REAL TIME FACIAL EXPRESSION RECOGNITION

[1]Mohit Jindal, [2]Harshit, [3]Himanshu, [4]Ms. Princy Jain

1,2,3 B.tech Students, 4 Assistant Professor

Department of Information Technology

Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi

Abstract

As we move towards a digital world, Human Computer Interaction becomes very important. A lot of research has been done in this field over the past decade. Face expressions are a key feature of non-verbal communication, and they play an important role in Human Computer Interaction. This paper presents an approach of Facial Expression Recognition (FER) using Convolutional Neural Networks (CNN). This model created using CNN can be used to detect facial expressions in real time. The system can be used for analysis of emotions while users watch movie trailers or video lectures. The KDEF dataset is used to solve image recognition problem with different Convolutional Neural Networks. The Neural Network used here has filter size of 64 in first and 32 in second Convolution layer,64 in third and 32 in fourth layer, 64 in fifth and 32 in sixth layer, 128 in seventh and eighth convolution layer and 64 units in two dense layers and 0.4 normalization layer, and the kernel size is consistently being used as (3, 3).

Literature Survey

In last few years, FER has become an increasingly researched topic, mainly because it has a lot of applications in the fields of Computer Vision, robotics, and Human Computer Interaction. In 1994 Paul Ekman, has presented six 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 of FER.

The Facial Action Coding System (FACS), developed by Swedish anatomist Carl-Herman Hjortsjö, is a coding system used to taxonomize human facial movements based on their appearance on the face. This system, which was later adopted by Ekman & Friesen (2003), is also a useful method of classifying human expressions. FER systems were mostly implemented using the FACS in the past. However, recently there has been a trend to implement FER using classification algorithms such as SVM, neural networks, and the Fisherface algorithm

(Alshamsi, Kepuska & Meng, 2017; Fathallah, Abdi & Douik, 2017; Lyons, Budynek & Akamatsu, 1999).There are several datasets available for research in the field of Facial Expression Recognition, such as the Japanese Female Facial Expressions (JAFFE), Extended Cohn Kanade dataset (CK+), The Karolinska Directed Emotional Faces (KDEF) and the FER2013 dataset. The type and number of images, the method of labelling the images varies in each dataset. The CK+ dataset uses the FACS system for labelling faces and contains the Action Units (AU's) for each facial image.

There are several challenges with 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.

There are many FER systems, such as Affectiva, and Microsoft's Emotion API (McDuff et al., 2016; Linn, 2015). These systems have become very popular in applications where FER is required.

Motivation

Within increase of technology, it is observed that Computer and Human Interaction has become important. As a result, FER is heavily researched by machines has become the most searched field by experts over the last decade. For classify human expressions in real time, there is a need for an application. for understanding the human mind in the field of psychology, or to help machines to understand user requirements, the classification of emotions is used.

Proposed System

This paper intends to elaborate a method to develop a FER system using CNN. The system will classify the expression of a human face into one of seven expressions - anger, happiness,

sadness, surprise, fear, neutral, disgust. to categorize human faces in real time, The model is used using a webcam. This FER system can be used for analysis of user expressions, to help the system understand human requirements better.

Methodology

All subjects acquired written information in progress. This information entailed a summary of the seven variant expressions that they were to pose during the photo session. The subject was asked to rehearse the different expressions for 1 hour before coming to the photo session. It was emphasized that the subject should try to evoke the emotion that was to be expressed, and - while maintaining a way of expressing the emotion that felt natural to them - try to make the expression strong and clear.

They were seated at a distance of approximately three meters from the camera. The absolute distance was adapted for each subject by adjusting the camera position until the subject's eyes and mouth were at specific, pre-defined vertical and horizontal positions on the camera's grid screen. The lights were set to cast a soft indirect light evenly distributed at both sides of the face. After a session of rehearsal, the subjects were shot in one expression at the time until all seven expressions had been shot (series one). The subjects were the shot once again in all expressions and angles (series two).

Process of Facial Expression 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 input image into one of seven classes. These stages are described using in a flowchart:

Input----- Pre-processing----- Face Detection----- Emotion Detection----- Output

A. Preprocessing

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 were done on the image. The preprocessing strategies used are conversion of image to grayscale, normalization, and resizing of image.

1) Normalization - Normalization of an image is done to remove illumination variations and obtain improved face image.

2) Grayscaling - Grayscaling 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.

3) Resizing - The image is resized to remove the unnecessary parts of the image. This reduces the memory required and increases computation speed.

