Classification of Facial Expressions using Machine Learning

Vatsal Patel
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Pratik Kanani
Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

ABSTRACT
Recognition of facial expressions is one of the most powerful and challenging tasks in Non-verbal communication. Normally major part of communication involves verbal Channels. But Non-verbal gestures are majorly expressed through facial expressions. Our project is based on classification of various human expressions using various types of Face Expression Recognition (FER) techniques which include the three major stages such as preprocessing, feature extraction and classification. We have carried out all these techniques using Convolutional Neural Networks (CNN). Our project is inspired by VGG and Xception model. Datasets used are FER 2013 (for emotion classification), IMDB (for gender classification), FEC (Google facial expression comparison). Using CNN, we classify 7 different expressions like Happy, Sad, Anger, Disgust, Fear, Surprise and Neutral.

Keywords
Preprocessing, Feature Extraction, CNN, VGG-16, Xception Model, Transfer Learning

1. INTRODUCTION
Facial Expressions are the minor signals that involves major communication. Humans mostly communicate through speech to express their emotions. But the emotions are also expressed through Non-verbal means like eye movements pressed eyebrows, slim and extended eyelids. Lips, eyebrow positions, body gestures, paralanguage, etc. Eye to eye connection is the significant period of correspondence which gives the combination of thoughts, Eye contact controls the conversations and makes a connection with others. Facial gestures include the smile, sad, anger, disgust, surprise, and fear. A smile on human face shows their bliss and it communicates eye with a bended shape. The sad expression is the inclination of detachment which is regularly communicated as rising slanted eyebrows and frown. The anger on human face is identified with terrible and disturbing conditions. The emotion of anger is expressed as pressed eyebrows, slim and extended eyelids. The disgust expressions are communicated with pull down eyebrows and wrinkled nose. The surprise or shock emotion is shown when some unpredicted situation has occurred. This is communicated with eye-enlarging and mouth opening wide and this articulation is an effortlessly distinguished one. The expression of fear is related with surprise gesture which is communicated as developing slanted eyebrows. In this project, FER is divided into 3 steps important stages which are Pre-processing, feature extraction and classification. Pre-processing involves detection of face which is carried out by Region of Interest Segmentation. Which is followed by smoothening of edges and converting image to the desired size. Feature extraction incorporates two types and they are geometric based and appearance based. Lastly, classification of expressions is carried out using Convolutional Neural Networks (CNN). Our project involves a 46 layer CNN which includes different layers like Input layer, Normalization layer, Activation layers, hidden layers, Convolutional layers, Max-pooling layers, Softmax layer, etc. We have used two different models namely VGG16 and Xception model. This is performed using Transfer Learning. Training of images involves steps like Image classification and Back propagation. In past few years there have been various advancements for facial expression classification. We observed that most of them involved 6 basic expressions which excluded the neutral expression. There are some models which detects neutral expression but that involves very less accuracy and a large amount of output delay. The aim of our project is to classify expressions like Happy, Sad, Anger, Disgust, Fear, Surprise and Neutral using CNN that gives high accuracy, robustness and negligible delay. We have used various databases FER, FEC and IMDB. The libraries used are OpenCV and Keras.

2. RELATED WORK
In recent years there have been various developments in the field of Facial Expression Recognition. Researchers have made a considerable progress in developing various models. Some involves training of CNN models with different gray scale images. Using Graphic Processing Unit (GPU) for increasing quality of training processes. This method allowed training of mixed feature data that included raw pixel data and Histogram Oriented Gradients features. Although the overall accuracy achieved by them was 51%. And had very high output delay [1].

In [2] researchers developed a real-time CNN model for facial expressions. They merged various datasets like JAFFE, KDEF and their own dataset. They trained images on LeNet architecture for classification of expressions. Intel 17 was used for training datasets and other experiments. The overall accuracy of the model was 91.81% but the accuracy of neutral expression was approximately 30%.

