many fields, like smart teaching systems, emotionally sympathetic robots, driver fatigue monitoring, interactive gamming experience, and emotion based data retrieval, categorization and management. Anger, disgust, fear, happiness, sadness, neutral and surprise are the 7 universally accepted and recognized emotions. The three main stages in an automatic Facial Emotion Detection System are face detection, facial feature extraction and emotion recognition. Real Time face detection is done using Open CV and feature extraction and emotion recognition is performed using Deep learning (CNN).

II. Related Work

Life is made smarter and easier as robots are able to communicate with humans by detecting their emotion and responding or performing specific task as a result of that emotion. Most of the emotions by humans are expressed using facial expressions. In this, the authors have proposed Bayesian networks to perform emotion detection from facial features[1]. The studies have had the method to fill in the gaps of obstructed features after capturing facial features from each image. Bayesian network classifiers works from the reliabilities among the target attribute and obtain explanatory variables.

Vydana et al. [2] proposed Spectral features from speech segments, depending on the location in the utterance and used them to learn about the existence of emotion in speech. Speech data from IITKGP-SESC was adopted in the study which made use of Gaussian mixture modeling with a universal background model (GMMUBM) to boost the conventional GMM due to less number of samples in small speech segments.

Schuller et al. [3] proposed two methods and compared their findings. In the initial method a world wide statistics composition of a sample is categorized using Gaussian mixture model with the help of derived features of the original pitch and energy curve of the speech signal. The next method results in the increase of temporal complexity by implementing continuous hidden Markov models considering many states using low-level features from samples at that particular instant instead of world wide statistics. The paper depicts the structure of current recognition engines and results obtained from the indicated options.

Poria Soujanya et al.[4] in this paper, the authors mainly stressed on how we can utilize audio, visual and text information for multimodal affect analysis, since almost 90 percentage of the related literature seems to include these three modalities. If there is more than one channel or mode present like visual, audio, text, gestures etc, only then it can be considered for multimodal affect analysis.

Caridakis George et al. [5] in this paper, the authors presented a multimodal solution for recognising 8 emotions from the combination of facial expressions, movement of the body and speech using the Bayesian classifier for trainning and testing the model.Ten subjects and a corpus with 8 emotions was used to build the model. Initially the classifiers were trained separately for every modality. The data from the classifiers was fused

at the feature level and the recognition level. There was an increase in the recognition rates after fusing of the mutimodal data when compared to the results of the classifiers trained on every mode separately. The improvemnet after fusing the data was greater than ten percent in comparison to the best unimodal system.

Yang Y et al. [6] discussed about FER2013 challenge which was one of the public challenge that was organized to spread the news of advances made in the field of facial emotion recognition based on images. The winner of the FER2013 challenge had build one of the first CNN model to perform facial emotion recognition. The ensemble CNN model was trained to minimize the square hinge loss. The optimization of the model was done by performing data augmentation at the time of training as well as testing.

Balahur Alexandra et al. [7] built several models which were capable of recognizing 7 common emotions like Happiness, Sadness, Disgust, fear, neutral, angry, and surprise from facial expressions using the CK+ dataset. The model trained on images was capable to detect emotions in a real-time emotion recognition system which would take a video stream as input and continuously detect faces and classify the principal emotion of the faces detected in them. Though the real time system is able to distinguish between some of the seven emotions, more research and optimization is required to develop a robust real system that performs outside of laboratory conditions.

Medhat Walaa et al. [8] in this article, the authors compared and analysed the results of already known methods for emotion detection (supervised and lexical knowledge based) and a proposed method, based on common sense knowledge kept in the EmotiNet knowledge base. The present approaches are mainly focused on word-level study of texts and are mostly able to find only few expressions of sentiment.

Madhoushi Zohreh et al. [9] represents a complete, multilateral and structured review of opinion mining and analysis of emotions to classify currently present methods and compare their advantages and drawbacks, in order to understand the type of different challenges present and provide a solution to overcome these challenges and guide for the future direction.

Hemmatian Fatemeh et al. [10] provided a deep learning method focused on attentional convolutional network, which is capable to stress on important parts of the face, and obtaines better results over past models on many datasets, like FER-2013, CK+, FERG, and JAFFE. Most of the work perform considerably well on dataset of images captured in a stable condition, but does not provide good results when tested on complex datasets consisting of image variations and partial faces.

