Facial Recognition Based Emotion Analysis

Mr. Saurav Kalaskar
Dept. of Computer Engineering VESIT
Mumbai, India
2018.saurav.kalaskar@ves.ac.in

Mr. Mohit peshwani
Dept. of Computer Engineering VESIT
Mumbai, India
2019mohit.peshwani@ves.ac.in

Mr. Abhishek Odrani
Dept. of Computer Engineering VESIT
Mumbai, India
2018.abhishek.odrani@ves.ac.in

Mrs. Anjali Yeole
Assoc. Prof., Dept. of Computer Engineering VESIT
Mumbai, India
anjali.yeole@ves.ac.in

***Abstract*—Human emotions are mental states of feelings that arise spontaneously rather than through conscious effort. Some of the critical emotions are happy, sad, anger. Facial expressions play a key role in non-verbalcommunication which appears due to internal feelings of a person that reflects on a person's face. In this paper, we are trying to predict human emotions using the Convolution Neural Network (CNN). In this algorithm, the FER-2013 dataset has been applied for training. We recognized emotions such as Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.**

***Keywords*—Facial Expression Recognition (FER), Facial Landmarks (FL), Facial Action Units (AU), Facial Action Coding System (FACS).**

I. INTRODUCTION

Facial expressions play a key role for understanding and detecting emotion. Studies have shown that reading of facial expressions can significantly alter the interpretation of what is spoken as well as control the flow of a conversation. The ability for humans to interpret emotions is very important to effective communication.

During the past decades, various methods have been proposed for emotion recognition. Many algorithms were suggested to develop systems/applications that can detect emotions very well. The basic emotions can be recognized from human facial expressions. This means that regardless of language and cultural barriers, there will always be a set of fundamental facial expressions that people communicate with.

Facial expression recognition, as it extracts and analyzes the information taken from the images or videos, it will be able to deliver the exact or unbiased emotional responses as data. We achieve facial recognition by detecting the faces and by

analyzing the movement of our eyes, nose, lips etc. and analyzing changes in the appearance of the facial features and classifying various expressions.

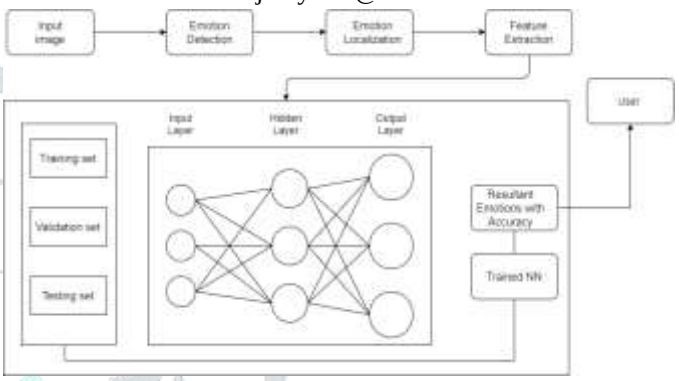


Fig 1. System Architecture

The current approaches primarily focus on facial investigation, keeping background intact and hence building up a lot of unnecessary and misleading features that confuse the CNN training process. Reported techniques on facial expression detection can be described as two major approaches. The first one is distinguishing expressions that are identified with an explicit classifier, and the second one is making characterization dependent on the extracted facial highlights. In the facial action coding system (FACS), action units are used as expression markers. These AUs were discriminable by facial muscle changes.

