Facial Emotion Detection Using Convolutional Neural Network

K. Shirisha , Mahidhar Buddha

Abstract— **Facial emotion detection is an emerging field which is used in many applications such as social robots, to find customer satisfaction and games. Non-verbal communication methods like facial emotion expressions, eye movement and hand gestures are used in many applications of human computer interaction, which among them facial emotion expression is widely used because it tells about the feelings of persons. The facial emotion detection is not an easy task because there is no perfect distinction between the emotions on the face and also there are a lot of complexity and variability. This model is developed using deep learning approach using a Convolutional Neural Network, on a dataset of fer2013 which consists of 48x48 pixel grayscale images of faces. The detection of facial emotions is done by comparing these physiological signals with the data of this dataset, to classify the affective state of a person.**
This model uses a Convolution Neural Network for feature extraction of the physiological signals and fully connected neural network layers, the emotion prediction is done. The experimental results on the dataset fer2013, showed that the proposed method using CNN achieves a better precision of the classification of the emotional states, in comparison with the general machine learning algorithms.
Index Terms— ***facial emotion; non verbal communication; convolutional neural network; detection;***

---

1. INTRODUCTION

Facial emotion is an important part of nonverbal communication. Human emotion recognition is influenced by certain context. When there is a feedback reviews, the customers might divert using voice tone or argument and may forget to keep track of the facial emotion. Automatic facial emotion recognition systems are used in such cases. These systems can be used in many fields, like gaming applications, criminal investigations, medical practices, animations etc. Facial emotion recognition system mainly identifies seven basic facial emotions such as anger, disgust, fear, happy, neutral, sad, and surprise. Facial emotion recognition techniques are based on either appearance features or geometric features. Geometric features [9] are derived from the shape of the face and other parts such as eyebrows, mouth, nose, lips, and eyes.

Appearance features are derived using the texture of the face which is caused by expression, furrows, wrinkles etc. In 1970s Paul Ekman and Wallace V. Friesen, developed Facial Action Coding System which is the most widely used method for detecting facial emotion. Facial Action Coding System is a system designed for human observers to describe changes in facial expression in terms of noticeable facial muscle movements known as facial action units. Facial Action Coding System is determined to be a powerful means for detecting and measuring facial emotions and is recently used for feature extraction in combination with other methods such as Dynamic Bayesian Network and Local Binary Pattern. Histograms of oriented gradients, Scale Invariant Feature Transform, Local Binary Pattern are few state-of-art techniques used for extracting facial features.

Most of the above methods use simple features for facial emotion recognition, so it requires a lot of efforts both in terms of computation and programming effort. In recent years, deep learning using convolution neural networks for feature extraction of image data has become so popular. This popularity originated from its ability to extract good representations from image data. CNN's computation intensive tasks can run on GPU, which leads to high performance at low power consumption. CNN is majorly used for facial feature extraction for determining gender, age, emotion etc.

2. RELATED WORK

The research for deep learning has brought up development of various applications in the field of computer science. In this literature review, it shows the implementation of emotion analysis using various techniques.

Prudhvi Raj Dachapally (2018) [1] has proposed two independent methods for the purpose of emotion detection. The first one uses representational autoencoders to construct a unique representation of any emotion. Autoencoders are unsupervised deep learning neural network algorithms that are used to reduce the number of dimensions in the data to encode.The second one is to create a convolutional neural network and train it from scratch using fer2013 dataset. His proposed model consists of 8-layer CNN with three convolutional layers, three pooling layers, and two fully connected layers.

Sushmitha, Anand, Chetan Kumar (2019) [2] has proposed that identification of human facial emotion is determined using facial muscles movements. There are various

methods for recognizing facial emotions but before recognizing facial emotion, it is important to detect faces. Since there are many variations in faces, detecting the face is a challenging task. Even in facial detection there are various methods, one such method used is Haar classifier. Haar function is used for face, eyes, lips and mouth detection. Edge detection techniques are used for sharpening and detecting boundaries of the image. Once the face is detected pre existing mobile net architecture is used for implementation.

Shervin Minaee, Amirali Abdolrashidi (2019) [3] has proposed an end-to-end deep learning framework, based on attentional convolutional network, to classify the underlying emotion in the human face. Mostly, improving a deep neural network depends on adding more layers, facilitating gradient flow in the network, or by using better regularizations on data, especially for classification problems with a large number of classes. However, for facial emotion recognition, due to the small number of classes, they developed a model using a convolutional network with less than 10 layers and attention is able to achieve promising results.