Sharef Nurfadhlina Mohd et al. [11] provides a complete feedback of the methods that have been put forward for emotion recognition in songs. A great effort has been made in the music information retrieval community to train a machine learning model to predictably classify the type of emotion just from the music signal.

Minaee Shervin et al. [12] proposed Automated facial expression analysis which is generally studied to look into the topics like behavior understanding and human-computer interface. The method they applied which is similar to the standard method which makes use of dynamic dense appearance descriptors and statistical machine learning techniques. The combined classification result of the emotion detection reached 70 percent which is noteworthier than the 56 percent accuracy obtained by the standard method demonstrated by the challenge organizers.

Ajay B. S et al. [13] in his article, several classification methods are analyzed to recognize applicable emotional feelings from prosodic, disfluency and lexical cues taken from the real time conversations between humans. The studies of real time spoken dialogs from two call center services show the existance of many combined emotions, based on the dialog reference.

Francois Chollet et al. [14] showed that the xception architecture performs better in comaprison with the Inception V3 model on the ImageNet dataset and performs remarkably better than Inception V3 model on a greater image classification dataset consisting of 350 million images and 17,000 classes.

III. PROPOSEDMETHODOLOGY

Proposed methodology can be split into two stages i.e Model Training and Real Time detection. Below Figure.1 displays the generic methodology in the form of a flowchart.

A. ModelTraining

The dataset is imported into a pandas dataframe using the pandas library on which all pre-processing operation was conducted. The pixel values represent the image. The pixel values and emotion values are stored in a separate list. Since keras does all computation only on numpy arrays. The list of pixel and emotion are converted into numpy arrays. Feature engineering techniques like normalization of the pixel values and one-hot coding is done to achieve faster computation. Then the numpy arrays are reshaped into the desired shape as input to the keras API. For the current version the format is channels last.The keras documentation can be referred for such information. The input is given to the convolution network and the model is trained to classify the emotions.

B. Real Time Detection

The frame from the live video input is saved and converted into grayscale. The haar-cascade classifier by

Open CV helps in detecting the face present in thegrayscale image. After face detection, the ROI is selected using the pixel locations of the detected face received from the detectMultiScale() method. Then the ROI is reshaped to the same size of the input image which was used during the training of the model (i.e 48x48) and then the trained model will make a prediction on the image. The prediction will then be displayed on the screen.

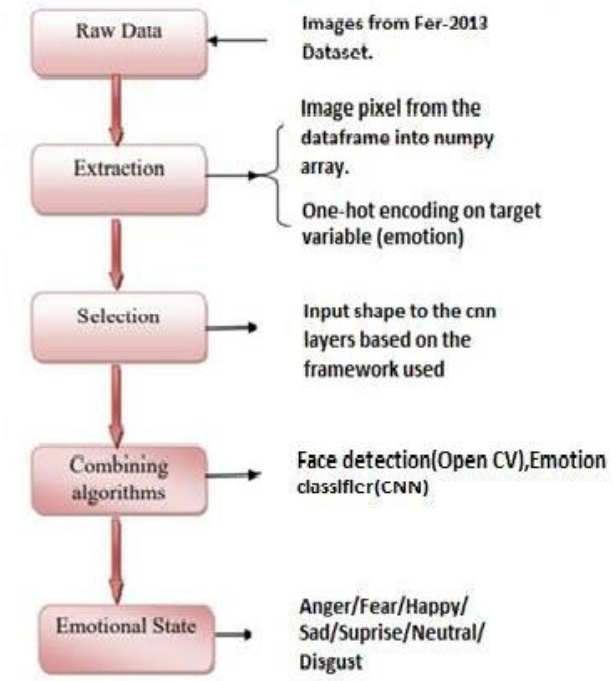

Fig. 1. Generic Process of Emotion Detection

IV. Dataset Description

The Fer-2013 dataset from kaggle is used in this paper. The dataset was created by Pierre-Luc Carrier and Aaron Courville, as part of their current study. The dataset consist of 35,887 images which are 48x48 pixel grayscale images of faces with different emotions. This dataset has already been pre-processed and hence all the faces are mostly in the centre and have the same dimensions.The images consist of emotion from these seven categories such that 0,1,2,3,4,5,6 maps to Angry, Disgust, Fear, Happy, Sad, Surprise and neutral respectively. The number of images for every emotion is shown in Fig.2.