II. LITERATURE SURVEY

Facial Recognition has become an increasingly researched topic in recent years, mainly because it has a lot of applications in the fields of Computer Vision, robotics, and Human Computer Interaction. Paul Ekman, 1994 has presented seven universal expressions. He has described the positioning of faces, and the muscular movements required to create these expressions in his study (Ekman, 1997). This study has proved to be very useful in the research. It is also used for classifying human expressions. However, recently there has been a trend to implement FER using classification algorithms such as neural networks. There are several datasets available for research in the field of Facial Expression Recognition. The type and number of images, the method of labelling the images varies in each dataset. The CK+ dataset uses the system for labelling faces and contains the Action Units (AU's) for each facial image. There are several challenges for implementing the FER

system. Most datasets consist of images of posed people with a certain expression. This is the first challenge, as real time applications require a model with expressions which are not posed or directed. The second challenge is that the labels in the datasets are broadly classified, which means that in real time there might be some expressions which the system might be able to classify correctly. These systems have become very popular in applications where FER is required.

III. METHODOLOGY

In this paper, deep learning with the convolutional neural network approach is used. The Keras Application programming Interface and OpenCV framework were used. OpenCV is used for the automatic detection of faces and drawing bounding boxes around them. OpenCV consists of many pre-trained classifiers for face, eyes, lips etc. This dataset can be taken from the Kaggle website. All the images of the dataset are of size 48*48. Some images from every category, utility functions are :

Fig 2. FER2013 Database

A. Dataset

The dataset used for implementing the system is the FER2013 dataset from Kaggle on FER (Goodfellow et al, 2013). The dataset consists of 35,887 labelled images, which are divided into 3589 test and 28709 train images. The dataset consists of private test images, on which the final test was conducted. The images in the FER2013 dataset have size 48x48 and are black and white images. The FER2013 dataset contains images that vary in viewpoint, lighting, and scale. Some sample images from the FER2013 dataset.

Fig 3. Dataset sample

B. Process of Facial Recognition

The process of FER has three stages. The preprocessing stage consists of preparing the dataset into a form which will work on a generalized algorithm and generate efficient results. In the face detection stage, the face is detected from the images that are captured real time. The emotion classification step consists of implementing the CNN algorithm to classify the input image into one of seven classes. These stages are shown in the following figure.

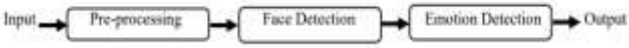


Fig 4. Stages of Facial Recognition

Pre-Processing

The input image to the FER may contain noise and have variation in illumination, size, and color. To get accurate and faster results on the algorithm, some preprocessing operations are done on the image. The preprocessing strategies used are conversion of image to grayscale, normalization, and resizing of image.

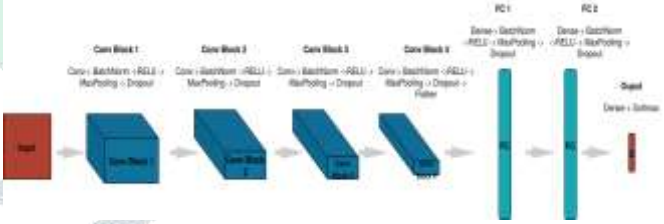


Fig 5. CNN Model

Normalization - Normalization of an image is done to remove illumination variations and obtain improved face image. Grayscaling - It is the process of converting a colored image input into an image whose pixel value depends on the intensity of light on the image.

Grayscaling is done as colored images are difficult to process by an algorithm. Resizing - The image is resized to remove the unnecessary parts of the image. This reduces the memory required and increases computation speed.

Face Detection

Face detection is the primary step for any FER system. The classifiers which detect an object in an image or video for which they have been trained. They are trained over a set of positive and negative facial images. They have proved to be an efficient means of object detection in images and provide high accuracy.

It features three dark regions on the face, for example the eyebrows. The computer is trained to detect two dark regions on the face, and their location is decided using fast pixel calculation. It then removes the unrequired background data from the image and detects the facial region from the image. The face detection process using classifiers is implemented in OpenCV.

Emotion Detection

System classifies the image into one of the seven universal expressions - Happiness, Sadness, Anger, Surprise, Disgust, Fear, and Neutral. The training was done using CNN, which is a category of neural networks proved to be productive in image processing. The dataset was first split into training and test datasets, and then it was trained on the training set. Feature extraction process was not done on the data before feeding it into CNN. The approach followed was to experiment with different architectures on the CNN, to achieve better accuracy with the validation set.

