Real Time Emotion Recognition Using Convolutional Neural Network

Bah Abdoulaye

MSc in Computer Engineering,
Istanbul Sabahattin Zaim University

***Abstract*—Due to the evolution of technology, Human Computer Interaction has become a crucial field of interest. Over the past decade various research has been accomplished in this field and still are ongoing. Facial expressions play an important role in non-verbal communication and also in Human Computer Interaction. Emotions on a human face say a lot about our thought process and give a hint on what is going on inside our brain. In this Paper a method of Facial Emotion Recognition is implemented (FER) using Convolutional Neural Networks (CNN) which can be used for the classification of facial emotion expressions in real time. This work and experiment can be used for emotion analysis while people watch movie trailers or video lectures or other purposes.**

***Index Terms*—Convolutional Neural Network, Transfer Learning, Facial Expression Recognition, Neural Network**

I. Introduction

Human Computer Interaction has gained consideration due to the high usage rate of technology. Therefore, FER using machines has become a trend topic of interest for experts and researchers over the last decade.

Furthermore, an application or device able to detect and classify human expressions in real time is required. This emotion classification can be used later in psychology to understand the human mind or also help devices to have a better idea about the user needs or requirements.

FER is not limited to that, it can be used in association with other systems in order to furnish safety. For instance, ATMs could be set up such that they won't dispense money when the user is scared. In the gaming industry, emotion-aware games can be developed which could vary the difficulty of a level depending on the player's emotions. By judging their expressions during different points of the game, a general understanding of the game's Software for cameras can use emotion recognition to take photos whenever a user smiles[11].

This paper aims to discuss about a method developed to implement a FER system using CNN. The system will be able to do the classification of seven human face expressions such as happiness, neutral, fear, disgust, anger, surprise and sadness.

Furthermore, with the help of a webcam, the so-called model will be used for the categorization or classification of human faces in real time. This experiment can then be used later to analyze user expressions in order to make easier the understanding of user requirements and needs.

The rest of this paper is as follows: in Section 2, there is an introduction about the Related Works. Section 3 about the Methodology. In Section 4 we will show our experiment result, and the experiment settings. Then we finish with our conclusion in Section 5 and Future Works in Section 6.

II. Related Works

In the last previous years FER has been and still a trend topic for researchers because of its various usage in robotic, Computer Vision, and especially Human Computer Interaction[1][2][3]. 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). In order to promote researchers to improve the FER2013 which was designed by Goodfellow et al. Kaggle organized a competition. Using CNN with some image transformation was the technique which led the top three teams to success [7]. The winner, Yichuan Tang, achieved a 71.2% accuracy by using the primal objective of an SVM as the loss function for training and additionally used the L2-SVM loss function [8].

A paper presented by Pramerdorfer and Kampel [9] describes the approaches taken by six current state-of-the art papers and ensembles their network to achieve 75.2% test accuracy on FER2013, which is, to our knowledge, the highest reported in any published journal paper.

Among the six papers, Zhang et al. achieved the highest accuracy of 75.1% by employing auxiliary data and additional features: a vector of HoG features was computed from face patches and processed by the first FC layer of the CNN (early fusion).

They also employed facial landmark registration, suggesting its benefits even in challenging conditions (facial landmark extraction is inaccurate for about 15% of images in the FER dataset) [4]. The paper with the second highest accuracy by Kim et al. utilized face registration, data augmentation, additional features, and ensembling [5].

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

A. The Dataset

We used FER2013 dataset from the Kaggle competition (Goodfellow et al., 2013)to implement the FER system. The images in this dataset are black and white images, having 48x48 size grayscale images. The dataset contains 38,887 images varying in lighting, scale, and viewpoint.

The dataset file is in a csv format containing columns and such as "Label", "Number of images" and "Emotion"[10].Moreover the so-called dataset has 28,710 images as training set, as for the public and final test, 3587, 3590 respectively[10]. We can have a look at some examples of the dataset in Fig. 1, and the explanation of the dataset in Table 1.

An image of each class is shown.

Fig. 1 FER2013 dataset sample images

Table 1. FER2013 dataset description [6]

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

The Convolutional Neural Network was implemented with the help of Keras and Tensorflow. This Network can be improved with the help of a stronger CPU [10].

