Facial Expression Recognition robust to partial Occlusion using MobileNet

Sreelakshmi P
M. Tech Student
Dept. of. CSE
LBS Institute of Technology for Women
Trivandrum, India

Sumithra. M. D
Associate Professor
Dept. of. CSE
LBS Institute of Technology for Women
Trivandrum, India

***Abstract*—Facial Expression Recognition have attracted the attention of researchers over a few decades. Many existing automatic Facial Expression Recognizing system use standard machine learning approaches for the purpose of feature extraction and classification and found difficult to handle partial occlusions and merely generalize unseen data. This paper proposes a Convolution Neural Network (CNN) based Facial Expression Recognition system which can handle both partial occlusions and pose variations. The proposed method efficiently interpret the information available in the face images automatically without explicitly providing the feature descriptors. Here we aim to develop an efficient Facial Expression Recognition model using MobileNet architecture of CNN by appling the transfer learning technique.**

Keywords*— *Facial Expression Recognition, partial occlusion, CNN, MobileNet

I. INTRODUCTION

Facial Expressions are the non-verbal way of communication between humans to convey ones intentions or emotional state. The process of identifying the emotions expressed by the human-beings is termed as Facial Expression Recognition. Humans have the ability to recognize the emotions of others and can respond to them naturally. In the field of artificial intelligence systems that can recognize human emotions have shown an interesting progress over the few years. Automatic facial expression Recognition is a very challenging and difficult task which has a vast range of applications such as gaming, emotionally sensitive robots, personal assistance provision, patient monitoring, security, criminal interrogation,online tutoring, human-computer interactions and many more. Eventhough human beings can perceive and able to recognize these emotions virtually, an automatic facial expression recognition system still lacks the ability for reliable emotion recognition. channels,potholes on the roads etc are also major threats and almost people are prone to such severe accidents irrespective of their age,health and other matters.

The facial expressions are developed by the movement of some facial muscles which may cause the movement of face skin or changes in the facial feature appearances. The recognition of the human emotions is not an easy task since there many be variations for the same emotions shown by different people and can also vary due to pose variations, the lightning conditions, shape, partial occlusions etc. But there is some universality in expressing some emotions by human.

The facial expressions are classified to six universally accepted emotions that are happy, sad, anger, disgust, fear and surprise. The state-of-art methods consists of mainly three steps in automatic Facial Expression Recognition such as face identification, feature extraction and classification. But these processes find difficult when they subjected to some challenges like pose variations, partial occlusions, gender variations, illumination changes and also when considering the real-world scenarios. So, if a well trained system is developed considering these challenges and the univerality among emotions it can be effectively used for human-computer interactions.

An automatic Facial Expression Recognition system have a wide variety of applications in many areas like medical diagnosis, entertainment, stress level assessment, personal service provision and importantly in human-computer interactions. For the Artificial Intelligence systems to attain human-level interactions they need also the emotion intelligence inorder to effectively communicate, interact and coexist with the humans. Such intelligence will be beneficent for many real-world applications.

Convolutional Neural Networks(CNN) recently shown striking progress in the area of object classification. The existing CNN based Expression Recognition systems are complex in nature and difficult to use in the mobile devices and embedded vision applications. For many real-world applications like robotics, driver fatigue detectors, patient monitoring etc, the facial expression recognition process has to be done well-timed and on computationally limited platforms. In this paper, we aims to design a Facial Expression Recognition system which can jointly address the challenges partial occlusion and pose variations using MobileNet, a class of Convolutional Neural Networks and suitable for implement in mobile devices. This projects deals with seven expression categories happy, sad, angry, fear, disgust, surprise and neutral

II. RELATED WORK

Facial Expression Recognition methods that can handle partial occlusions have studied by researchers during past few years. Towner et al.[1] introduced a feature reconstruction based approach which uses three Principal Component Analysis(PCA) methods to reconstruct the missing parts on the faces. Then the features are fed to a Support Vector

Machine (SVM) for classifying among six basic emotions.

Facial Expression Recognition system introduced by Haung et al.[2] integrates the component-based feature descriptors, sparse-representation based occlusion detection and a video sequence based weight learning for facial expression. Spatio-Temporal Local Binary Pattern(STLBP) and Edge Map is used for the feature extraction from face regions on the video sequences. The obtained feature vectors are concatenated into one by using a multiple feature fusion. This method achieved 93%, 79.08% and 73.54% recognition rate for eyes, mouth and lower face regions with occlusions on CK+ dataset.

