IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FACIAL EXPRESSION RECOGNITION USING DEEP LEARNING

[1]Yogesh S. Ambhore, [2]Prakash J. Kulkarni

[1]M.Tech Student, [2]Professor

[1]Department of Computer Science and Engineering,

[1]Walchand College of Engineering (An Autonomous Institute), Sangli, India

Abstract: Facial expressions are considered to be the fastest way of communication while delivering any type of information. Facial expressions play a key role in getting an understanding of the mental views of any person. With changing time, enhancement in technology is significantly increasing day by day. In neural networks, CNN is well known for carrying out image recognition and image classification tasks. To improve human-computer interaction, it is needed to build an efficient scheme for human expression recognition. In facial expression recognition system, feature extraction is the first step. This is followed by the classifier which is used to classify facial expressions. This paper aims to provide an automated facial expression recognition system which can efficiently classify and recognize the facial expressions. This proposed work runs on top of a convolutional neural network, image data of human facial expressions is fed to the neural network build with the help of Keras. The proposed methodology aims to build a neural network to recognize facial expressions by using the architectural properties of pre-trained CNN architectures such as VggNet and Alexnet**.**

Index Terms **– Convolutional Neural Network(CNN), Facial Expression Recognition(FER), Alexnet, VGGnet.**

I. INTRODUCTION

Arrangement of facial muscles is considered as facial expressions and each arrangement of facial muscle conveys a certain emotion to the observer[10]. Facial expression is one of the finest ways to express our emotions inefficient way [6]. The ability to detect and classify an emotion correctly opens many doors to Advanced Human-Computer Interaction[1]. Facial expressions can be recognized by one of three methods i.e geometric feature extraction based, appearance-based method and the third one is the neural network-based method[8]. In the geometrical feature-based approach, the face is represented by a set of landmarks which used to denote the facial angle and distance between the facial points[8]. The main difficulty in geometrical based FER system is to find out the facial points and needs to be manually labeled with landmarks[8]. In an appearance-based approach, features are extracted from the pixel intensity values in the facial image.

The neural network-based recognition strategies are seen as especially encouraging since the neural network can without much of a stretch execute the mapping from the feature space of face images to the facial expression space. In a neural network, a convolutional neural network which is a deep learning method is one of the main categories to do image classification and recognition. CNN has the ability of automatic feature extraction, each layer of the CNN extracts a unique feature of the given input image.

Pre-trained CNN architectures like VggNet, AlexNet, ResNet, etc. are trained with the image-net dataset of over 15 million high resolution labeled images and achieved the highest accuracy and performed well with classification and recognition tasks. This study aims to provide an automated facial expression recognition system which can efficiently classify and recognize the six basic emotions of a human being.

Facial expressions can be efficiently recognized by using a convolutional neural network, as each layer of CNN extracts unique features of the input image which makes it more accurate for prediction of output. The objective of CNN is to convert the set of the input image into meaningful output. The FER system follows two steps to detect facial expressions efficiently firstly it detects face from the input image then emotion detection on the detected face[11]. In the proposed methodology, OpenCV is used for image processing tasks where Haar feature-based cascade classifiers are used to detect face from a given image which is then processed and fed into a trained neural network for emotion detection. The facial emotions that can be detected and classified by this system are Happy, Disgust, Sad, Anger, Surprise, and Neutral.

II. LITERATURE SURVEY

Facial expression recognition can be carried out with the help of one of three methods such as the geometrical feature extraction method, appearance-based feature extraction method, and neural network-based feature extraction method. Anima Majumder et al. proposed a facial expression recognition method using a deep network framework consisting of autoencoders. Geometric and local binary pattern features are extracted and fused using autoencoders. Kohonen-self organizing map based classifier is used for facial expression classification[2].