In [3] researchers worked on local and global features for classification of Facial Expressions. This system involved fusion of global and local features with CNN. Modified AlexNet, VGG Net, ResNet were used for verifying training. They used parallel convolution to integrate global AND local features. FER 2013 datased was used which composed of 28,709 images. As a result accuracy obtained by AlexNet, VGG Net and ResNet were 63.57%,64.69% and 68% respectively. This model required improvements in accuracy and output delay rate.

In [4] researchers developed a Facial Expression Recognition model using CNN and image edge computing. This system involved classification of 6 basic expressions. This system wasn't suitable for detection of neutral expression. This model was developed to avoid complex processes in feature

extraction. Here, the images were normalized and edges of each images underwent through convolutional layers. These extracted images were convolved with the other texture images and then Max-pooling was performed. Then finally expressions were classified in the Softmax layer. This proposed algorithm had a recognition rate of 88.56%.

3. METHODOLOGY

3.1 Proposed Design

The FER system comprises of the major stages such as face image pre-processing, feature extraction and classification.

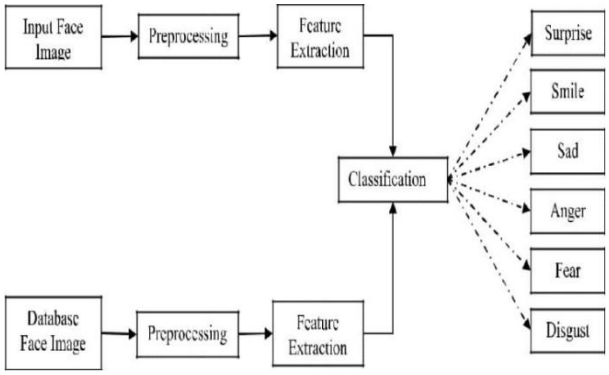

Fig1: Architecture of face expression recognition system

3.2 Preprocessing using ROI segmentation

Preprocessing is a method which can be utilized to improve the exhibition of the FER framework and it tends to be completed before including extraction measure. Picture preprocessing incorporates various sorts of steps, for example, picture lucidity and scaling, contrast change, which involves various enhancement steps to improve the articulation outlines.

In Preprocessing, ROI (Region of Interest) Segmentation is one of the significant kindof preprocessing technique which incorporates three significant capacities, for example, directing the face measurements by partitioning the shading segments and of facial picture, eyes and mouth area division. In FER, ROI Segmentation is most used because it carries out most efficient division of face organs from the given images.

In FER, more preprocessing techniques are utilized however the ROI Segmentation method is more appropriate on the grounds that it recognizes the face organs precisely which are the organs responsible for expression recognition. Next the histogram equalization is likewise another significant preprocessing strategy for FER on the grounds that it improves the image quality.

Preprocessing involves following steps: First Load an Image. Apply Skin Color Segmentation. Detection of Connected Region. Recognition of Face from the Image. Creating Binary Image from RGB Image. Face Detection from Binary Image. Region of Interests (ROI). Identification of Feature Points. Objectives of ROI based segmentation are: To recognize face from the image supplied. To identify the desired region to work with different parts of facial image. To find the shape curves from each desired region [5].

3.3 Feature Extraction and Classification using CNN

Classification is the last step of FER architecture in which the system classifies the expression such as smile, sad, surprise, anger, fear, disgust and neutral. Support Vector Machine (SVM) is one of the classification techniques in which two types of approaches are involved [6]. They are one against one and one against all methodologies. One against all arrangement implies it develops one example for each class. One against all classification means it constructs one sample for each class. One against one classification means it constructs one class for each pair of classes and SVM is one of the strongest classification methods for advanced dimensionality troubles. SVM is the directed Artificial Intelligence method and it utilizes four sorts of kernels for its better presentation. They are direct, polynomial, Radial Basis Function (RBF) and sigmoid.