Recently Deep learning techniques attained increased attention in the field of facial expression recognition. Xu et a.[3] came with an Facial Expression Recognition model which has robustness to partial occlusion. The framework rely on transfer features from trained deep convolutional networks (ConvNets) which has four convolutional layers. The high-level features of the trained two ConvNets are merged and fed into an SVM classifier to recognize the six emotions plus neutral. The model achieved an average of 81.50% accuracy on a self-build dataset. Mollahosseini et al.[4] trained used Inception architecture of CNN to acheieve state-of-art results in Facial Expression Recognition. The network consists of two convolutional layers, max-pooling, and 4 Inception layers and they tested on publically available datasets.

III. CONVOLUTIONAL NEURAL NETWORK

Deep neural networks are inspired by the connectivity patterns of neurons in human brain. Conventional neural networks are not suitable for the image processing tasks and also fed images in reduced-resolution pieces. A convolutional Neural Network (CNN or ConvNets) is a class of deep neural network used in image recognition and many computer vision applications. CNN has shown better accuracy rates on difficult classification tasks which requires concrete details of the images. The best thing about CNN is that they can learn relevant features of an image or video at various levels just like the human brain. A ConvNet is able to capture spatial and temporal features of an image by using different filters regardless of where they are located. Since the same feature extractor can be learned for every location in an image the parameters requires for the feature extraction process can be reduced tremendously. The aim of CNN is to convert the images into an easily processing form without compromising the important features that are crucial for a good prediction process. The minimal parameter requirement and reusability of weights provide better fitting to the image datasets. The efficiency of CNNs provide major advances in Computer Vision applications like robotics, medical field, secutiry, entertainments and many more.

A Convolutional Neural Network consists of an input layer, output layer and hidden layers. The hidden layer includes convolution layer, pooling layer and fully connected layers. The color images are perceived by the network as rectangular box where the number of pixels along that direction represents its width and height and depth will be three which represent each layer of RGB. Instead of focusing on single pixels a ConvNet takes square patches of pixels and fed to the filter. In convolution layer the input matrix is passed to a convolutional filter which slides over the image inorder to find out their dot product. The filter is a square matrix of size lesser than the input matrix which is also called kernel. As the filter is sliding or convolving over the input image it multiplies the values in filter with the input image pixel values and the values are summed up to produce a new matrix called feature map or activation map. The Pooling layer is used inorder to reduce the spatial size of feature maps by retains the most important information and hence speedup the computation process. Max pooling outputs the maximum value of the subregion while Average pooling outputs its average value. The fully connected layer is similar to the multi-layer perceptron which connect every neuron in the previous layer to every neuron in other layer. The purpose of this layer is to receive high-level features from previous layers and use them to classify the images to different classes.

IV. MATERIALS AND METHOD

A. *Image Preprocessing*

In the preprocessing step, from the raw image the face detection is done and it is cropped inorder to get maximum area of face. The face detection is done using the Haar Cascade classifier which is present in the opencv library. The bounding box is adjusted around the image to get maximum area of face by cropping the face image. MobileNet architecture takes input image size 224x224 as well as the features should be scaled between [-1,1]. The loaded input image is resized and we feature scaled the image tensor values using center value of an image pixel range [0,255], ie.127.5. Then we subtract the offset value for each pixel in the image and divide it by the offset value inorder to scale between [-1,1].

B. *Convolution Neural Network Architecture*

The tremendous progress of deep learning and neural networks in the image processing and computer vision pave the way for its increased users. While some recognition and detection technologies are provided for on-devices with the help of internet connection, they poses many challenges like the computational power and their availability at anytime, anywhere despite of the need of internet. In 2017 google introduced the MobileNet which can be used on on-devices and embedded applications with the advantages of minimum computational power, time, space and better accuracy. MobileNets has many advantages over the state-of-art architectures like ResNet, VGG16, Inception and Xception that they have a reduced network size of 17 MB and a reduced number of parameters ie,4.3 million.