B. Face Detection

Face detection is the primary step for any FER system. For face detection, Haar cascades were used (Viola & Jones, 2001). The Haar cascades, also known as the Viola Jones detectors, are 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. Haar cascades have proved to be an efficient means of object detection in images and provide high accuracy.

Haar features detect 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. Haar cascades successfully remove the unrequired background data from the image and detect the facial region from the image.

The face detection process using the Haar cascade classifiers was implemented in OpenCV. This

method was originally proposed by Papageorgiou et al, using rectangular features which are

shown in figure 3 (Mohan, Papageorgiou & Poggio, 2001; Papageorgiou, Oren & Poggio, 1998).

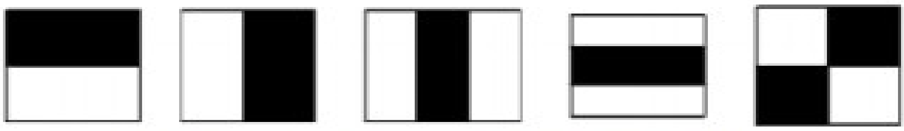


Fig.3. Haar features (Shan, Guo, You, Lu, & Bie, 2017)

C. Emotion Classification

In this step, the system classifies the image into one of the seven universal expressions -

Happiness, Sadness, Anger, Surprise, Disgust, Fear, and Neutral as labelled in the KDEF dataset. The training was done using CNN, which are 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, with minimum overfitting. The emotion classification step consists of the following phases:

1) Splitting of Data

The dataset was split into 3 categories according to the "Usage" label in the KDEF dataset:

Training, Public Test, and Private Test. The Training and Public Test set were used for generation of a model, and the Private Test set was used for evaluating the model.

2) Training and Generation of model

The neural network architecture consists of the following layers:

i. Convolution Layer

In the 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. The output is a 3D volume of multiple filters, also called the feature map.

ii. Max Pooling

The pooling layer is used to reduce the spatial size of the input layer to lower the size of input and the computation cost.

iii. Fully connected layer

In the fully connected layer, each neuron from the previous layer is connected to the output

neurons. The size of final output layer is equal to the number of classes in which the input image is to be classified.

iv. Activation function

Activation functions are used in to reduce the overfitting. In the CNN architecture, the swish activation function has been used. Like ReLU, Swish is unbounded above and bounded below. Swish is smooth and nonmonotonic. In fact, the non-monotonicity property of Swish makes it different from most common activation functions. [15] [16].

$f(x) = x * \text{sigmoid}(x)$ Equation 1: Equation of Swish Activation Function

v. SoftMax

The SoftMax function takes a vector of N real numbers and normalizes that vector into a range of values between (0, 1).

vi. Batch Normalization

The batch normalizer 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 70 individual videos with 7 emotion images.

4) Using 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 Fig:

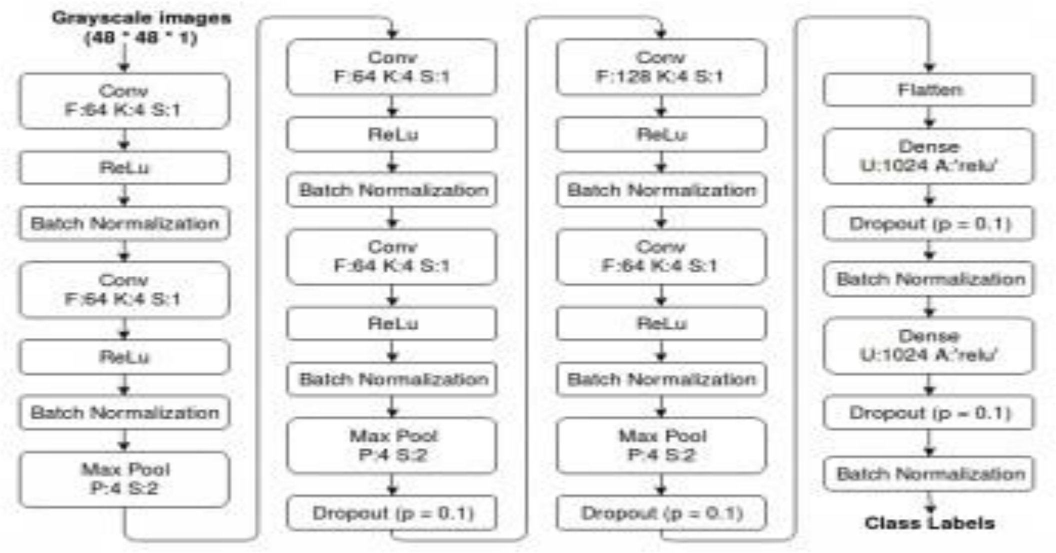


Fig.4. CNN architecture

Experiments and Results

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.