Convolution Neural Network (CNN) comprises of two layers, for example, convolutional layer and subsampling layer in which the two-dimensional pictures are taken as information. In convolutional layer the element maps are created by perplexing the convolution bits with the two-dimensional pictures where as in the subsampling layer, pooling and redeployment are performed. The CNN likewise contains two significant discernments probably shared weight and inadequate network. In FER, the CNN classifier is utilized as various classifiers for the distinctive face locales. On the off chance that CNN is outlined for whole face picture, at that point first casing the CNN for mouth region and next for eye zone likely for one another zone CNNs are outlined Deep Neural Network (DNN) contains different hidden layers and are more difficult to be trained. According to several classifiers CNN gives better recognition accuracy and it provides better classification. In FER, CNN classifier is more exploitable comparing with other classifiers for recognition of expressions [7].

Following are **Layers of CNN:**

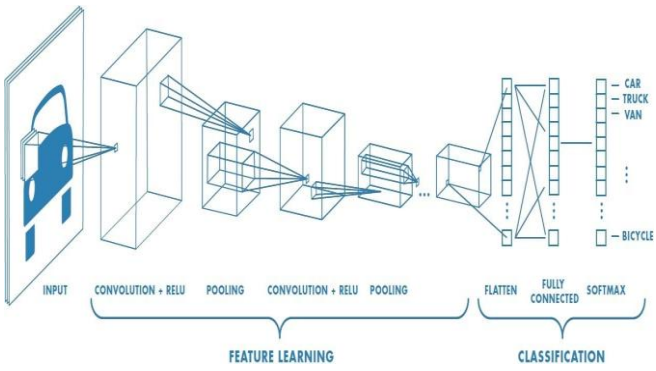

Fig2: Layers of CNN

3.3.1 Input layer

It will read the given image. There is one input layer in our system.

The first layer is always convolutional layer. The input is 480*480*3 array. A filter or kernel is moved over this array and the area over which it is moved is called as receptive field. The output of this layer forms a feature map which is also called as activation layer. The more filters are used the more spatial features are detected. There are 6 convolutional layers in our system [8].

3.3.2 Convolutional layers (2d)

The first layer is always convolutional layer. The input is 480*480*3 array. A filter or kernel is moved over this array and the area over which it is moved is called as receptive field. The output of this layer forms a feature map which is also called as activation layer. The more filters are used the

more spatial features are detected. There are 6 convolutional layers in our system [8].

3.3.3 Activation layer

The activation layer used hers is ReLU (Rectified Linear Unit). The output of the first convolutional layer becomes input for this activation layer. The output of this layer will be activations that represent higher level features such as semicircles, squares, combination of different curves and many more. This layer adds non linearity to the image. It also avoids overfitting. This term alludes to when a model is so tuned to the training models that it can't sum up well for the approval and test sets. A manifestation of overfitting is having a model that gets 100% or 99% on the preparation set, however just half on the test information. There are 14 hidden layers in our framework [9].

3.3.4 Normalization layer

A batch normalization layer normalizes each input channel across a mini-batch. To speed up training of convolutional neural networks and reduce the sensitivity to network initialization, use batch normalization layers between convolutional layers and nonlinearities, such as ReLU layers. There are 9 normalization layers in our system.

A. Separable convolutional layer:

There are 9 separable convolutional layers in our system. This process is broken down into 2 operations:

B. Depth-wise convolutions

In depth-wise convolution, convolution is applied to a single kernel at a time unlike standard CNN's wherein it is accomplished for all the K channels. So here the channels/portions will be of size Dm x Dm x 1. Given there are K channels in the input, at that point K such channels are required. Yield will be of size Ds x Ds x K.

3.3.5 Point-wise convolutions

In point-wise method, a 1×1 convolution activity is applied on the K channels. So the channel size for this activity will be 1 x 1 x K. Let's assume we use P such channels, the yield size becomes Ds x Ds xP.

3.3.6 Maxpooling:

It is also called as a down sampling. It considers a kernel (normally of size 2x2) and a stride of the equal length. It is then integrated with the input data and outputs the maximum number in every sub region that the kernel convolves around. There are 4 Maxpooling layers in our system.

3.3.7 Average Pooling

An average pooling layer performs maxpooling by partitioning the data to rectangular pooling locales and registering the average estimations of every pixel. There is 1 Average Pooling layer in our system.