The face can be recognized by using one of three methods, geometric feature-based method, appearance feature-based method and neural network-based methods[8]. L. Ma and K. Khorasani Introduced a facial expression recognition system that uses two-dimensional discrete cosine transform as a feature detector and for classifying the facial expression one hidden layer feed-forward neural network is used[1]. A facial expression recognition scheme based on hierarchical deep learning is proposed by JI-hae Kim et al. which firstly extracts the facial features on the basis of appearance and then this feature network is fused with geometric facial features in a hierarchical structure[3]. The softmax function is used in the prediction of emotions and a ten-fold cross-validation technique is used to validate the accuracy[3]. Method for facial expression recognition on the basis of depth volume and neural network is proposed by Minjun Wang et al., the proposed method takes an image of expression as input to the CNN i.e. deep convolutional network is used to extract the facial features and trained CNN is then used to recognize the facial expressions, the softmax classifier is used to classify the facial expressions[6]. Features related to facial expressions are extracted and fine-tuned with partial VGG16, softmax function is used to classify the emotions[9]. In this paper, Guan Wang and Jun Gong propose a method based on the cross-connected multilayer LeNet-5 CNN model for facial expression recognition. Network structure extracts low-level features which are then combined with high-level features to efficiently build the classifier[7]. The goal of the method proposed in the paper is to use CNN to analyze the facial emotions in real-time. This is a special kind of deep learning technique that gives the solution to many facial emotion recognition problems after a significant amount of training[12]. The key benefit of CNN is the total elimination or substantial reduction of reliance on physics-based models and/or other preprocessing techniques[13].

From the literature review, it is found that traditional approaches such as geometric feature-based, appearance-based approach for facial expression recognition are not robust due to lack of image information, The neural network-based approach works fine with the classification and recognition of facial expressions.

III. METHODOLOGY

The diagrammatic representation of the proposed methodology for FER is given in figure 1. The methodology proposes the implementation of a neural network which will acquire the architectural properties of pre-trained CNN models. The proposed methodology uses the architectural properties of pre-trained CNN architectures such as VGGNet, AlexNet, etc., architectural properties of both pre-trained CNN models are going to used to build a neural network so more accuracy could be achieved.

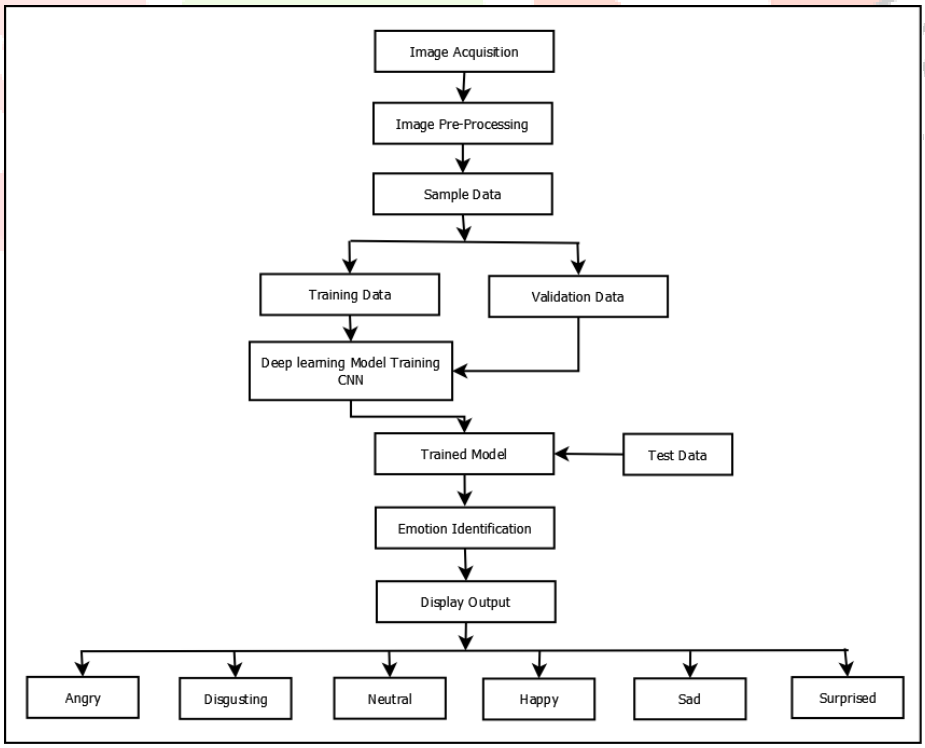
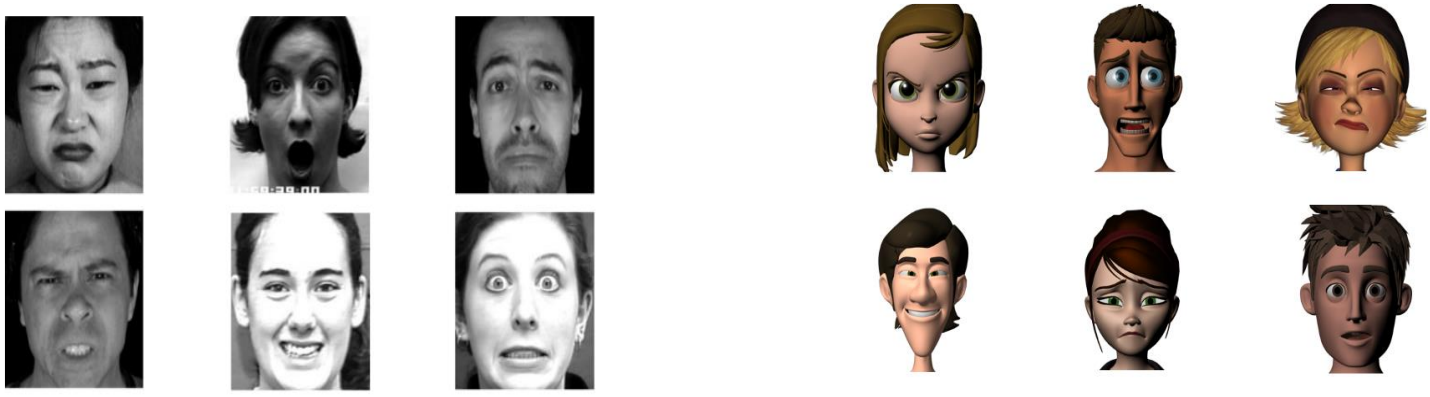


