Karishma Sunvaiya. 2020, 8:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Real Time Facial Expression Recognition using Deep Learning

M. Tech. Student Karishma Sunvaiya, Assistant Professor Megha Jat
Department of Computer Science
Patel College of Science & Technology, Indore, India
ksunvaiya@gmail.com, Gmegha424@gmail.com

***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.**

***Keywords*- Facial Expression Recognition, Convolutional Neural Networks, Deep Learning, Transfer Learning.**

I. INTRODUCTION

1.1. Motivation

With an increase in the use of technology, it is observed that Human Computer Interaction has become important. As a result, FER by machines has become a heavily researched field by experts over the last decade. There is a need for an application which will be able to detect and classify human expressions in real time. This classification of emotions can then be used for understanding the human mind in the field of psychology, or to help machines to understand user requirements.

1.2. 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. The model thus developed can be used to categorize human faces in real time using a webcam. This FER system can be used for analysis of user expressions, to help the system understand human requirements better.

FER 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 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+), and the FER2013 dataset (Canade, Cohn & Tian, 2000; Lucey et al., 2010; Goodfellow et al., 2013). 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.

III. METHODOLOGY

1. Dataset: The dataset used for implementing the FER system was the FER2013 dataset from the Kaggle challenge 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 another 3589 private test images, on which the final test was conducted during the challenge. 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. Fig. 1 shows some sample images from the FER2013 dataset, and Table 1 illustrates the description of the dataset.

Fig. 1. Sample images from the FER2013 dataset

Table 1. Description of the FER2013 dataset

Label	Number of images	Emotion
0	4593	Angry
1	547	Disgust
2	5121	Fear
3	8989	Happy
4	6077	Sad
5	4002	Surprise
6	6198	Neutral

2. 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 in Fig. 2:

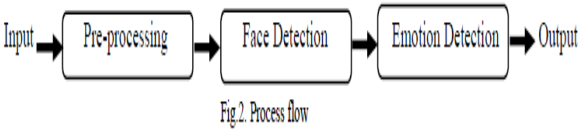

Fig.2. Process flow

2.1. 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. Gray scaling – Gray scaling 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.

2.2. 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).

3.3. 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 FER2013 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:

3.3.1. Splitting of Data

The dataset was split 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 set was used for evaluating the model.

3.3.2. Training and Generation of model

The neural network architecture consists of the following layers:

3.3.2.1. 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.

3.3.2.2. 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.

3.3.2.3. 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.

3.3.2.4. Activation function

Activation functions are used in to reduce the overfitting. In the CNN architecture, the ReLu activation function has been used. The advantage of the ReLu activation function is that its gradient is always equal to 1, which means that most of the error is passed back during back-propagation [15] [16].

$$f(x) = \max(0, x)$$

Equation 1: Equation of ReLu Activation Function

3.3.2.5. Softmax

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

3.3.2.6. 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.

3.3.3. Evaluation of model

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

3.3.4. 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. 4:

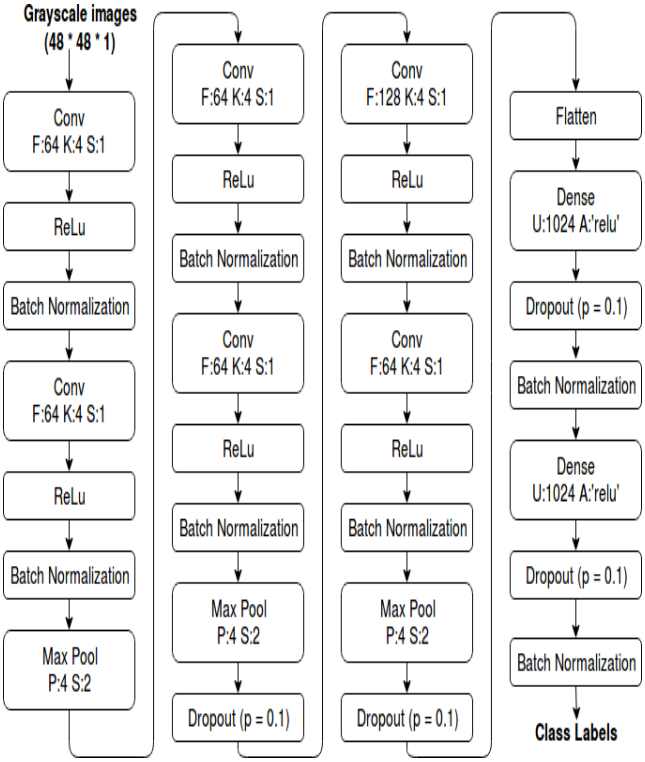

Fig.4. CNN architecture.

IV. 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 eight epochs resulted in minimum overfitting and high accuracy.

2. Number of layers: The neural network architecture consists of three hidden layers and a single fully connected layer. A total of six convolution layers, were built, using 'relu' as the 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 for the first two layers of the network was 64, and it was kept 128 for the third layers of the network.

1. Accuracy

The final, state-of-the-art-model gave a training accuracy of 79.89% and a test accuracy 60.12% 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. Table. 2 shows the results of some of the experiments conducted on CNN.