3.3.8 Softmax:

Softmax is an actuation signal like sigmoid, tanh, and ReLU commonly applied on the yield of the last layer. It is especially helpful in multiclass arrangement when the information must be one, and just one class. This is on the grounds that softmax restores a discrete probability distribution over all the classes. There is 1 Softmax layer in our framework.

3.4 Datasets and Features

In this project, we have considered three datasets namely FER 2013 for facial expression recognition, IMDB for gender classification and FEC [10]. FER dataset comprises of 23,000 images with 48*48 dimension. The pictures are prepared so that the appearances are practically focused and each face possesses about the same measure of room in each picture. Each picture has to be classified into one of the seven classes that express unique facial expression. These facial emotions have been ordered as: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral.

Figure 3 portrays sample images from the dataset. Notwithstanding the picture class number (a number somewhere in the range of 0 and 6), the given pictures are partitioned into three unique sets which are training, validation and test sets. There are around 13,000 training pictures, 5,000 validation pictures, and 5,000 pictures for testing. Subsequent to perusing the raw pixel information, we standardized them by deducting the mean of the training pictures from each picture involving those for the validation and test sets.

Fig 3: Sample images from dataset

4. EXPERIMENTS AND RESULTS

Preprocessing involves various steps like detection of face, cropping, batch normalization, conversion of RGB image to gray scale image, resizing of image.

1. ROI segmentation and crop: We load an image.Then skin color segmentation is applied followed by detection of connected region. Lastly, face is recognized from the image. OpenCV cascade classifier is utilized to distinguish face from the given pictures. After recognizing the face, that part has been cropped out to keep away from foundation unpredictability so the model training turns out to be more proficient [11].

2. Batch normalization: Batch normalization has been applied to demonstrate dataset which is a measure that changes the scope of pixel force esteems to a certain breaking point. It is a cycle by which difference or histogram of the pictures can be extended with the goal that it empowers deep networks to break down the pictures in a superior manner [12].

3. RGB to gray scale: Pictures have been resized into 48*48 pixels having 3 channels red, green and blue. To diminish the unpredictability in pixels, dataset pictures have been changed over into grayscale having as it were one channel. So it becomes essentially simple for the model to learn [13].

4. Resizing of image and denoising: Image is converted to the required size. Denoising results in blurring of image using Gaussian features.

Fig 4: Dataset Preprocessing

For this project, we shall use Transfer Learning to train CNN models. Transfer Learning is a process of training pre-trained CNN architectures with our own developed dataset. Due to this the pre-trained model will perform as feature extractor. We eliminate outermost layer and replace it with our won classification layer. The pre-trained models used are VGG-16 and Xception architecture.

a. VGG-16 architecture: It is 19 layer CNN. It uses 3*3 filter with stride and padding of 1. It has Maxpooling of 2*2. The main advantage of this architecture is it uses a small sized filter, so if 2 filters are used in place of one filter it will give better results and the number of parameters required will be less [14].

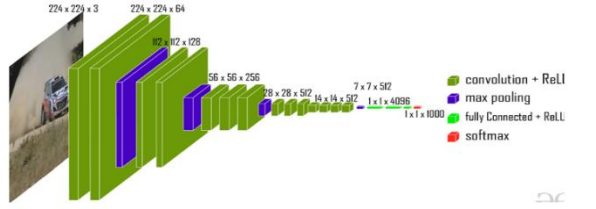

Fig5: VGG-16 Architecture

b. Xception architecture: It is a deep convolutional network which has depth wise separable convolution followed by pointwise convolution. The information first flows through the entry flow followed by middle flow. This step is repeated 8 times, then finally data flows through exit flow [15].