Figure 1 Block diagram of proposed system

Image Acquisition: Image acquisition is defined as the action of retrieving images from various sources[15]. For the facial expression recognition system using deep learning techniques, the performance is tested and validated on the merged dataset of Ck+ and FERG-DB which are known to be widely used datasets for FER.

Ck+ and FERG-DB: To build a neural network that acquires the properties of pre-trained CNN architectures CK+ and FERG-BD datasets are used, both datasets are downloaded from their respective sources. Extended Cohn-Kanade Dataset (CK+) includes human facial images whereas FERG-DB includes facial images of six animated characters. random 10,548 images from both datasets are selected and segregated under six labels of human expressions. Fig 2(A) and 2(B) show data samples of Ck+ and FERG-DB respectively. The performance of the CNN architecture is trained and tested on CK+ and FERG-DB datasets.



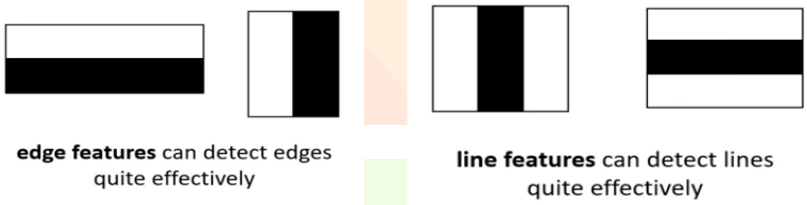
(A) (B)

Figure 2 (A)CK+ data samples (B) FERG-DB data samples

Image Preprocessing: In order to improve image quality, image preprocessing is done by ignoring unwanted data and enhancing important features of the image. It allows a wider range of algorithms to applied over the image. In order to recognize and classify facial expressions from image data image prepossessing needs to be done to get efficient results. For facial expression recognition following image preprocessing, steps need to be followed.

Face Detection Using Haar Cascades Classifier: Haar cascades classifier is an effective technique for object detection. To detect an object from the input image, line features, and edge features shown in fig 3 are used. Haar classifier is a machine learning-based approach and is trained with a large number of positive as well as negative images.

– **Positive images** – Positive images are images with faces.
– **Negative Images** – Negative images are images without faces.