Table 2. Accuracy obtained over the three experiments

Experiment	Training Accuracy	Test Accuracy	Validation Accuracy
Experiment 1	63.22	56.56	89.01
Experiment 2	68.37	58.03	89.61
Experiment 3	79.89	60.12	89.78

Table 3 shows a brief comparison of the proposed system with other related works:

Related work	Algorithm	Dataset	Results
Kumar, Kumar, & Sanyal, 2016	CNN	FERC-2013	Around 90%
Amin, Chase & Sinha, 2017	CNN	FER-2013	60.37
Shan, Guo, You, Lu & Bie, 2017	KNN	JAFFE, CK+	65.11, 77.27
Kulkrni, Bagal, 2015	Gabor, Log Gabor	FACES	82%, 87%
Minaee, & Abdolrashidi, 2019	Attentional CNN	FER2013	70.02%
Proposed	CNN	FER2013	89.78%

2. Loss and accuracy over time

It can be 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 obtained for the FER2013 dataset using CNN are given in Table.3

Table 4. Accuracy per epoch.

Epoch	Training Accuracy	Validation Accuracy
1	29.10	43.33
2	47.81	50.65
3	55.60	56.90
4	60.13	57.65
5	64.07	57.95
6	67.00	59.63
7	69.95	59.01
8	72.88	60.13

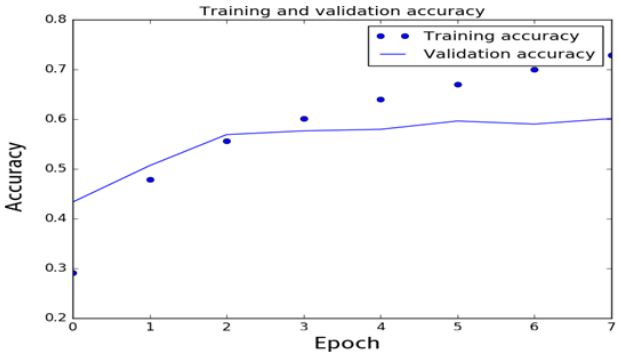

Fig.5. Graph of training and validation accuracy per epoch

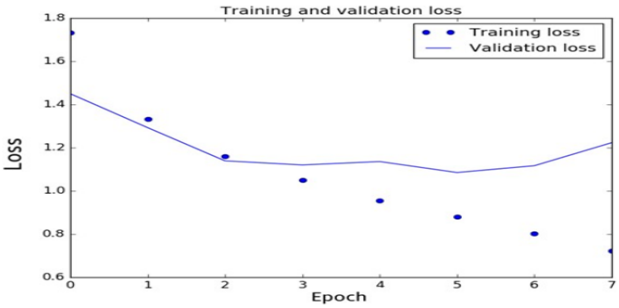

Fig.6. Graph of training and validation loss per epoch

3. Confusion Matrix

The confusion matrix generated over the test data is shown in figure 7. The dark blocks along the diagonal show that the test data has been classified well. It can be observed that the number of correct classifications is low for disgust, followed by fear. The numbers on either side of the diagonal represent the number of wrongly classified images. As these numbers are lower compared to the numbers on the diagonal, it can be concluded that the algorithm has worked correctly and achieved state of the art results.

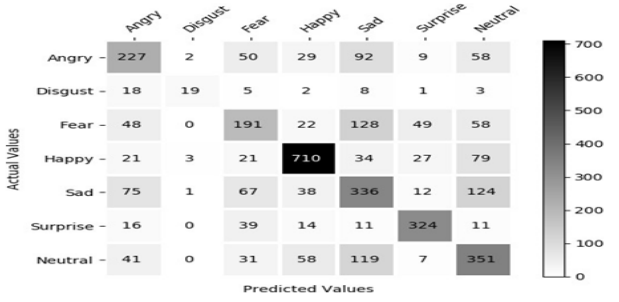

Fig.7. Confusion matrix represented as a heatmap.

V. CONCLUSION

In this paper, an approach for FER using CNN has been discussed. A CNN model on the FER2013 dataset was created and experiments with the architecture were conducted to achieve a test accuracy of 0.6012 and a validation accuracy of 0.8978. 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.

[4] Alshamsi, H., Kepuska, V., & Meng, H. (2017, October). Real time automated facial expression recognition app development on smart phones. In 2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 384-392). IEEE.

[5] 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.

[6] 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.

[7] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580) (pp. 46-53). IEEE.

[8] 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.

[9] 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 International Conference on Neural Information

Processing (pp. 117-124). Springer, Berlin, Heidelberg.

[10] P. Shakyawar, P. Choure, and U. Singh, "Eigenface method through through facial expression recognition," in Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017, 2017, vol. 2017-Janua, doi: 10.1109/ICECA.2017.8212714.

[11] K. Kushwah, V. Sharma, and U. Singh, "Neural network method through facial expression recognition," in Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017, 2017, vol. 2017-Janua, doi: 10.1109/ICECA.2017.8212721.

[12] Y. Mathur, P. Jain, and U. Singh, "Foremost section study and kernel support vector machine through brain images classifier," in Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017, 2017, vol. 2017-Janua, doi: 10.1109/ICECA.2017.8212726.

[13] V. S. Tomar, N. Gupta, and U. Singh, "Expressions recognition based on human face," in Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, 2019, doi: 10.1109/ICCMC.2019.8819714.

[14] Papageorgiou, C. P., Oren, M., & Poggio, T. (1998, January). A general framework for object detection. In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271) (pp. 555-562). IEEE.

[15] Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature, 405(6789), 947.

[16] Patil M.N., Brijesh Iyer, Rajeev Arya (2016) Performance Evaluation of PCA and ICA Algorithm for Facial Expression Recognition Application. In: Pant M., Deep K., Bansal J., Nagar A., Das K. (eds) Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol 436. Springer, Singapore

[17] 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).

[18] 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.

[19] Amin, D., Chase, P., & Sinha, K. (2017). Touchy Feely: An Emotion Recognition Challenge. Palo alto: Stanford.

[20] 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.

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

[22] Minaee, S., & Abdolrashidi, A. (2019). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. arXiv preprint arXiv:1902.01019.