MD. Zia Uddin et.al. (2017) [17] has used a depth camera-based novel technique for recognizing the facial expressions. For efficient emotion detection, the rank of each pixel in the depth picture is calculated and used here by eight local directional pixels. For each pixel in a depth image, the eight surrounding directions and eight histograms are implemented by using the corresponding calculated ranks. They have then concatenated these histograms for depth image feature representation.

Wingenbach et al. (2016) [6] has developed a set of video stimuli depicting three levels of intensity of emotional expressions, from low to high intensity. The videos were taken from the Amsterdam Dynamic Facial Expression Set Bath Intensity Variations (ADFES-BIV) dataset; this project included six basic emotions along with contempt, embarrassment and pride, and these are expressed at three different intensities of expression. The three levels of intensity were validated as distinct categories, with higher accuracies and faster responses.

3. BACKGROUND

The proposed model firstly starts with pre processing of data on fer2013 dataset. The fer2013.csv file consists of three columns namely emotion, pixels and usage. The column in pixels contains pixels stored in a list format. As high computational power is needed for computing pixel values in the range of 0-255, the data in pixel field is normalized to values between 0-1 using batch normalization techniques. The face objects stored are reshaped and resized to the predefined size of 48 X 48. The corresponding emotion [4] labels and their respective pixel values are stored in objects. Once data pre-processing is done data augmentation is applied where more data is generated using the training set by applying various transformations. The image data is generated by transforming the existing images [11] by rotation, crop, horizontal shifts, vertical shifts, shear, zoom, flip, reflection, normalization etc. once we get sufficient amount of data we extract needed features from the image using various filters in convolution model. Here we are considering dominant edges and regions as features like detecting eyes, lips. Using these features we train our model for 25 epochs and try to minimize loss function for each epoch. If loss function is not getting minimized for 3 consecutive epochs [10] we stop our model training which is called early stopping phase. For each epoch we test our model using public test dataset which is used for determining accuracy of model. Once training of model is completed we use this model by giving input from web camera and check corresponding results.

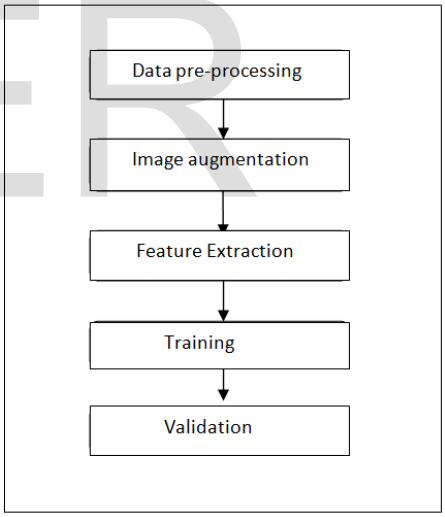


Fig. 1: Architecture of the proposed system

4. METHODOLOGY

The setup used for this model is Anconda3. Anaconda is data mining software which provides data mining apparatus such as jupyter, spyder, orange which manage environments with conda consisting of all the packages predefined. The process flow of the model is illustrated in

the four main steps starting with the dataset preparation and validating the model. These steps are explained as A, B, C, and D.

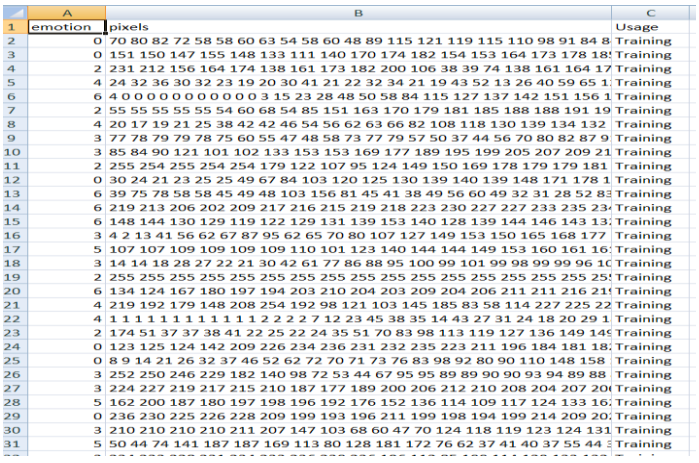
A. *Dataset information*

The data set [13] consists of 48*48 pixel gray scale images of faces. The faces have been categorized into facial expression in to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).The training set of this data set contains of 28,709 examples. The public test data set contains of 3,589 examples. The final private test data set, which was used to determine the winner of the competition, which contains 3,589 examples. The dataset was prepared by Aaron Courville and Pierre-Luc Carrier, for their ongoing research project.