B. Convolutional Neural Networks

A Convolutional neural network is a neural network comprised of convolution layers which does computational heavy lifting by performing convolution. Convolution is a mathematical operation on two functions to produce a third function. It is to be noted that the image is not represented as pixels, but as numbers representing the pixel value. In terms of what the computer sees, there will simply just be a matrix of numbers. The convolution operation takes place on these numbers. We utilize both fully-connected layers as well as convolutional layers. In a fully-connected layer, every node is connected to every other neuron. They are the layers used in standard feedforward neural networks.

Unlike the fullyconnected layers, convolutional layers are not connected to every neuron. Connections are made across localized regions. A sliding "window" is moved across the image. The size of this window is known as the kernel or the filter. They help recognise patterns in the data. For each filter, there are two main properties to consider - padding and stride. Stride represents the step of the convolution operation, that is, the number of pixels the window moves across. Padding is the addition of null pixels to increase the size of an image. Null pixels here refers to pixels with value of 0. If we have a 5x5 image and a window with a 3x3 filter, a stride of 1 and no padding, the output of the convolutional layer will be a 3x3 image.

This condensation of a feature map is known as pooling. In this case, "max pooling" is utilized. Here, the maximum value is taken from each sliding window and is placed in the output matrix. Convolution is very effective in image recognition and classification compared to a feed-forward neural network. This is because convolution allows to reduce the number of parameters in a network and take advantage of spatial locality. Further, convolutional neural networks introduce the concept of pooling to reduce the number of parameters by downsampling. Applications of Convolutional

neural networks include image recognition, self-driving cars and robotics. CNN is popularly used with videos, 2D images, spectrograms, Synthetic Aperture Radars.[11]

C. Process of FER

The implementation of FER is divided into three steps. In the first step which is the preprocessing, we are preparing the dataset such that it works on a generalized algorithm and also works efficiently. In the second step, which is the face detection, we are detecting face from the images being captured in real time.
And our last step is the emotion classification, we are classifying the input image into our seven classes by the implementation of CNN. A description of the steps is shown in fig. 2.

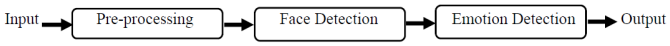

Fig.2 Implementation phases

Pre-processing:
To get rid of the noise, the variation in illumination, the color and size, that the input image may contain in order to get faster and accurate results on the algorithm, some preprocessing implementations were done on the images. The so-called used techniques were first of all Normalization, then Grayscaling, and lastly image resizing.

Normalization was done to remove the illumination variations in order to get better face image. Grayscaling was done to convert the 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[6]. Image resizing was done to remove the useless or unnecessary image parts. As a result it increases the computational speed and also diminish the memory.

Face detection:
Lets start with to the primary step of FER which is the face detection. Haar cascade are kind of classifiers able to detect an object in a video or an image for which they have been trained. They are trained over a set of positive and negative facial images[6].

The reason behind choosing Haar cascade is because of its fame in object detection and also its high accuracy results.

Haar features detect three dark regions on the face [6], the eyebrows for example.

The computer is trained to detect two dark regions on the face and using fast pixel calculation their location is decided. The unnecessary background data is successfully removed by Haar cascade from the image and detect the facial region from the image.

OpenCV was used for the implementation of Haar cascade classifiers for the face detection step. 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)

Emotion Classification:
Here our model will be classifying the image into one of the seven categories – Sadness, Happiness, Anger, Disgust, Surprise, Fear, and Neutral. A category of neural network known as CNN was used for the training because of its high accuracy in image processing. First we split the dataset into training an testing datasets, and then trained on the training set. Feature extraction was done after feeding it into CNN.
The following steps were used for the emotion classification phase:

- *Data Splitting:* According to the usage label in the FER2013 dataset, the data was split into 3 categories such as Training, PublicTest and PrivateTest. For the generation of the model Training and PublicTest were used, and for the model evaluation PrivateTest.

- *Training and Model Generation:* The architecture of our neural network was as follows.

  - *Convolutional layer:* here a learnable filter which is randomly instantiated is convolved or slid over the input. The operation implements the dot product between the filter and each local region of the input. As output we have a 3D volume of multiple filters, which we can call feature map also.

  - *Max Pooling:* To diminish the spatial size of the input layer we use the pooling layer, to reduce the size of the computational cost and input.

  - *Fully Connected Layer:* In the fully connected layer, there is connection between each neuron from the previous layer and 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.

  - *Activation function:* The activation function is used for only one reason which is to reduce the overfitting. The ReLu activation function was used in the CNN, because its gradient is equal to 1 always. That is to say that most of the error is passed back while the implementation of back propagation.

- *Softmax:* Softmax is a function which takes a vector of N real numbers and normalizes it between (0,1) values.