The following changes were made in the neural network architecture to achieve good results:

1) Number of epochs:

It was observed that the accuracy of the model increased with increasing number of epochs. However, a high number of epochs resulted in overfitting. It was concluded that twenty five epochs resulted in minimum overfitting and high accuracy.

2) Number of layers:

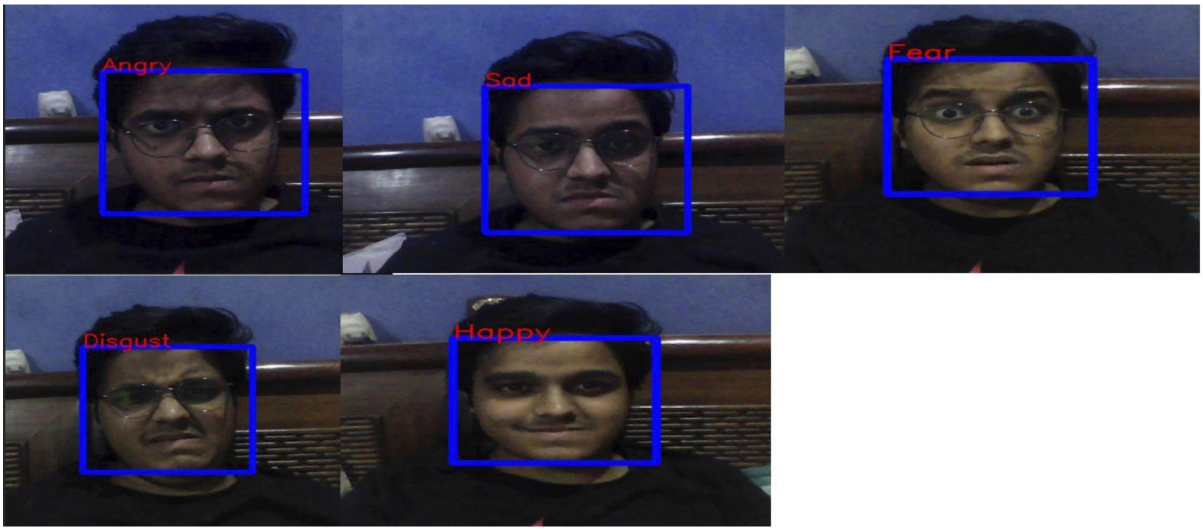
The neural network architecture consists of layers as follow: seven convolution layers, three batch normalization layers, four maxpooling layers, five dropout layers and three dense layers. The initially used 'sigmoid' function was replaced by 'swish' activation function.

3) Filters:

The neural network accuracy on the dataset varied on the number of filters applied to the image. The number of filters in the convolutional layers used were 32 and 64 used alternatively and 128 in sixth and seventh convolutional layer.

4) Data Augmentation:

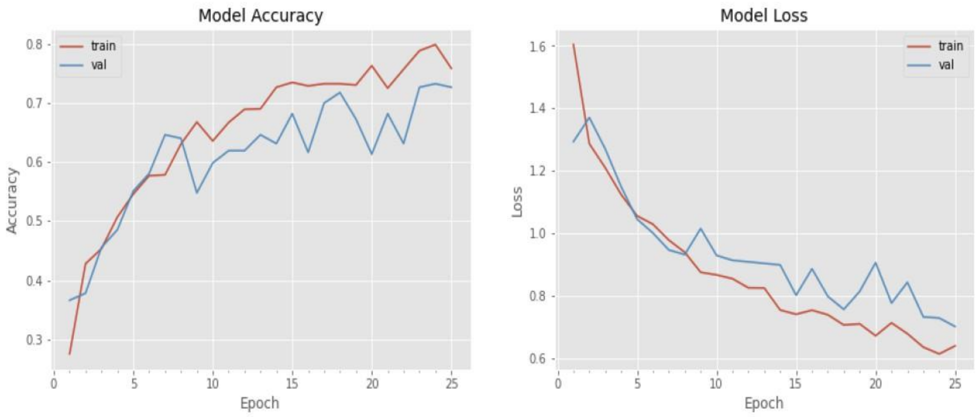
The data was augmented, viz., rotated, shifted, horizontally flipped, vertically flipped etc., which combined with swish activation function resulted into drastic changes into the accuracy.