Table 1: list of attributes

Attribute	Possible values
emotion	Class of corresponding image.
pixels	Contains image in form of array.
purpose	Whether image is used for training or Testing.

Fig. 2 displays the screenshot of the actual dataset used for this model in fer2013.csv file. It shows attributes mentioned in the Table 1.



	A	B	C
1	emotion	pixels	Usage
2	0	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121 119 115 110 98 91 84 8	Training
3	0	151 150 147 155 148 133 111 140 170 174 182 154 153 164 173 178 18	Training
4	2	231 212 156 164 174 138 161 173 182 200 106 38 39 74 138 161 164 17	Training
5	4	24 32 36 30 32 23 19 20 30 41 21 22 32 34 21 19 43 52 13 26 40 59 65 1	Training
6	6	4 0 0 0 0 0 0 0 0 0 0 0 3 15 23 28 48 50 58 84 115 127 137 142 151 156 1	Training
7	2	55 55 55 55 55 54 60 68 54 85 151 163 170 179 181 185 188 188 191 19	Training
8	4	20 17 19 21 25 38 42 42 46 54 56 62 63 66 82 108 118 130 139 134 132	Training
9	3	77 78 79 79 78 75 60 55 47 48 58 73 77 79 57 50 37 44 56 70 80 82 87 9	Training
10	3	85 84 90 121 101 102 133 153 153 169 177 189 195 199 205 207 209 21	Training
11	2	255 254 255 254 254 179 122 107 95 124 149 150 169 178 179 179 181	Training
12	0	30 24 21 23 25 25 49 67 84 103 120 125 130 139 140 139 148 171 178 1	Training
13	6	39 75 78 58 58 45 49 48 103 156 81 45 41 38 49 56 60 49 32 31 28 52 83	Training
14	6	219 213 206 202 209 217 216 215 219 218 223 230 227 227 233 235 23	Training
15	6	148 144 130 129 119 122 129 131 139 153 140 128 139 144 146 143 13	Training
16	3	4 2 13 41 56 62 67 87 95 62 65 70 80 107 127 149 153 150 165 168 177	Training
17	5	107 107 109 109 109 109 110 101 123 140 144 144 149 153 160 161 16	Training
18	3	14 14 18 28 27 22 21 30 42 61 77 86 88 95 100 99 101 99 98 99 99 96 1	Training
19	2	255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255 255	Training
20	6	134 124 167 180 197 194 203 210 204 203 209 204 206 211 211 216 21	Training
21	4	219 192 179 148 208 254 192 98 121 103 145 185 83 58 114 227 225 22	Training
22	4	1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 7 12 23 45 38 35 14 43 27 31 24 18 20 29 1	Training
23	2	174 51 37 37 38 41 22 25 22 24 35 51 70 83 98 113 119 127 136 149 14	Training
24	0	123 125 124 142 209 226 234 236 231 232 235 223 211 196 184 181 18	Training
25	0	8 9 14 21 26 32 37 46 52 62 72 70 71 73 76 83 98 92 80 90 110 148 158	Training
26	3	252 250 246 229 182 140 98 72 53 44 67 95 95 89 90 90 93 94 89 88	Training
27	3	224 227 219 217 215 210 187 177 189 200 206 212 210 208 204 207 20	Training
28	5	162 200 187 180 197 198 196 192 176 152 136 114 109 117 124 133 16	Training
29	0	236 230 225 226 228 209 199 193 196 211 199 198 194 199 214 209 20	Training
30	3	210 210 210 211 207 147 103 68 60 47 70 124 118 119 123 124 131	Training
31	5	50 44 74 141 187 187 169 113 80 128 181 172 76 62 37 41 40 37 55 44	Training

Fig. 2: Sample part of dataset in excel sheet

B. *CNN model construction*

The CNN model contains two convolutional layers along with batch normalization .Then output of these is given as input to other batches among them one batch is treated with a convolutional layer and other batch is treated with a seperable-convolutional layer. Then the output from these layers are followed my max pooling and global average pooling and then finalized by the softmax algorithm.