The dataset consist of 3 columns and 35,887 rows. Every row represents an image and contains the features of the respective image. The columns are named as emotion, pixels and Usage. The emotion column contains the label (0-6) of the particular image. The pixels column consists of the pixel values of every image. The Usage column is used to denote if the particular image is used for training or testing. The images are divided into three sets. Table.1 displays that the Training set consist of 28,709 images and the Public and Private Test sets consist of 3,589 images each respectively. The Public and Private sets can be used for validation and testing.

V. Convolution Model Architectures

A. Simple Convolution Model

The Convolutional Neural network is widely used on images. Images with highly complex features can be classified with the help of these networks. The network consist of different layers as shown in Fig .3. It reads the image pixels in the form of numpy array as a input. It consist of a Convolutional layer, pooling layer, fully connected layer, and activation layer.

The convolutional layer is used to perform feature extraction. The deeper convolutional networks tend to learn more complex features. For example, the convolutional layer close to the input learns features like vertical and horizontal edges, curves, simple colors. The activation layer is simply the output of a given function. Some of the activation layer functions which are used in CNN are identity function, tanh function and rectified linear unit(relu). The relu activation layer is used to introduce non linearity in the output. The network tends to learn faster and more accurately with the introduction of non-linearity. The function of the Pooling layer is to decrease the dimensions of the image as it passes through the different layers which decreases the number of parameters and thus the calculations required by the network in a nonconventional manner. Only the important features are extracted. The usually used pooling techniques are Average pooling and Max pooling. Generally the input to the fully connected layer is an input volume which can be the output of the convolution or activation or pooling layer pesent before it. and gives a N dimensional vector, where N is the number of classes that the model has to choose from. In this case N=7 as there are 7 type of emotions in the dataset.

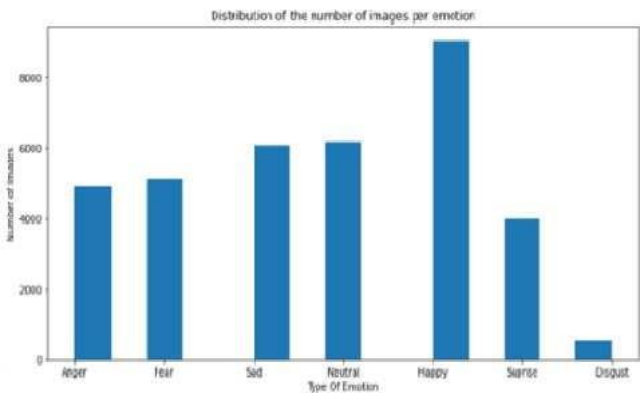

Fig. 2. Bar graph of no of imagesper emotion

Table 1. Numerical value of no of images in emotions and Usage

Emotions	
Happy	8989
Neutral	6198
Sad	8077
Fear	5121
Anger	4953
Surprise	4002
Disgust	547