MobileNet architecture is based on Depthwise separable convolution. This reduces the number of parameters while compared to other networks which has the same depth and thus results in a light-weight neural network. The normal convolution process is divided by a depthwise convolution followed by a pointwise convolution which is altogether termed as depthwise separable convolution. The mobileNet V2 is used in this model for the Facial Expression Recognition process. Its main building block consists of three convolutional layers. The depth-wise convolution layer filters the input followed by a 1x1 pointwise convolution layer. Pointwise convolution is also termed as Projection layer which will reduce the large number of channels into a tensor which has less number of channels. The first layer is a 1x1 convolution with the purpose of the number of channels in the data before it going for depth-wise convolution. Hence this layer is called Expansion layer which has larger number of output channels than input ones. The hyperparameter, expansion factor is used to decide how much data to be expanded. Here we use an expansion factor of 6. To the expanded tensor channels filters in depthwise convolution is applied. And finally, the projection layer projects the number of channels to a small value, ie, the initial number of tensor channels. Also, the MobileNet V2 has a residual connection which help in the flow of gradients through the network. As in all CNN, it has batch normalization and a ReLU6, which is the activation function. Since the projection layer produce low-dimensional data it doesn't have any activation function.

The full architecture consists of totally 17 blocks in a row which is followed by a 1x1 convolution, a global average pooling and a classification layer that give the output as the most probable emotion out of the seven (happy, sad, angry, fear, surprise, disgust and neutral).

C. Fine Tuning

Training the CNN model using a smaller dataset will affect its ability to generalize and thus result in overfitting problem. So the easier way is to fine-tune the existing network that are trained on some larger dataset. The MobileNet is pretrained on the ImageNet which has more than 1000 classes. But the data in it is different from the content that we require. MobileNET is trained to perform several object detection but not on faces. We fine-tuned the MobileNet using the datasets FER2013, JAFFE, FED-RO. Since it is pretrained it helps to save a lot of computation time and power by avoiding training from the scratch. The initial layers of the networks will learn very general features and when we go to the higher layers, they tend to learn more specific patterns of the task it is trained on. So, in fine-tuning the initial layers will remain same,ie, no change will be done to these layers and the last layers are retrained for our expression recognition task.

D. Training the model on partially occluded images

The classifiers where trained using Keras. Feature-wise mean subtraction and normalization is done on the input image data. Data augmentation like horizontal shifts, vertical shifts and random horizontal flips are done.

We trained for 15 epoch in each run. The full run history are saved and the models best weight is saved in the HDF5 format. We initially fine tuned on only small percentage of the total data and found out that these parameters is not generalized to the full dataset. So, last half of parameter tuning is done with more than 60% of the data in full datasets. When the classifier is trained using only unoccluded images the performance will be very severe. This is one of the case in domain adaptation problem where training and testing data are from different distribution types. Here we train our classifier on partially occluded images also inorder to make the classifier robust to partial occlusion. First the pretrained MobileNet model is loaded in Keras and indicate which layers are needed to be trained by setting the trainable parameters. This is the base network and the classifier is added on the top of the base network by adding a fully connected layer followed by a softmax layer. This produced seven output classes which are happy, sad, angry, fear, surprise, disgust and neutral. The datas to be given is separated and put in two folders for training which are 'train' and 'validation' folders. The ImageDataGenerator available in Keras is used to read the images in batches from these folders directly and data augmentation is done optionally for this. The batch size given here is 32. The learning rate is indicated by the programmer to indicate the weights convergence on suitable value. Learning rate of this model is 0.01. After the data setup is done the training is done and the model is saved. We have achieved greater accuracy by using this transfer learning approach.