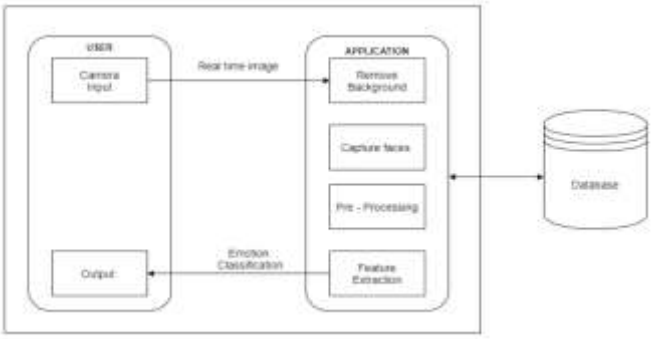


Fig 6. Block Diagram

1) Splitting of Data : The dataset splits into 3 categories according to the "Usage" label in the FER2013 dataset: Training, PublicTest, and PrivateTest. The Training and PublicTest set were used for generation of a model, and the PrivateTest for testing the model.

2) Training and Generation of model : The neural network architecture consists of the following layers: convolution layer, a randomly instantiated learnable filter is slid, or convolved over the input. The operation performs the dot product between the filter and each local region of the input.

Max Pooling, this is used to reduce the spatial size of the input layer and the computational cost. Fully connected layer, each network from the previous layer is connected to the output neurons. The size of the final output layer is equal to the number of classes in which the input image is to be classified.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 64)	640
batch_normalization (BatchNo	(None, 48, 48, 64)	256
activation (Activation)	(None, 48, 48, 64)	0
max_pooling2d (MaxPooling2D)	(None, 24, 24, 64)	0
dropout (Dropout)	(None, 24, 24, 64)	0
conv2d_1 (Conv2D)	(None, 24, 24, 128)	204928
batch_normalization_1 (Batch	(None, 24, 24, 128)	512
activation_1 (Activation)	(None, 24, 24, 128)	0
max_pooling2d_1 (MaxPooling2	(None, 12, 12, 128)	0
dropout_1 (Dropout)	(None, 12, 12, 128)	0
conv2d_2 (Conv2D)	(None, 12, 12, 512)	590336
batch_normalization_2 (Batch	(None, 12, 12, 512)	2048
activation_2 (Activation)	(None, 12, 12, 512)	0
max_pooling2d_2 (MaxPooling2	(None, 6, 6, 512)	0
dropout_2 (Dropout)	(None, 6, 6, 512)	0
conv2d_3 (Conv2D)	(None, 6, 6, 512)	2359808
batch_normalization_3 (Batch	(None, 6, 6, 512)	2048
activation_3 (Activation)	(None, 6, 6, 512)	0
max_pooling2d_3 (MaxPooling2	(None, 3, 3, 512)	0
dropout_3 (Dropout)	(None, 3, 3, 512)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 256)	1179904
batch_normalization_4 (Batch	(None, 256)	1024
activation_4 (Activation)	(None, 256)	0
dropout_4 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 512)	131584
batch_normalization_5 (Batch	(None, 512)	2048
activation_5 (Activation)	(None, 512)	0
dropout_5 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 7)	3591

Total params: 4,478,727
Trainable params: 4,474,759
Non-trainable params: 3,968

Table 1 - CNN Model

Activation functions are used to reduce the overfitting of the image. The advantage of the activation function is that its gradient is always equal to 1, which means that most of the error is passed back during propagation f(x) = max (0, x).

Softmax takes a vector of N real numbers and normalizes that vector into a range of values between (0, 1). Batch Normalization speeds up the training process and applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1.

3) Evaluation of model : The model generated during the training phase was then evaluated on the validation set, which consisted of 3589 images.