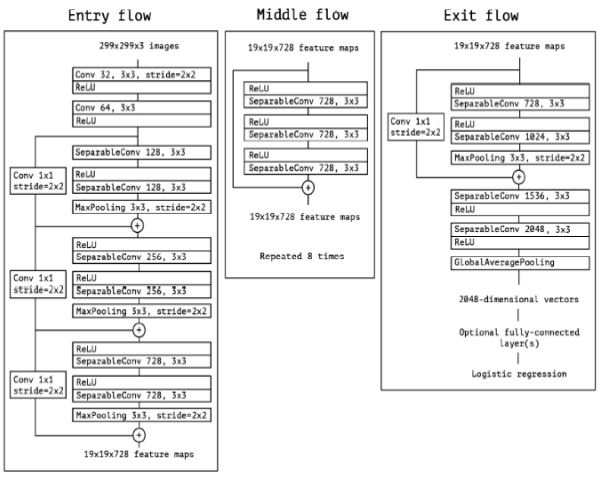

Fig6: Xception Architecture

Followed by training of proposed CNN model, the model underwent through real time testing. Most importantly, human images were recognized with the Haar Cascade library inside 30 pictures for each second of the camera. From that point forward, the distinguished pictures were shipped off the model and the classes they have a place with were questioned. Because of the prior assumptions, the chance of having a place with which class the outward appearance was appeared on a different screen and the expression in which class was higher was overwritten on the Haar Cascade outline. This cycle was performed on each 30 frames that happened each second of the camera picture captured in real time.

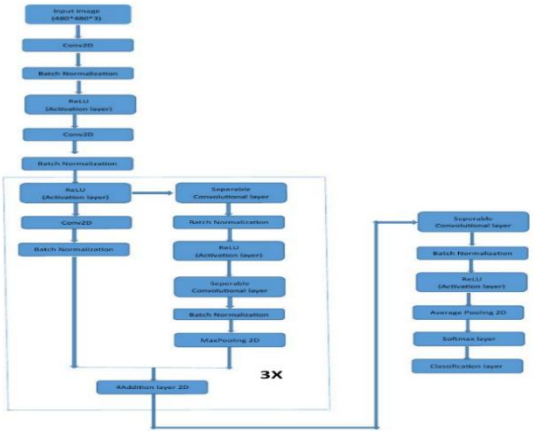

Fig7: Our CNN Model

Fig8: Real time testing outcomes

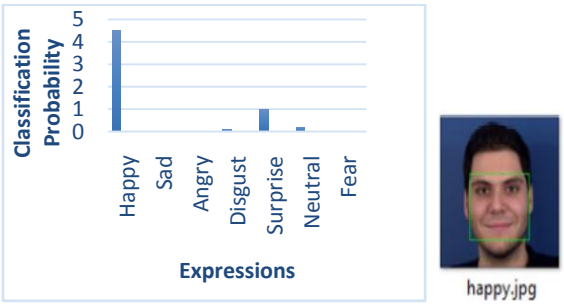

Fig9: Expression Classified: Happy

Here, in fig.9 we observed that accuracy for detection of a happy expression is 94%. As with happy the system shows a possibility of other expressions like surprise, neutral and disgust.

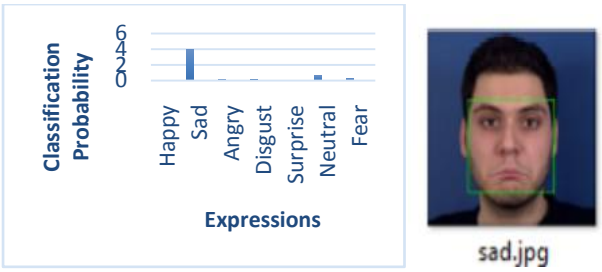

Fig10: Expression Classified: Sad

Here, in fig.10 we observed that accuracy for detection of a sad expression is 73%. As with sad the system shows a possibility of other expressions like angry, fear, neutral and disgust.

Fig11: Expression Classified: Angry

Here, in fig.11 we observed that accuracy for detection of a angry expression is 75%. As with angry the system shows a possibility of other expressions like surprise, sad, fear and disgust.