(A) (B)

Figure 3 (A)Haar edge features (B) Haar line features

Haar features are a sequence of rescaled square-shaped features, these features are very similar to convolutional kernels. The Haar feature helps to detect the most relevant features such as eyes, nose, lips, forehead, eyebrows, etc from the input image. So these are the most relevant features as far as the human faces are concerned. Ideal pixel intensity values for the black and white regions of haar features are 1 and 0 respectively. Each time when the input image is given, from input image their real pixel intensities are obtained. To detect features from the input image, each time the calculated average of white pixel intensities is subtracted from the calculated average of black pixel intensities. if the obtained value is the closer the value to 1, the more likely the haar feature has been found. To calculate the delta value for the real image the formula is as follows where the delta value for ideal haar feature is 1.

$$\triangle = Dark - White = \frac{1}{n}\sum_{dark}^{\infty} I(X) - \frac{1}{n}\sum_{white}^{\infty} I(X)$$

In the proposed methodology ”haar cascade frontal face default” is used to detect frontal face from the input image and crop it. Faces are cropped from image data in order to avoid irrelevant feature extraction from the non-facial part of an image. A non-facial part from the input image is removed and only cropped facial images are stored. All detected faces are stored under their respective emotion labels. This detected facial image data is fed to the neural network for classification.

Image Grayscaling :
The image grayscaling process is defined as it is a process of converting an image from other color spaces to shades of gray. FERG-DB dataset consists of colored images of six animated characters. As it contains colored images it will be difficult for the neural network model to extract important features from the input image. To avoid misclassification and wrong prediction of emotions, images from FERG DB are grayscaled.

Image Resizing :
CK+ and FERG-DB are two different datasets with varying image size, in order to build a neural network which acquires properties of pre-trained CNN architectures images are resized to 100 X 100. resized images are then fed to the neural network.

Data Augmentation :

Image data augmentation is a technique that can be used to expand the size of a training data artificially by creating new data i.e modified version of images with different orientations. Data augmentation is performed to prevent overfitting and to generate more training data from the limited data. Data augmentation is carried out by the following operations.

– Image Shift: Image shift is used to shift images either by width range or by height range.
– Image Flip: Image flip allows us to make a change in image orientations by flipping the image horizontally or vertically.
– Image Rotation: Image rotation allows us to make a change in image orientations by rotating images.
– Shearing: Shearing is also used to transform the orientation of the image.
– Zooming: Zooming allows us to make a change in image orientation by zoom in or by zoom out.

Split Image Data: CK+ and FERG-DB combined image dataset include a total of 10,548 grayscaled images of size 100X100 which are then split into training and validation datasets. The ratio of dataset split is 80:20 where 80% i.e. 8433 image data is fed to a Convolutional neural network for training purposes. The training phase of the neural network extracts important features from the grayscaled images. 20% i.e. 2115 image data is used to validation purpose of the neural network

CNN Architectures:

a. CNN-VGGNets:

VGGNet Come up with two versions VGG16 and VGG19 which consist of 16 and 19 convolution layers respectively and are very appealing because of its uniform architecture. The input image is passed through the stack of layers where each layer consists of a receptive field of size 3x3. RGB Image of size 48x48 is given as input to VGG, image is passed through the stack of layers. Preprocessing of the image is done by subtracting RGB mean from each pixel. Dimensionality reduction is handled by max-pooling. The following diagram shows the VGG network[4].

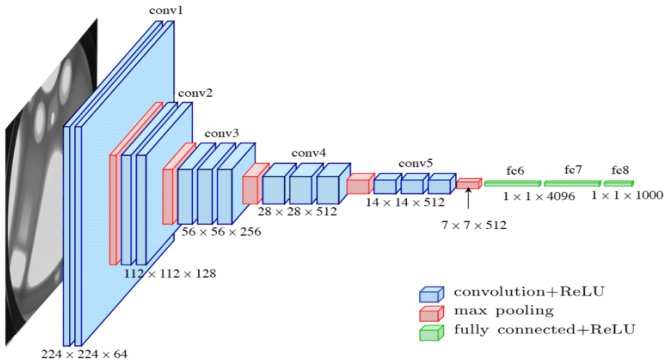


Figure 4 VGGnet architecture

b. Alexnet:

Alexnet is eight layered architecture out of which the first 5 layers are convolutional layers followed by 3 fully connected layers. Alexnet uses varying kernel sizes such as 11*11, 5*5, 3*3, 3*3, and 3*3 for its five convolutional layers respectively. Overlapping the max-pooling layer connected to the first two convolutional layers is used to extract the maximum number of features. The remaining convolutional layers such as C3, C4, C5 are directly connected to the fully connected layer. The rectified linear unit (ReLu) is used as an activation function, all outputs of convolutional layers and fully connected layers are connected to the ReLu activation function. fully-connected layer collects all inputs from convolutional pooling operation and gives as input to the softmax classifier, and finally, facial expression classification results are obtained.