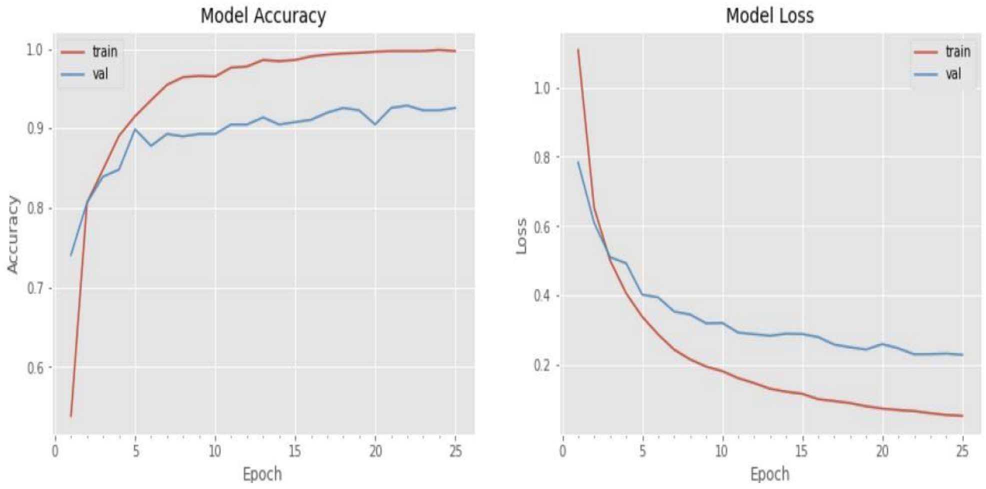
Results pictures

Accuracy

The Neural Network used here has 64 layers in first and 32 in second Convolutional layer, 64 in third and 32 in fourth layer, 64 in fifth and 32 in sixth layer, 128 in seventh and eighth convolution layer and 64 units in two dense layers and 0.4 normalization layer. The output/result obtained by this model is about 72% when used sigmoid activation function without Data Augmentation. This means there is a probability of .8 that the image will be recognized correctly. To make results better we use augmentation on the data and swish activation function. By augmentation we mean, creation The Neural Network used here has 64 layers in first and 32 in second Convolution layer, 64 in third and 32 in fourth layer, 64 in fifth and 32 in sixth layer, 128 in seventh and eighth convolution of new data with different orientations. For instance, flipping an image, rotation of image etc. After augmentation and using swish we see a jump of about 20% in our test accuracy making it around 92%.



WITHOUT AUGMENTATION AND SWISH



WITHOUT AUGMENTATION AND SWISH

Conclusion

In this paper, an approach for FER using CNN has been discussed. A CNN model on the KDEF dataset was created and experiments with the architecture were conducted to achieve a test accuracy of 0.7262 and a validation accuracy of 0.9256. This state-of-the-art model has been used for classifying emotions of users in real time using a webcam. The webcam captures a sequence of images and uses the model to classify emotions.

References

[1] Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique.

[2] 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

[3] Ekman, P., & Friesen, W. V. (2003). Unmasking the face: A guide to recognizing emotions from facial clues. Ishk. Alshamsi, H., Kepuska, V., & Meng, H. (2017, October). Real time automated facial expression recognition app development on smart phone.

[4] DEEP FACIAL EXPRESSION RECOGNITION : A SURVEY BY SHAN LI AND WEIHONG DENG

[5] International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580) (pp. 46-53). IEEE.

[6] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (pp. 94-101). IEEE.

[7] Fathallah, A., Abdi, L., & Douik, A. (2017, October). Facial expression recognition via deep learning. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) (pp. 745-750). IEEE.

[8] Lyons, M. J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. IEEE transactions on pattern analysis and machine intelligence, 21(12), 1357-1362.

[9] Hahnloser, R. H., & Seung, H. S. (2001). Permitted and forbidden sets in symmetric

threshold-linear networks. In Advances in Neural Information Processing Systems (pp. 217-223).

[10] Kumar, G. R., Kumar, R. K., & Sanyal, G. (2017, July). Facial emotion analysis using deep convolution neural network. In 2017 International Conference on Signal Processing and Communication (ICSPC) (pp. 369-374). IEEE.

[11] Shan, K., Guo, J., You, W., Lu, D., & Bie, R. (2017, June). Automatic facial expression recognition based on a deep convolutional-neural-network structure. In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 123-128). IEEE.

[12] Kulkarni, K. R., & Bagal, S. B. (2015, December). Facial expression recognition. In 2015 Annual IEEE India Conference (INDICON) (pp. 1-5). IEEE.