- *Batch Normalization:* To speed the training process we use batch normalization; another benefit of batch normalization is that it applies a transformation that conserve the mean activation close to 0 and the activation standard deviation close to 1.

D. Model evaluation

The obtained model during the Training phase was later evaluated on the validation test, that contains 3589 images.

E. Model usage to classify real time images

By using the concept of transfer learning, we can detect emotion in images captured in real time. The generated model from the training phase is composed of pretrained values and weights, that can be used to implement a new facial expression detection problem. Since the so-called model already contains weights, FER becomes faster for real time images. In fig. 4 the CNN architecture is shown.

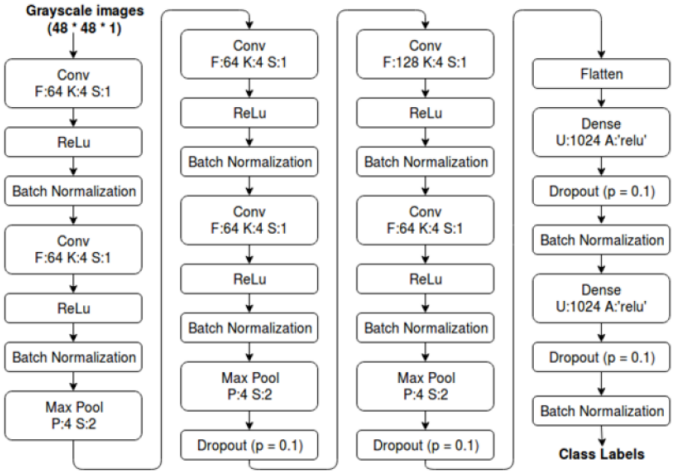

Fig. 4 CNN Architecture [6]

IV. Experimental Results

The results were obtained by the implementation of CNN algorithm. It was remarked that the training and testing set was decreasing after each epoch. The batch size was 256, which did not change during all the experiment.
To get better results the following changes were done in the neural network:

- *Epoch number:* We observed that the accuracy of the model increased when we increased the number of epochs. Furthermore, a high number of epochs causes overfitting. In the end we concluded that eight epochs resulted in minimum overfitting and high accuracy.
- *Layer numbers:* The neural network is composed of three hidden layer and one fully connected layer. Totally six convolutional layers were built using ReLu activation function.
- *Filters:* We observed some changes in the accuracy while we changed the number of filters. For the first two layers the number of filters applied on the network was 64, and for the third layer it was left as 128 constantly.

A. Accuracy:

The final state-of-the-art model gave a test accuracy of 60.12% and a training accuracy of 79.89% as we can see in the table. The used architecture was able to classify correctly 22936 out of 28709 images from the train set and 2158 out of 3589 images from the test set. In table 2 you can see the results of some experiments done on CNN.

Table 2. Accuracy from three experiments[6]

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[6].

Table 3. Comparison with related works [6]

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%

B. Loss and accuracy over time:

We remark that the loss and the accuracy decrease after each epoch. The training versus testing cure for the accuracy remains the same over the first five epochs. The training and test accuracy along with the training and validation loss obtained from our dataset FER2013 using CNN is given in Table.3.

Table 3. Accuracy per epoch [6]

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

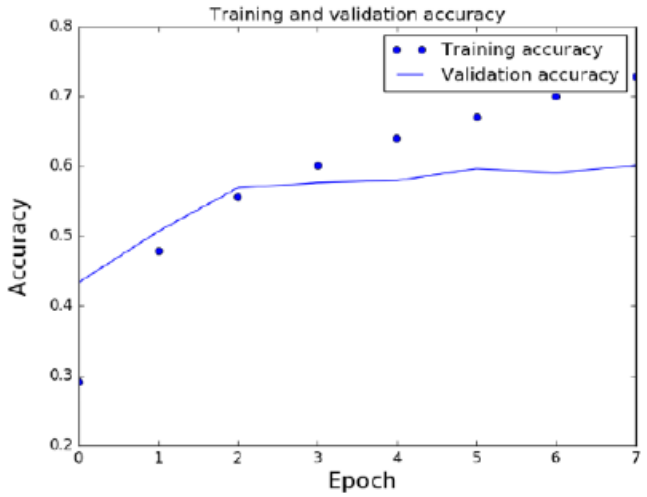

Fig.5. Graph of training and validation accuracy per epoch[6]