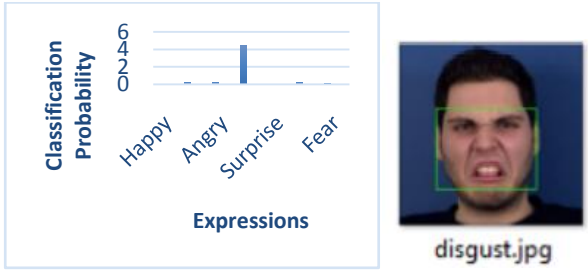

Fig12: Expression Classified: Disgust

Here, in fig.12 we observed that accuracy for detection of a disgust expression is 81%. As with disgust the system shows a possibility of other expressions like fear, sad, neutral and angry.

Fig13: Expression Classified: Surprise

Here, in fig.13 we observed that accuracy for detection of a surprise expression is 90%. As with surprise the system shows a possibility of other expressions like happy, angry, fear and disgust.

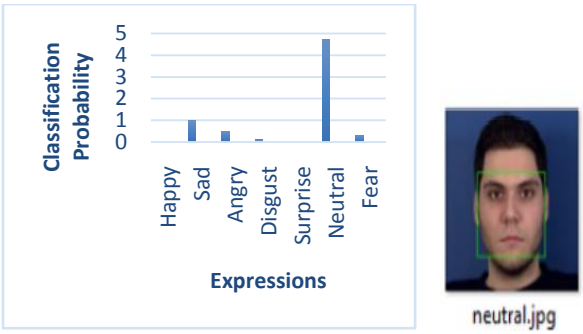

Fig14: Expression Classified: Neutral

Here, in fig.14 we observed that accuracy for detection of a neutral expression is 93%. As with neutral the system shows a possibility of other expressions like sad, angry, fear and disgust.

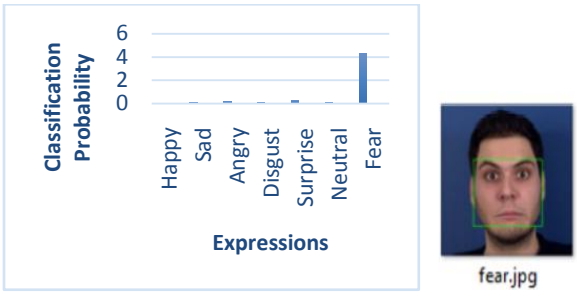

Fig15: Expression Classified: Fear

Here, in fig.15 we observed that accuracy for detection of a fear expression is 86%. As with fear the system shows a possibility of other expressions like angry, sad, surprise, neutral and disgust.

Table 1: Accuracy rates of each expression

Expression	Accuracy
Happy	94%
Sad	73%
Disgust	81%
Fear	86%
Anger	75%
Surprise	90%
Neutral	93%

According to our survey depending upon the complexity rates, accuracies, availability of processing methods and number of features extracted the techniques used in our method are:

1. Preprocessing: ROI Segmentation method will be used as it gives 99% accurate results.
2. Feature Extraction: CNN have less complexity which gives the accuracy always between 82.5% and 99%.
3. Classification: The highest recognition accuracy of 99% is provided by the CNN classifier and it recognizes the several expressions such as disgust, sad, smile, surprise, anger, fear and neutral

effectively.

4. In 2D FER, mostly IMDB and FEC database are used for efficient performance than the other databases.

5. CONCLUSION

Thus, we have designed an algorithm for Human Facial Recognition using various techniques such as ROI segmentation for preprocessing and Convolutional Neural Networks for feature extraction and classification of expressions with much less output delay than the other algorithms and with a higher accuracy. This system will classify expressions like happy, sad, anger, disgust, fear, neutral and surprise for still images as well as real time images and videos with a accuracy of 93.33% approximately.