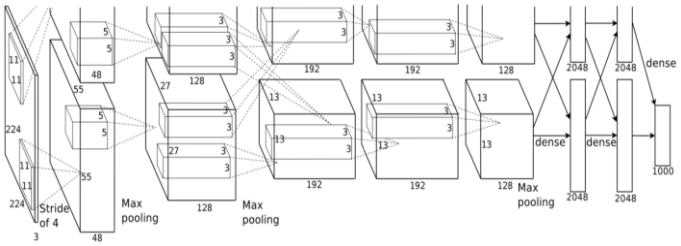


Figure 5 Alexnet architecture

Convolutional Neural network:

After the advancement in the neural network, the convolutional neural network is one of the main categories to do the image recognition and classification[14]. From previous studies, CNN has been found to be more effective and quicker for computer vision-related problems than a regular deep neural network[5].

This methodology proposes the building of a neural network that will acquire the architectural properties of pre-trained CNN models such as VGG and Alexnet. The formation of the neural network is a collection of three parts: The first part is the convolutional layer then the second layer is the pooled layer and finally fully connected layer. The structure of a neural network is derived from Alexnet neural network, it includes a total of 8 layers. The proposed neural network is comprised of five pooled layers along with convolutional layers and out of three fully connected layers, one is the output layer. Similar to the pre-trained VGG neural network the size of convolutional kernels of all convolutional layers is 3*3, and the size of the pooling layer is 2*2. The size of the input image to the

convolutional layer is 84*84. fully-connected layer collects all inputs from convolutional pooling operation and gives as input to the softmax classifier, and finally, facial expression classification results are obtained. FER model trained on a dataset obtained after combining CK+ and FERG-DB datasets. This network will have the following components:

– Input Layer: This layer is used to take a raw 84*84 image as an input to the layer. Preprocessing operations such as face detection, gray-scaling, and image resizing are done on the image before it is fed to this layer.

– Convolution layer: These layers are the building blocks of the neural network, that computes a dot product between weights and small regions to which the neurons are connected to the input layer. Using convolutional kernels to achieve feature extraction from the input is a key role of a convolutional layer. Kernel of size 3X3 is used so the extracted features will be highly local. Smaller kernel size helps to capture smaller and complex features from the input image. As the kernel size is smaller more numbers of features can be extracted from the input.

– Activation functions: In the proposed neural network model Relu and Softmax are used as activation functions. This is applied to the output of each layer of the neural network. It is used to remove nonlinearities from the output of each layer of the neural network.

– Pooling Layers: In the proposed neural network pool size of 2*2 max-pooling is used. A large number of convolutional layers cause more computation time at each layer so to reduce that computational time pooling layer is used. The pooling layer reduces the computational time by down-sampling the dimensions. Down-sampling is done by combining similar features from input.

– Batch Normalization: This layer boosts the overall performance of the neural network. Batch-normalization normalizes the output of previous layers. This speeds up the training process.

– Dropout Layers: This layer is used to prevent model overfitting.

– Dense layers: This layer is present at the end of CNN. It performs decision making on the basis of feature data generated by all convolutional layers.