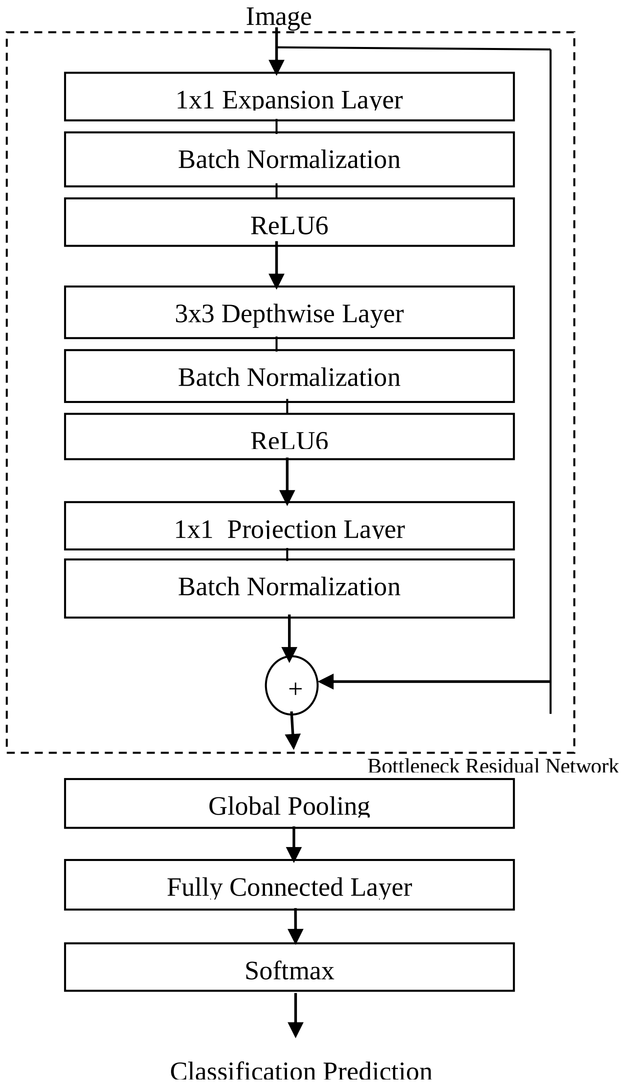

Fig 1: Working of MobileNet V2

The full architecture consists of totally 17 blocks in a row which is followed by a 1x1 convolution, a global average pooling and a classification layer that give the output as the most probable emotion out of the seven (happy, sad, angry, fear, surprise, disgust and neutral).

E. Fine Tuning

Training the CNN model using a smaller dataset will affect its ability to generalize and thus result in overfitting problem. So the easier way is to fine-tune the existing network that are trained on some larger dataset. The MobileNet is pretrained on the ImageNet which has more than 1000 classes. But the data in it is different from the content that we require. MobileNET is trained to perform several object detection but not on faces. We fine-tuned the MobileNet using the datasets FER2013, JAFFE, FED-RO. Since it is pretrained it helps to save a lot of computation time and power by avoiding training from the scratch. The initial layers of the networks will learn very general features and when we go to the higher layers, they tend to learn more specific patterns of the task it is trained on. So, in fine-tuning the initial layers will remain same,ie, no change will be done to these layers and the last layers are retrained for our expression recognition task.

F. Training the model on partially occluded images

The classifiers where trained using Keras. Feature-wise mean subtraction and normalization is done on the input image data. Data augmentation like horizontal shifts, vertical shifts and random horizontal flips are done.

We trained for 15 epoch in each run. The full run history are saved and the models best weight is saved in the HDF5 format. We initially fine tuned on only small percentage of the total data and found out that these parameters is not generalized to the full dataset. So, last half of parameter tuning is done with more than 60% of the data in full datasets. When the classifier is trained using only unoccluded images the performance will be very severe. This is one of the case in domain adaptation problem where training and testing data are from different distribution types. Here we train our classifier on partially occluded images also inorder to make the classifier robust to partial occlusion. First the pretrained MobileNet model is loaded in Keras and indicate which layers are needed to be trained by setting the trainable parameters. This is the base network and the classifier is added on the top of the base network by adding a fully connected layer followed by a softmax layer. This produced seven output classes which are happy, sad, angry, fear, surprise, disgust and neutral. The datas to be given is separated and put in two folders for training which are ‘train’ and ‘validation’ folders. The ImageDataGenerator available in Keras is used to read the images in batches from these folders directly and data augmentation is done optionally for this. The batch size given here is 32. The learning rate is indicated by the programmer to indicate the weights convergence on suitable value. Learning rate of this model is 0.01. After the data setup is done the training is done and the model is saved. We have achieved greater accuracy by using this transfer learning approach.

V. RESULTS

The accuracy of the model will be increased with the increase in the epoch. The proposed model attained an accuracy of 99% on training. The loss factor decreased with the increase in the epochs and reached near zero indicating the efficiency of the model. The bad prediction of a model is given by the loss factor. Here it is near zero, as shown in fig.2, which shows the model efficiency. The proposed model is tested on the real-world images available on internet and images we collected from real-world scenarios.