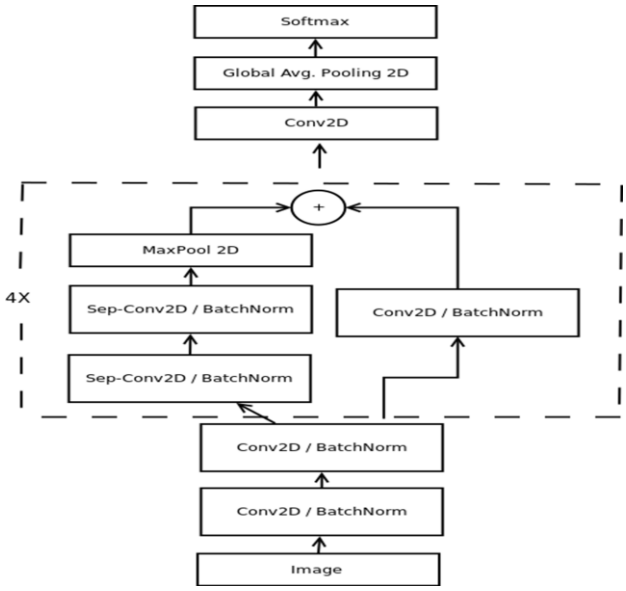


Fig. 3: flow chart of CNN model

i) Convolutional 2D

The convolution 2D is an easiest operation where you start with a kernel, which consists of all small matrix of weights. This kernel slides by single step each time over the input data and performs matrix [6] multiplication with feature matrix, and then summing up the results into a single output pixel. In order to get output in same dimension as input image we use zero padding technique.

ii) Batch Normalization and Max pooling 2D

Batch normalization is a method used to decrease internal covariance shift in neural networks and also improves the speed, stability and performance in convolutional neural networks [8][9]. In max pooling, a kernel of size n*n is moved across the matrix and for each position the maximum value is taken and kept at the corresponding position of the output matrix.

iii) Global Average Pooling

Global average pooling layers tries to minimize over fitting of data by reducing the total number of parameters in the model. Global average pooling

layers are used to reduce the spatial dimension of a three-dimensional tensor.

iv) Softmax

The softmax function takes a vector of N real numbers as input and normalizes that vector into values ranging from 0 to 1.

v) Activation function

To reduce over fitting of data activation functions are used. In this model we used ReLu [10] activation function. The main advantage of ReLu is its gradient is always equal to 1. Negative values in matrix input are always changed to zero and all other positive values remain constant.

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

C. *Optimizer, Loss function and Metrics*

Loss function is used to measure the absolute difference between our prediction and the actual value that is present in validation dataset. Loss function [5] used in this model is categorical cross entropy. Cross-entropy loss indicates the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss function output value changes as the predicted probability differs from the actual output.

Optimizers are used to minimize loss function by updating attribute values of neural network. Optimizer used in our model is the Adam () optimizer [12]. Adam stands for Adaptive Moment Estimation. Adaptive Moment Estimation is used to compute adaptive learning rates for each attribute.

D. *Training the model*

After all the initial steps of data collection, cleaning and pre-processing we need to fit the data into the proposed CNN model. In order to train the model the following 3 steps are performed on the encoded dataset:

1) Dividing the data into features and target: Firstly, to train our model we need the training set of data. The train data contains complete set of feature variables (independent variables) and target variable (dependent variable). So, we need to separate all the features that are used for predicting the target variables in our dataset. Class label is the only target variable that is predicted for all the feature variables.

2) Partitioning the data into training and testing set: There are many ways for the train-test splitting ratio but as out data set consists of usage column which distinguishes entire data set from training data set and testing dataset.

3) Fit the model: The proposed model is build using the CNN image classifier. Then the training set is supplied to the model. The model learns the corresponding emotions from the training set and trains itself accordingly. Now, the time for validating the model starts where the test set features are fitted in the model and the target variable value is predicted.

5. EXPERIMENTAL RESULTS

Once the final model is trained it is used for predicting the results. The test set is shuffled and given as input to the model. After the training, the test set is given and the output produced by the model is compared with test set to compute the performance of the trained model. The designed model was implemented using keras. The training process was applied for 25 epochs with a batch size set to 32. The training took around 120 minutes. In the implementation phase, various OpenCV functions and Keras functions have been used. First the video frame is stored in a video object. A Haar cascade classifier is used to detect facial region of an image. The image frame is converted into gray scale and resized and reshaped with the help of numpy. This resized image is given as input to the model which is loaded by keras.models.load_model() function. The max argument is output. A rectangle is drawn around the facial regions and the output is shown above rectangular box.