The workflow of the proposed FER model is shown in fig 6

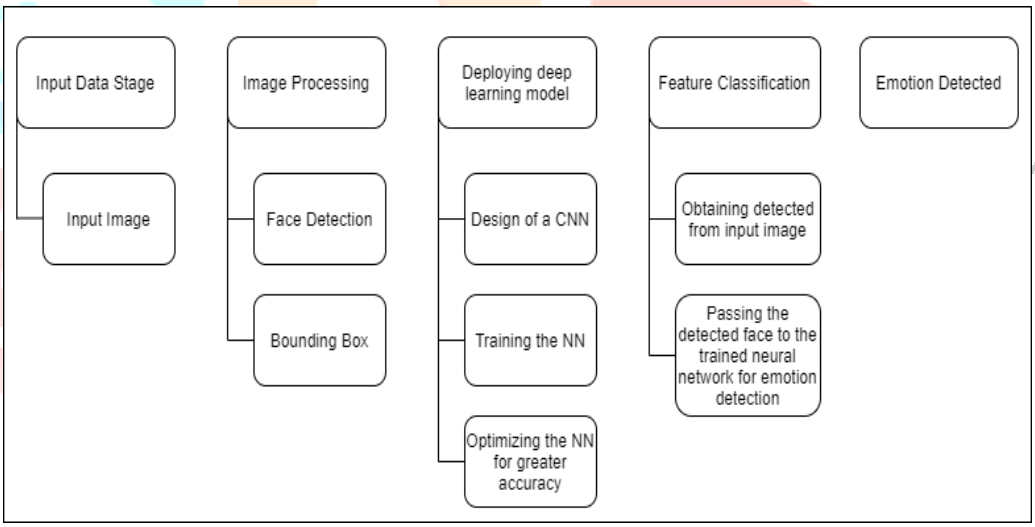
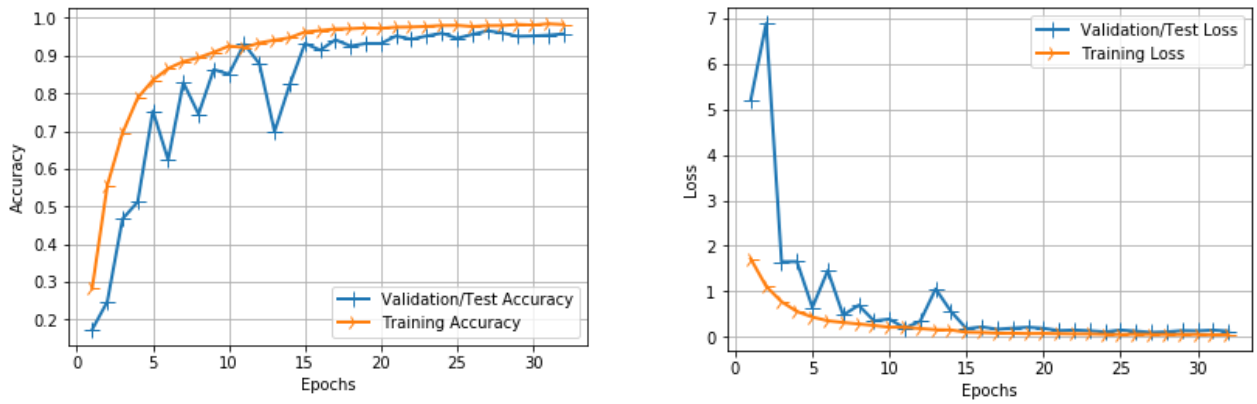


Figure 6 FER model working overview

IV. Results and Analysis

In order to prove the effectiveness of the proposed neural network for FER. The experiments were performed on the merged dataset of CK+ and FERGDB dataset, CK+ is a human facial dataset of 123 subjects whereas FERG-DB consists of facial data of six animated characters. The combined dataset includes a total of 10,548 images, that are subdivided into six basic emotions of human beings. This dataset includes six basic emotions such as Angry, Disgust, Happy, Neutral, Sad, and Surprise. Various preprocessing operations such as face detection and cropping, image resizing, and grayscaling are performed on data before it is being fed to neural networks. To prepare the dataset faces are cropped so that that the dataset is more concentrated on the facial image of a human being. The combined dataset is then divided into an 80% training set and a 20% validation set. The training set includes a total of 8433 images of six emotions and the validation set contains a total of 2115 images of six emotions. After selecting the image data, 8433 training images of size 84*84 are fed to the proposed neural network, and the model is trained for 40 epochs with batch size 62. While the training process of the neural network it is observed that, During the training phase of our model, it is trained on the same image data frequently and this frequent training on the same data helps it to learn features from the input data.