The classification results are evaluated using the outcomes of true positive, true negative, false positive and false negative. True Positive(TP) gives the outcomes where the model predicted correctly the positive class. Similarly, the outcomes where the model correctly predicted the negative class is termed as True Negative(TN). Incorrect predictions of positive class is given by the False Positive(FP) and incorrect prediction of negative class is given by False Negative(FN).The recognition accuracy is the percentage of total items classified corrected as is given by

$$\text{RecognitionAccuracy}=[(TP+TN)/(TP+TN+FP+FN)]*100$$

The method attained an average accuracy rate of 93.5% on the classification of our test set which contain real-world images having partial occlusion.

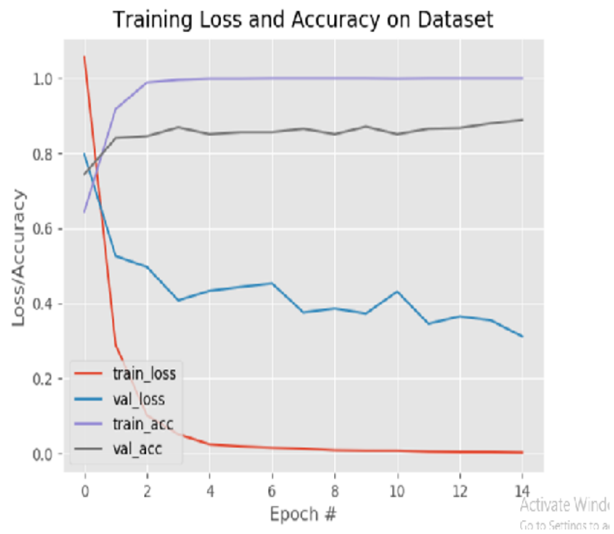

Fig.2. Training accuracy and loss for 15 epochs

Emotion	TP	TN	FP	FN	Recognition rate
Anger	42	15	1	2	95
Disgust	38	18	2	2	93.3
Fear	34	20	2	4	90
Happy	48	10	0	2	96.6
Sad	29	28	2	1	95
Surprise	31	24	2	3	91.6
Neutral	40	16	3	1	93.3
Avg. accuracy					93.5

Table.1. Analysis of our recognition rates per emotion with partially occluded images

VI. CONCLUSION

This paper proposes an effective Facial Expression Recognition model which can handle partial occlusions and pose variation using the CNN architecture MobileNet V2. We focused on classifying the emotion into universally accepted seven emotional categories..We experimented using the techniques like fine-tuning the MobileNet , training it on partially occluded images and finally tested on real-world images. The model achieved an accuracy of 92.5 % on the occluded images. By training the CNN with a huge dataset the accuracy level can be increased.This model can be successfully used on mobile devices and many embedded applications.

REFERENCES

[1] Towner. H. and Slater. M. 2007. *Reconstruction and Recognition of Occluded Facial Expressions Using PCA.* Affective Computing and Intelligent Interaction, 36-47.

[2] Xiaohua Huang, Guoying Zhao, Wenming Zheng and Matti Pietikinen. 2012.*Towards a dynamic expression recognition system under facial occlusion*. Pattern Recognition Letters 33, 16, 2181-2191.

[3] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, Facial expression recognition based on transfer learning from deep convolutional networks, in Natural Computation (ICNC), 2015 11th International Conference on.IEEE, 2015, pp. 702708..

[4] D. C. Ali Mollahosseini and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks.IEEE Winter Conference on Applications of Computer Vision, 2016.

[5] EKMAN, P., FRIESEN, W. 1978. The Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, CA, USA, 274-280.

[6] Rui Li, Pengyu Liu, Kebin Jia, Qiang Wu.2015.Facial Expression Recognition under Partial Occlusion Based on Gabor Filter and Graylevel Co-occurrence Matrix. IEEE 2015 International Conference on Computational Intelligence and Communication Networks.

[7] S.-Y. D. Bo-Kyeong Kim, Jihyeon Roh and S.-Y. Lee. Hi- erarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multi- modal User Interfaces, pages 1–17, 2015.

[8] C. W. Pablo Barros and S. Wermter. Emotional expression recognition with a cross-channel convolutional neural net- work for human-robot interaction.IEEE 15th International Conference on Humanoid Robots, 2015.

[9] Alex Krizhevsky et al. “ImageNet Classification with Deep Convolutional Neural Networks”, Neural Information Processing Systems (NIPS), 2012.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 2015.

[11] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.

[12] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[13] Mark Sandler Andrew Howard Menglong Zhu Andrey Zhmoginov Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381v4 [cs.CV] 21 Mar 2019.