We have proposed a general building design for creating real-time CNNs and especially for detection of neutral expression. Our proposed architectures have been systematically build in order to reduce the amount of parameters. We started by disposing completely the fully connected layers and by lowering the amountof parameters in the remaining convolutional layers via depth-wise separable convolutions to give a less output delay and better precision. We have demonstrated that our proposed models can be stacked for multi-class arrangements while handling continuous real-time situations which was lacking in prior systems. Specifically, we have developed an architecture that will performs face detection and emotion classification in a single model. We can accomplish useful execution of our model utilizing just one CNN that uses advanced tools. Our architecture reduces the amount of parameters 80x while obtaining favorable results. Our complete model can be fabricated in a Care-O-bot 3 robot. At last, we learned highlights in the CNN utilizing the guided back-propagation perception. This perception method can show us the complex features learned by our models and examine their interpretability.

6. FUTURE SCOPE

Today, one of the fields that utilizes facial expression recognition the most is security. Expression Recognition is an extremely efficient instrument that can help law masters perceive lawbreakers and programming organizations are utilizing this to help clients access their information. This technology can be additionally evolved to be utilized in different roads, for example, ATMs, getting to secret documents, or other delicate materials. This can make other safety efforts, for example, passwords and keys old. Another way that trend-setters are hoping to actualize expression recognition is inside metros and other transportation outlets. They are hoping to use this tool to utilize faces as MasterCard's to pay for your transportation charge. Rather than going to a stall to purchase a ticket, the face recognition would take your face, run it through a framework, and charge the record that you've recently made. This might actually smooth out the cycle and advance the progression of traffic radically. Health psychological science is the investigation of activity measures in wellbeing, disease, and medical care. This field of health psychology research is fundamentally centered around wellbeing just as the expectation and treatment of sickness and ill health. This FER system can also be used as a feedback tool. It can be incorporated in shopping websites and apps to know customer's reaction on their product.

7. REFERENCES

[1] Z. Xie, Y. Li, X. Wang, W. Cai, J. Rao and Z. Liu, "Convolutional Neural Networks for Facial Expression Recognition with Few Training Samples," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 9540-9544, doi: 10.23919/ChiCC.2018.8483159.

[2] K. Liu, C. Hsu, W. Wang and H. Chiang, "Real-Time Facial Expression Recognition Based on CNN," 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2019, pp. 120-123, doi: 10.1109/ICSSE.2019.8823409.

[3] Shengtao, Gu & Chao, Xu & Bo, Feng. (2019). Facial expression recognition based on global and local feature fusion with CNNs. 1-5. 10.1109/ICSPCC46631.2019.8960765.

[4] Zhang, Hongli & Jolfaei, Alireza & Alazab, Mamoun. (2019). A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2949741.

[5] Bashyal, S., Venayagamoorthy, G.K.V., 2008. Recognition of facial expressions using Gabor wavelets and learning vector quantization. Eng. Appl. Artifi Intelli 21, 1056—1064.

[6] Proc. 17th ACM Int. Conl Multimed. pp. 569-572. Chang, H.T.Y., 2017. Facial expression recognition using a combination of multiple facial features and support vector machine. Soft Comput. 22, 4389-4405. https://doi.org/l().100%00500-017-2634-3.

[7] S. Singh and F. Nasoz, "Facial Expression Recognition with Convolutional Neural Networks," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0324-0328, doi: 10.1109/CCWC47524.2020.9031283.

[8] Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9.

[9] Nwankpa, Chigozie & Ijomah, Winifred & Gachagan, Anthony & Marshall, Stephen. (2020). Activation Functions: Comparison of trends in Practice and Research for Deep Learning.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in null. IEEE, 2001, pp. 511.

[11] R. C. Gonzalez, "Digital image processing/richarde," Woods. Interscience, NY, 2001.

[12] M. Grundland and N. A. Dodgson, "Decolorize: Fast, contrast enhancing, color to grayscale conversion," Pattern Recognition, vol. 40, no. 11, pp. 2891–2896, 2007.

[13] Srikanth Tammina (2019); Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images; International Journal of Scientific and Research Publications (IJSRP) 9(10) (ISSN: 2250-3153), DOI: http://dx.doi.org/10.29322/IJSRP.9.10.2019.p942.

[14] Chollet, Francois. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 1800-1807. 10.1109/CVPR.2017.195.