Usage	
Training	28709
PrivateTest	3589
PublicTest	3589

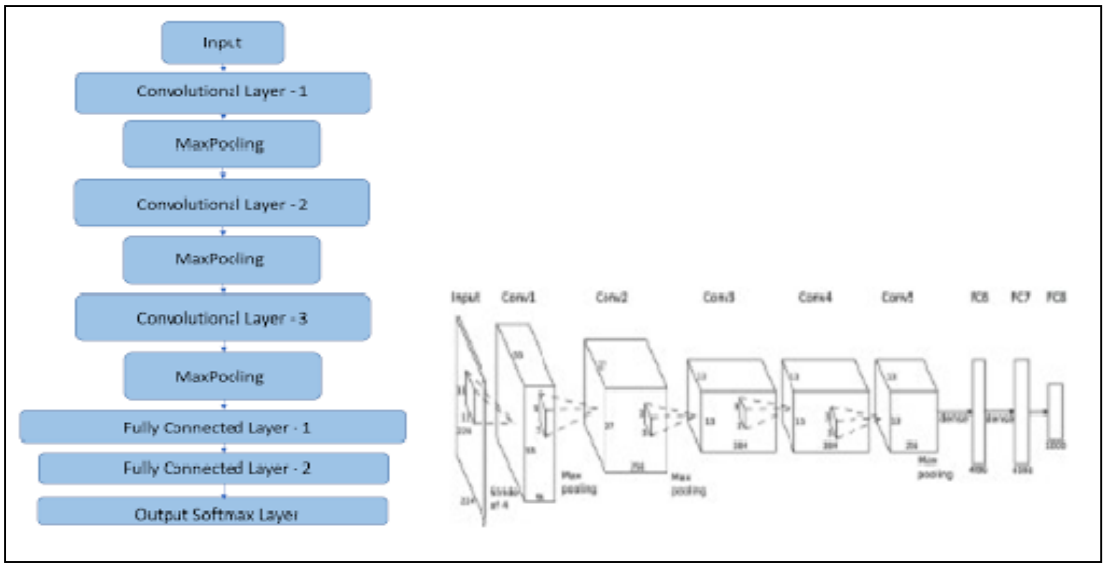

Fig. 3. Simple convolutional network architecture

B. Inception Model

The Simple model of CNN as shown in Fig.3 would have to stack up convolution layers to learn complex features and produce accurate results. Naively stacking large convolution operations is computationally expensive and will take more time to train. Such networks are prone to over-fitting. Thus the inception model was introduced to solve this problem. As every kernel extracts different information, an Inception module executes various different computations over the same input map by various kernels at the same time and concatenates their results into a single output. In other words, for each layer, Inception does a 5x5 and 3x3 convolutional transformation in a single layer and a 3x3 maxpool layer. Thus this model is able to take the advantage of Multi-level feature extraction. Thus instead of stacking up convolutional layers and making this network deeper the inception module makes the network wider. The computational cost and the number of feature maps produced by every layer will increase drastically if dimensions are not reduced at the output of every layer. Hence in order to reduce the dimensions 1x1 Convolutional filters are used while concatenation. A convolution layer with a filter size of 1x1 looks at only one value at a time, but across all channels, it can extract spatial information and reduce it down to a lower dimension. For example, using 10,1x1 filters, an input of size 48x48x100 with 100 feature maps can be lowered to 48x48x10. Fig.4 which displays the Inception module with dimension reductions. The inception v3 model is also used. In this model two 3x3 convolutional layers in place of one 5x5 convolutional layer is used to get better computational advantage. 21 layers of convolution and the architecture includes three centers is applied for concatenation. The axis must be set as axis=1 during concatenation along with this batch normalization and max pooling before beginning and is utilized to widen the network to speed up the training. The model could be visualized with the help of plot model function from keras.utils.vis_utils. The model architecture between two concatenation blocks as shown in Fig.5 helps in

understanding how the inception v3 helps to widen the network and reducing dimensions becomes easy. Increasing the depth of the model did not provide any improvement in the results. The use of plot_model() and model.summary() helps you visualize the network and get the shape of input and output of every layer. The functions are useful to analyse the construction of the network and is useful in preventing overfitting.