(A) (B)

Figure 7 Results of CNN: (A) Epoch Vs Accuracy (B) Epoch Vs Loss

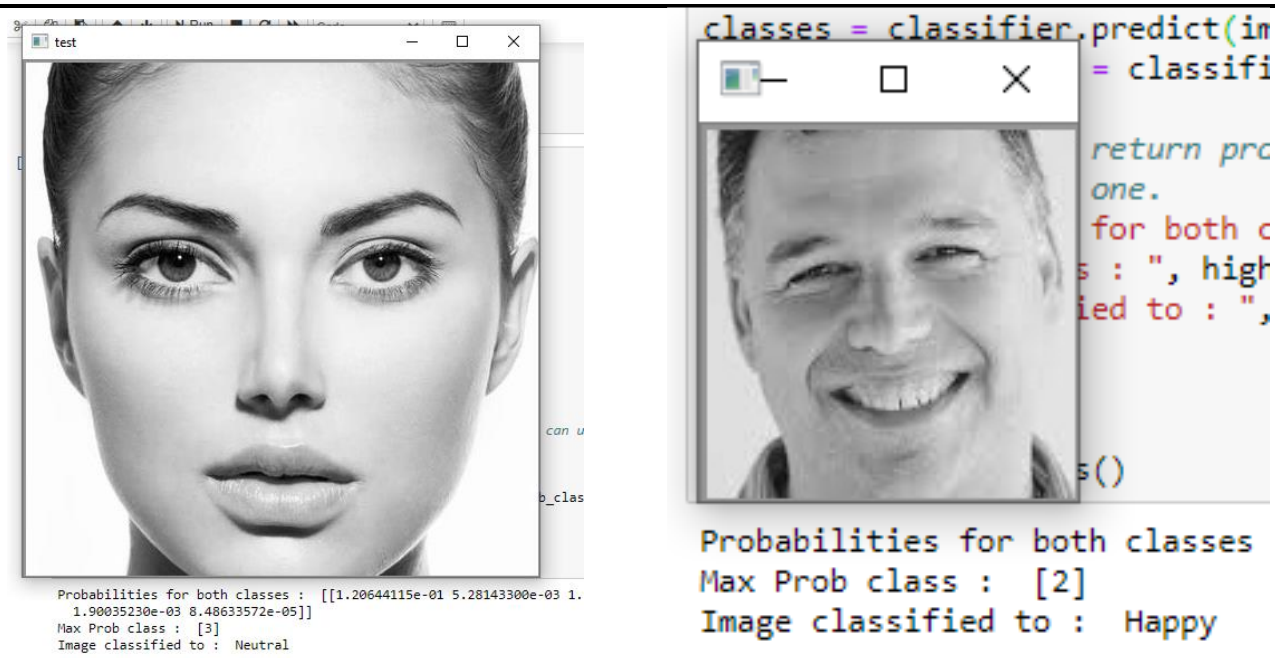
As can be seen from figure 7(A) Epoch Vs Accuracy, as with an increase in the number of epochs noted at the X-axis of the graph, validation accuracy denoted by the blue line is also increasing. Ups and downs in the validation line denote the change in learning rate. The orange line in the graph denotes the training accuracy, from the graph it is observed that at the end of 40 epochs validation accuracy reached very close to training accuracy it means our model has learned all important features for facial expression recognition from training data. At the end of 40 epochs model has achieved 95.60% accuracy.

As can be seen from figure 7(B) Epoch Vs loss, as with an increase in the number of epochs noted at the X-axis of the graph, validation loss denoted by the blue line is also decreasing. During the training process at the end of each epoch, the loss is calculated on the model's predictions. the model calculates error on each input by looking at what output is predicted for that input and taking the difference of that output value and then creates a label for that input. it is observed that with the increase in the number of epochs the loss is reducing gradually hence increase the accuracy of the model. The orange line in the graph denotes the training loss. At the end of 40 epochs validation loss of our model is 0.1176.