4) Using the model to classify real time images : The concept of transfer learning can be used to detect emotion in images captured in real time. The model generated during the training process consists of pretrained weights and values, which can be

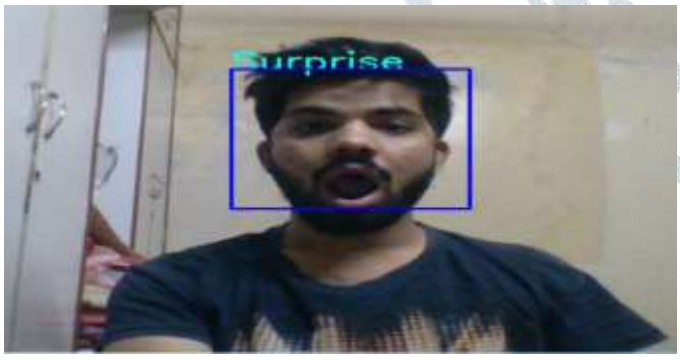
used for implementation of a new facial expression detection problem. As the model generated already contains weights, FER becomes faster for real time images. The CNN architecture is shown in the following figure.

IV. RESULTS AND ANALYSIS

Results were obtained by experimenting with the CNN algorithm. It was observed that the loss over training and test set decreased with each epoch. The batch size was 256, which was kept constant over all experiments.

It was observed that the accuracy of the model increased with an increasing number of epochs. However, a high number of epochs resulted in overfitting the image. It was concluded that fifteen epochs resulted in minimum overfitting and high accuracy. A total of six convolution layers were built, using the activation function.

The neural network accuracy on the dataset varied on the number of filters applied to the image. The number of filters for the first two layers of the network was 64, and it was kept 128 for the third layers of the network Validation of the network is done using different real facial expressions in which the faces are detected and bounded by a box and the facial expressions are named according to the emotion groups.



Loss and accuracy over time

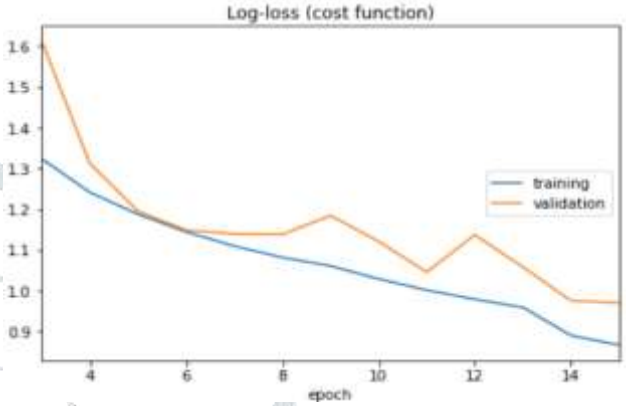


Fig 7. Graph of loss per epoch

It is observed that the loss decreases, and the accuracy increases with each epoch. The training versus testing curve for accuracy remains ideal over the first five epochs, after which it begins to deviate from the ideal values. The training and test accuracy along with the training and validation loss was obtained for the FER2013 dataset using CNN.