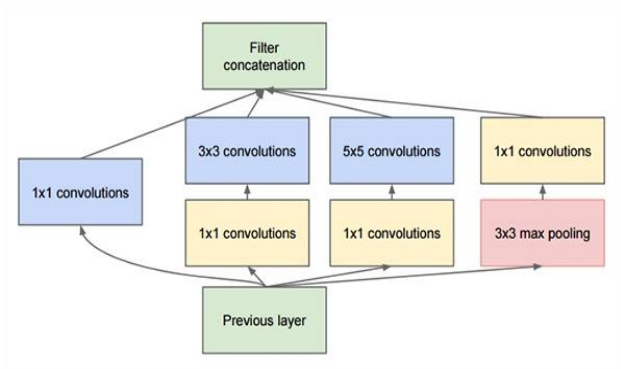

Fig. 4. Inception network architecture

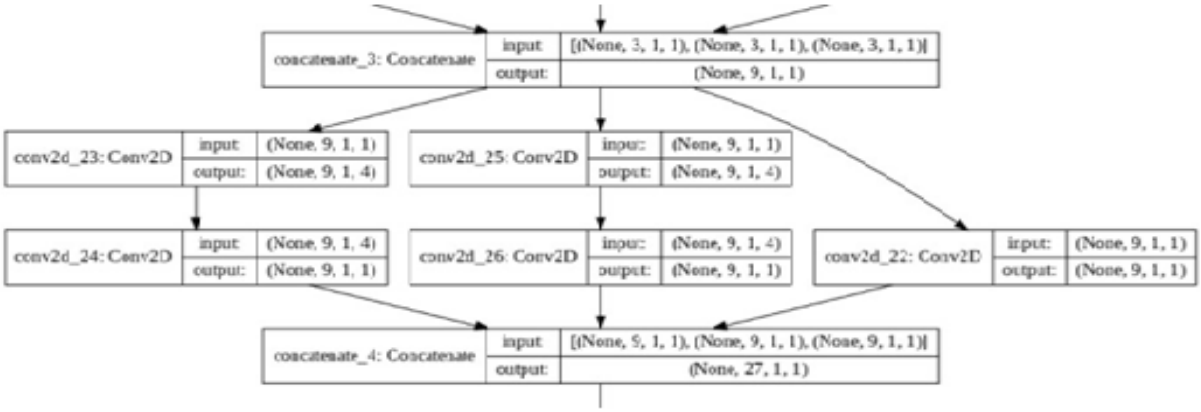

Fig. 5. Model plot between two concatenation blocks

C. Xception Model

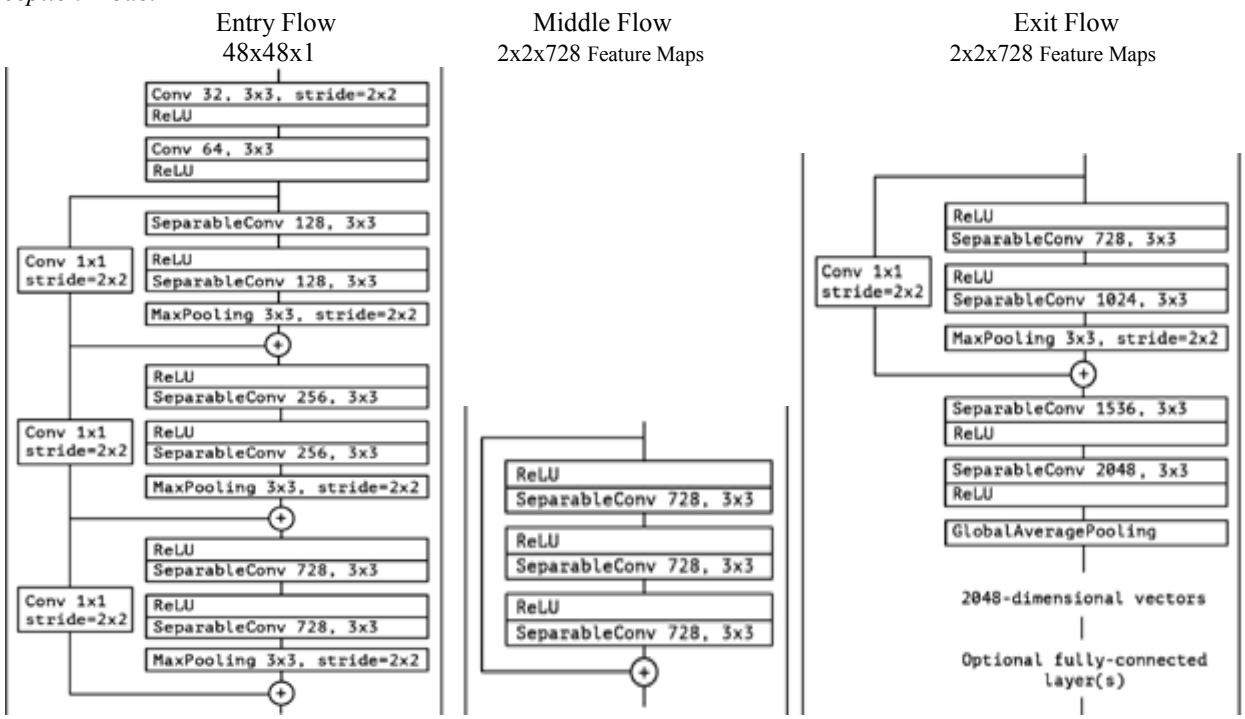

Fig. 6. The Xception Architecture

The Xception architecture is based entirely on the concept of depthwise seperable convolutional layers. Thus this model does not follow the idea of performing cross channels correlations and spatial correlations in the feature maps of convolutional neural network like the inception model. Thus this model is named Xception which means Extreme inception. The use of depth- wise convolution gives higher computational advantage to the model. Point wise convolution and depthwise convolution are the two types of convolution that take place in the Xception model. The complete description of the specification of the model is depicted in Fig.6.