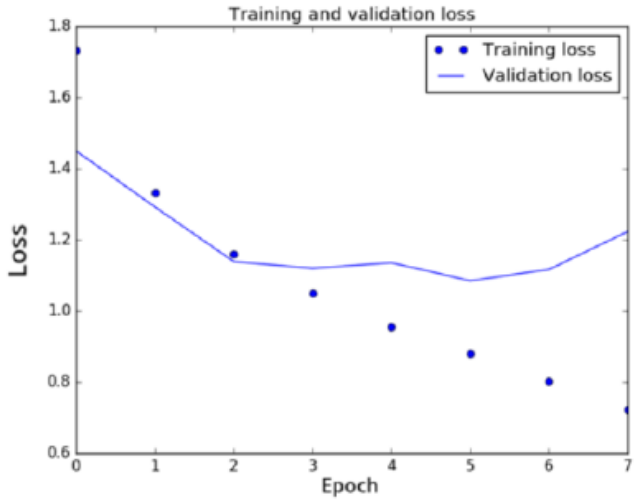

Fig.6. Graph of training and validation loss per epoch[6]

C. Confusion Matrix

The confusion matrix that was created for the test data is discussed in Figure 7. The dark squares around the diagonal indicates that the test data is conducting the classification correctly. Whereas the number of right classifications for disgust and fear is poor, as we observed.

The numbers on each side of the diagonal indicate the number of images that have been correctly identified. We can assume that the algorithm has been efficient and obtained state-of-the-art results as these numbers are lower relative to the numbers on the diagonal.

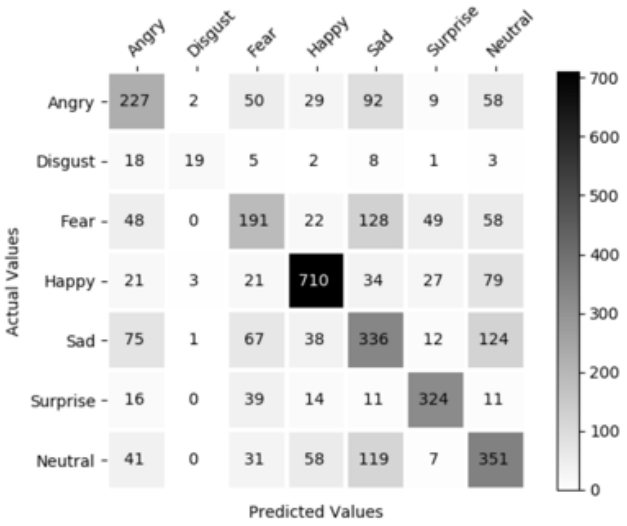

Fig. 7. Confusion Matrix represented as a heatmap [6].

V. Conclusion

In this paper we discussed about an approach for FER using CNN. We created a CNN model on the FER2013 dataset and experiments with the architecture were done to achieve a test accuracy of 0.6012 and a validation accuracy of 0.8978. This state-of-the-art model were used to classify emotions of users in real time with a webcam. The webcam captures images and uses them to classify the emotions.

VI. Future Works

Our remark is that to improve this model, we need more better images with more quality and more specific, we do not forget to mention that also the webcam and the background has a huge impact on the real time classification. Another opinon is that OpenCV should be improved for best image capturation. Also there were less numbers of images compared to the overall such as disgust, so this can affect the accuracy also. So more pictures should be used approximately for all emotions.

REFERENCES

[1] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 2001.

[2] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE Conference on Automatic Facial and Gesture Recognition, 2006.

[3] M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man and Cybernetics, 34(3), 2004

[4] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," in Proc. IEEE Int. Conference on Computer Vision (ICCV), 2015, pp. 3631–3639.

[5] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach," in IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2016, pp. 48–57.

[6] Real Time Facial Expression Recognition Using Deep Learning, Isha Talegaonkar, Kalyani Joshi, Shreya Valunj, Rucha Kohok, Anagha Kulkarni 2019.

[7] Facial Expression Recognition, Amil Khanzada, Charles Bai, Ferhat Turker Celepcikay.

[8] Y. Tang, "Deep Learning using Support Vector Machines," in International Conference on Machine Learning (ICML) Workshops, 2013.

[9] Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. Preprint arXiv:1612.02903v1, 2016.

[10] Human Emotion Recognition using Convolutional Neural Network in Real Time by Rohit Pathar,Abhishek Adivarekar,Arti Mishra,Anushree Deshmukh.

[11] Facial Emotion Recognition using Convolutional Neural Networks by Akash Saravanan, Gurudutt Perichetla, and Dr. K.S.Gayathri