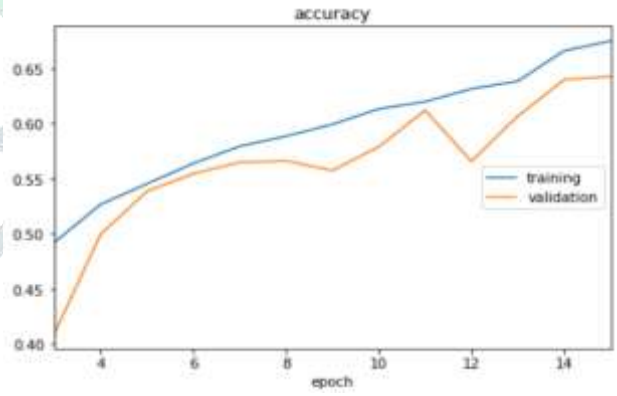


Fig 8. Graph of accuracy per epoch

Accuracy

The final, state-of-the-art model gave a training accuracy of 67.5% and a test accuracy 64.3% as shown in the table. The architecture used could correctly classify 22936 out of 28709 images from the train set and 2158 out of 3589 images from the test set. The results of some of the experiments conducted on CNN are :

```
Log-loss (cost function):
training   (min:    0.866, max:    1.786, cur:    0.866)
validation (min:    0.970, max:    1.705, cur:    0.970)

accuracy:
training   (min:    0.313, max:    0.675, cur:    0.675)
validation (min:    0.381, max:    0.643, cur:    0.643)

Epoch 00015: saving model to model_weights.h5
448/448 [==============================] - 27s 60ms/step
CPU times: user 6min 50s, sys: 57.4 s, total: 7min 47s
Wall time: 6min 46s
```

Fig 9. Accuracy of the model obtained

V. APPLICATIONS

Emotion recognition has applications in different sectors, for example, retail and education. Based on our current customer and partner engagements, one of the most promising use cases is marketing/advertising. Our customers want to know how people respond to ads, products, packaging, and product design. Education applications measure real-time learner responses and engagement with the educator. For medical purposes, doctors can use the system to understand the intensity of pain or illness of the patients.

VI. FUTURE SCOPE

Future work should entail improving the robustness of the classifiers by adding more training images from different datasets, investigating more accurate detection methods that still maintain computational efficiency, and considering the classification of more nuanced and sophisticated expressions. We can also test the dataset with different angles, lighting and background..

VI. CONCLUSION

In this paper, an approach for FER using CNN has been discussed. Facial Expression Recognition (FER) has attracted increasing attention in recent years. The past decade has witnessed the development of many new FER algorithms. We divided the conventional methods into three major steps: pre-processing, face detection and emotion detection. We analyzed emotions according to their facial expressions from the perspective of computer simulation. With reference to the FER2013 dataset, the emotions were classified into 7 emotion groups. The background removal adds a great advantage in accurately determining the emotions. We achieved a test accuracy of 0.643 and a validation accuracy of 0.675.

VII. REFERENCES

https://www.kaggle.com/deadskull7/fer2013

http://cs231n.stanford.edu/reports/2016/pdfs/005_Report.pdf

https://web.stanford.edu/class/ee368/Project_Autumn_1617/Reports/report_pao.pdf

https://cs224d.stanford.edu/reports/AhresY.pdf

https://link.springer.com/article/10.1007/s42452-020-2234-1

https://towardsdatascience.com/emotion-recognition-4fba48dabb6e

F. Ahmed, H. Bari, and E. Hossain. Person-independent facial expression recognition based on compound local binary pattern (clbp). Int.Arab J. Inf. Technol., 11(2):195–203, 2014.

F. Bashar, A. Khan, F. Ahmed, and M. H. Kabir. Robust facial expression recognition based on median ternary pattern (mtp). InElectrical Information and Communication Technology (EICT), 2013International Conference on, pages 1–5. IEEE, 2014.

P. Carcagn, M. Coco, M. Leo, and C. Distante. Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus, 4(1):1, 2015.

P. Ekman. An argument for basic emotions. Cognition & emotion,6(3-4):169–200, 1992.

Ekman, R. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In the International Conference on Neural Information Processing (pp. 117-124). Springer, Berlin, Heidelberg.

Mehrabian A (2017) Nonverbal communication. Routledge, London

Bartlett M, Littlewort G, Vural E, Lee K, Cetin M, Ercil A, Movellan J (2008) Data mining spontaneous facial behavior with automatic expression coding. In: Esposito A, Bourbakis NG, Avouris N, Hatzilu Geroudis I (eds) Verbal and nonverbal features of human–human and human–machine interaction. Springer, Berlin, pp 1–20

Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. Psychol Bull 115(1):102

Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998).