The data enters via the entry flow as input , then travels towards the middle flow which is executed eight times, and finally reaches the exit flow. Note that all Convolution and Separable Convolution layers are followed by batch normalization (not included in the diagram). All Separable Convolution layers use a depth multiplier of 1 (no depth expansion). Feature extractionis performed using 36 convolutional layers in the Xception architecture. The ideological concept that every layer receives the information produced by its preceding layers was implemented by this model architecture. This would help the layer learn better and produce more accurate results as they would be exposed to more information. The task was implemented using a residual layer.The 36 convolutional layers are organised as 14 modules which contain a linear residual connection except for the first and last modules.The use of plot_model() and model.summary() helps you visualize the network and get the shape of input and output of every layer. The functions are useful to analyse the construction of the network and is useful in preventing overfitting.

VI. Experimental Evaluation Of The Models

Table 2. ExperimentalResults

	Simple Model	Inception Model	Xception Model
Training Accuracy	99.6%	65.4%	93.2%
Validation Accuracy	53.58%	60.52%	64.4%
Test Accuracy	54%	61.42%	65.2%
Number Of Epochs	30	150	150

A. Training Infrastructure

All the models were trained, validated and tested on google colaboratory (for its GPU assistance) using Tensorflow and KerasAPI. The Xception model outperformed the inception v3 and the simple model with its results. Thus we choose this model for the task of real time emotion detection. The plot of the graph between the training accuracy and validation accuracy and training loss and validation loss is given below in Fig. 7 and Fig.8 respectively. The curves are useful in providing better intuition on the model performance which also makes it easier to decide the different optimization techniques to apply which would improve the performance of the model

B. Data Description

The training was done using 28,709 distinct images.The private and public test set consist of 3,589 distinct images which waas used for validation and testing respectively Each image is a 48x48 pixel grayscale image of different faces.

C. Data Augmentation

The models were trained using Data augmentation technique to reduce overfitting. In this technique the parameters like zoom range, rotation range, horizontal flip, vertical flip and height and width shift range can be adjusted to create new images from the original image.

D. Experimental Results

The Table 2. displays the accuracy of all the three models that were trained, validated and tested using the FER2013 Dataset.

VII. Best Model Plot

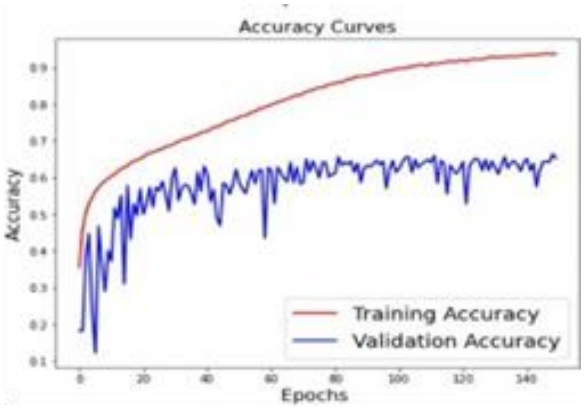

Fig. 7. Accuracy Curves

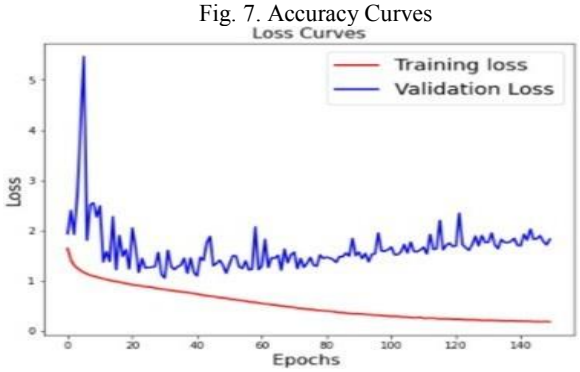

Fig. 8. Loss Curves

VIII. Result Of Real Time Implementation

The results are obtained using the internal webcam of the

laptop. The model detects multiple faces in a frame and successfully detects their emotions. The Figures 9(a), 9(b), 10(a), 10(b), 11(a) and 11(b) are the results of the Happy, Sad, Surprise, Neutral, Fear and Angry emotions respectively.