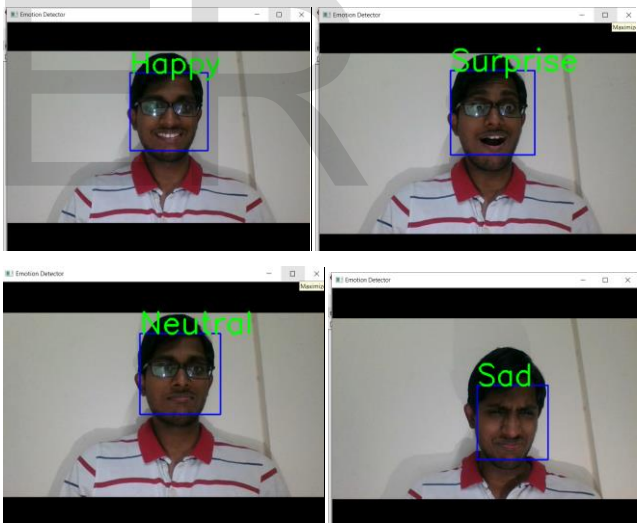


Fig. 4: Results of Emotion detection using CNN

An accuracy of 62 % was achieved. It is evident that the overall accuracy is not too high. These may be due to transfer learning and less dataset for each class of emotion. It is proposed to use the transfer learning with large amount of data for better accuracy and try various combinations in designing convolution layers.

6. CONCLUSIONS

Many researches and studies about Emotion Recognition, Deep learning techniques used for recognizing the emotions are conducted on data set. It is required in future to have a model like this with much more accuracy and reliability, which has many applications in many fields. This paper contains a study of some of the facial emotion recognition systems based on CNN. This study helps to understand different kinds of models for facial emotion recognition and to develop new CNN architectures for better performance and accuracy. Tensor Flow and keras are used to train the model. Accuracy rate of about 62% is achieved. In future, any real time emotion recognition can be developed using the same architecture with minor modifications.

REFERENCES

[1] Prudhvi Raj Dachapally, "Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units" arXiv:1706.01509v1 [cs.CV] 21 Jun 2018.

[2] Sushmitha, Anand, Chetan Kumar, " Facial Emotion Detection using Convolutional Neural Network " IJSTE - International Journal of Science Technology & Engineering | Volume 5 | Issue 9 | March 2019

[3] Shervin Minaee , Amirali Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network" arXiv:1902.01019v1 [cs.CV] 4 Feb 2019.

[4] "Automatic Emotion Recognition Using Facial Expression: A Review" – IRJET by Monika Dubey, Prof. Lokesh Singh.

[5] "Emotion Recognition from Facial Expression Using Multilevel HMM" by Ira Cohon, Ashutosh Garg, Thomas.S. Huang

[6] T. S. Wingenbach, C. Ashwin, and M. Brosnan, "Correction: Validation of the amsterdam dynamic facial expression set–bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions," PloS One, 2016, vol. 11, no. 12, p. e0168891.

[7] Recognizing Facial Expressions Using Deep Learning by Alexandru Savoiu, Stanford University and James Wong Stanford University [http://cs231n.stanford.edu/reports/2017/pdfs/224.pdf]

[8] W. Cheng, L. Jingtian, "Facial expression recognition based on geometric features and geodesic distance," Int J Signal Process, vol. 7(1), 2014, pp. 323–330.

[9] Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu. "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition." arXiv preprint arXiv:1804.10892, 2018.

[10] Shima, Yoshihiro, and Yuki Omori. "Image Augmentation for Classifying Facial Expression Images by Using Deep Neural Network Pre-trained with Object Image Database." Proceedings of the 3rd International Conference on Robotics, Control and Automation. ACM, 2018

[11] Abir Fathallah, Lofti Abdi, Ali Douik., "Facial Expression Recognition via Deep Learning", 2017 IEEE/ACS 14th AICCSA.

[12] Tom McLaughlin, Mai Le, Naran Bayanbat," Emotion Recognition with Deep-Belief Networks", September 2017.

[13] Raghuvanshi, Vivek Choksi, "Facial Expression Recognition with Convolutional Neural Networks" published on Semantic scholar 2016.

[14] Kaggle Dataset. https://www.kaggle.com/deadskull7/fer2013

[15] L. Cavidon ,H. M. Fayak, M. Lich,"Evaluating deep learning architectures for Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[16] P. Hairár, R. Burgeet, and M. K. Duitta, "Facial Emotion Recognition with Deep Learning," in Signal Processing and Integrated Networks (SPIN), 2017, pp. 4–7.

[17] Md. Zia Uddin,Weria Khaksar,Jim Torresen,"Facial Expression Recognition Using Salient Features and Convolutional Neural Network" 10.1109/ACCESS.2017.2777003,December 22, 2017.

[18] P. R. Khorrami, How deep learning can help emotion recognition, in Electrical & Computer Eng., 2017, University of Illinois, Urbana-Champaign, p. 92

[19] M. Olszanowski et al., "Warsaw set of emotional facial expression pictures: a validation study of facial display photographs," Frontiers in Psychology, 2015, vol. 5, p. 1516.