	No. of Images	Angry	Disgust	Happy	Neutral	Sad	Surprise	Accuracy
Angry	395	77	67	50	46	72	83	19%
Disgust	337	64	66	46	42	58	61	19%
Happy	351	55	47	71	52	75	51	20%
Neutral	294	62	49	45	51	45	42	17%
Sad	372	64	55	62	44	78	69	21%
Surprise	366	57	68	59	39	70	73	21%

Figure 8 Confusion matrix of CNN

This model has taken 2 days to complete 40 epochs on a dataset consisting of a total of 10,548. By looking at the confusion matrix from figure 8, it is observed that though the model has achieved accuracy seems to be promising but it does not perform well. More data need to be added to the dataset to improve the performance of the model. Form figure 8 it is observed that out of 2115 images of the validation set only 416 images are correctly classified to their respective emotions. Figure 9 shows test results of random images, our model firstly crops the face from the input image, convert it to grayscale then image fed to the model for the prediction, image is classified to max probability class. figure 9(A) shows ”Neutral” expression is detected similarly from figure 9(B) ”Happy” Expression is detected.



(A) (B)

Figure 9 Test results: (A) neutral expression (B) happy expression

V. CONCLUSION

Implementation details of an automated facial expression recognition system using deep learning technique are discussed in this paper. Facial expression recognition could help to improve the human-computer interaction so the computer could easily understand the human and could also mimic humans. How the architectural properties of the pre-trained neural network could be used to build a new neural network and how building a neural network by acquiring properties of two or more pre-trained CNN models will help to improve the accuracy of the model is also discussed in this paper.

REFERENCES

[1] L. Ma and K. Khorasani, "Facial Expression Recognition Using Constructive Feedforward Neural Network", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 34, NO. 3, JUNE 2004.

[2] Anima Majumder et al., "Automatic Facial Expression Recognition System Using Deep Network-Based Data Fusion", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 48, NO. 1, JANUARY 2018.

[3] Ji-Hae Kim et al., "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure", Received February 22, 2019, accepted March 20, 2019, date of publication March 25, 2019, date of current version April 11, 2019.

[4] Poonam Dhankhar, "ResNet-50 and VGG-16 for recognizing Facial Emotions", International Journal of Innovations in Engineering and Technology (IJIET).

[5] Burhanudin Ramdhani et al. "Convolutional Neural Networks Models for Facial Expression Recognition," 2018 International Symposium on Advanced Intelligent Informatics (SAIN).

[6] Minjun Wang, "Face Expression Recognition Based on Deep Convolution Network", 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2018).

[7] Guan Wang, Jun Gong, " Facial Expression Recognition Based on Improved LeNet-5 CNN", The 31th Chinese Control and Decision Conference (2019 CCDC).

[8] Sonali Sawardekar, Prof. Sowmiya Raksha Naik," Facial Expression Recognition using Efficient LBP and CNN ", International Research Journal of Engineering and Technology (IRJET).

[9] BIAO YANG, "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images", Received November 14, 2017, accepted December 9, 2017, date of publication December 15, 2017, date of current version February 28, 2018.

[10] FreeCodeCamp : How I developed a C.N.N. that recognizes emotions and broke into the Kaggle top 10 , https://www.freecodecamp.org/news/facial-emotion-recognition-develop-a-c-n-n-and-break-into-kaggle-top-10-f618c024faa7/

[11] Appliedmachinelearning : Demonstration of Facial Emotion Recognition on Real Time Video Using CNN : Python Keras ,https://appliedmachinelearning.blog/2018/11/28/demonstration-of-facial-emotion-recognition-on-real-time-video-using-cnn-python-keras/

[12] S. Mishra et al. "Emotion Recognition Through Facial Gestures — A Deep Learning Approach," Mining Intelligence and Knowledge Exploration, Nov. 28, 2017, Springer, Cham.

[13] R.Walecki et al. "Deep Structured Learning for Facial Action Unit Intensity Estimation," IJACSA, Apr. 14, 2017.

[14] Medium : Understanding and Implementing Architectures of ResNet and ResNeXt for state-of-the-art Image Classification: From Microsoft to Facebook , https://medium.com/@14prakash/understanding-and-implementing-architectures-of-resnet-and-resnet-for-state-of-the-art-image-cf51669e1624/

[15] Buzztech : Image Acquisition in digital image processing , https://buzztech.in/image-acquisition-in-digital-image-processing/