Fig. 9. (a) Happy Emotion (b) Sad Emotion

Fig. 10. (a) Surprise Emotion (b) Neutral Emotion

Fig. 11. (a) Fear Emotion (b) Anger Emotion

IX. Conclusion And Future Scope

The model performed extremely well in detecting all 7 emotions on an image provided by the user. During the Real Time Detection of emotion the model lacked robustness. It successfully detected 6 out of 7 emotions. The lack of robustness was found out to be because of the size of the images used in training data. Considering the fact that it is extremely hard for a human to detect the emotion from a 48x48 image and the winner of the Kaggle challenge had an accuracy of 34 percent, this model has given satisfactory results.

The model can be used in security systems to detect potential threat. Gaming experience can be increased extensively with the help of this model. The model can help develop interactive methods to teach small children about emotions. Analytical results on the emotions of the candidate in video interview can be produced. Fun activities such as clicking selfie can be upgraded with the help of this model.

References

[1] Y. Miyakoshi and S. Kato, "Facial emotion detection considering partial occlusion of face using Bayesian network," 2011 IEEE Symposium on Computers & Informatics, Kuala Lumpur, Malaysia, 2011, pp. 96-101, doi: 10.1109/ISCI.2011.5958891

[2] Hari Krishna Vydana, P. Phani Kumar, K. Sri Rama Krishna and Anil Kumar Vuppala, "Improved emotion recognition using GMM- UBMs", 2015,International Conference on Signal Processing and Communication Engineering Systems, pp. 53-57.

[3] B. Schuller, G. Rigoll M. Lang,"Hidden Markov model-based speech emotionrecognition", 2003, Proceedings. International Conference on Multimedia and Expo, volume 2, pp. 401-404

[4] Poria S, Cambria E, Bajpai R & Hussain A (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, pp. 98-125.

[5] Caridakis George, Castellano Ginevra, Kessous Loic, Raouzaiou Amaryllis, Malatesta Lori, Asteriadis Stelios, Karpouzis Kostas "Multimodal emotion recognition from expressive faces, body gestures and speech", 19 September 2007, IFIP Advances in Information and Communication Technology (AICT) , Volume 247, Issue 247, pages 375-388

[6] Yang Y.-H. and Chen H. H. "Machine recognition of music emotion: ACM transactions on Intelligent Systems and Technology, Volume 3, Issue 3, pp. 1-22, Article No.: 40

[7] Balahur Alexandra, Hermida Jesus M, Montoyo A. "Detecting implicit expressions of emotion in text: A comparative analysis", November 2012 Decision Support Systems, 53 (4), pp. 742-753

[8] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal,Volume 5, Issue 4,2014, Pages 1093-1113.

[9] Madhoushi Zohreh, Hamdan Abdul Razak, Zainudin Suhaila. "Sentiment analysis techniques in recent works". 2015 Science and Information Conference (SAI), pp. 288-291.

[10] Hemmatian Fatemeh, Sohrabi Mohammad Karim, "A survey on classification techniques for opinion mining and sentiment analysis" Artificial Intelligence Rev (2019), 52: pp. 1495–1545

[11] Sharef Nurfadhlina Mohd, Zin Harnani Mat, Nadali Samaneh "Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data". Journal of Computer Science, Volume 12 No. 3, 2016, pp. 153-168

[12] Minaee Shervin, Abdolrashidi Amirali, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network", 2019, Computer Vision and Pattern Recognition

[13] Ajay.B.S, Anirudh.C.R, Karthik Joshi.S, Keshava B.N,Mrs. Asha.N, "Emotion Detection Using Machine Learning", International Journal of Recent Trends in Engineering and Research(IJRTER), Volume 3, Issue 6, 2017, pp. 28-32

[14] Francois Chollet ,Google Inc. "Xception: Deep Learning with Depthwise Separable Convolution",2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),