IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. I (May. –June. 2022), PP 36-43
www.iosrjournals.org

Emotion Recognition Using Artificial Intelligence - A Review

Gulrukh Shahab, *M.Tech (ECE), DPGITM Gurugram*
Dr. Rakhi Dua *Asst. professor, DPGITM Gurugram*

Abstract- *Here, an overview of the work done in past related to emotion recognition using Artificial Intelligence is discussed. Emotion, a strong mental process involves n number of muscles. This n is decided by the type of emotion. Various emotions viz anger, happiness, sadness, grief, frowning, horror (terror, flight), surprise, doubt, grinning, contempt, whistling etc. require different number of muscle contraction, different pattern of muscle contraction. Human beings as well as several other living beings are built naturally to recognise and distinguish these and various other emotions not mentioned here. Sometimes there is a requirement that the different emotions can be recognised quickly and effectively many times faster than humans. Here comes the solution in the form of Artificial Intelligence. Some techniques like Support Vector Machine (SVM), Convolution Neural Network (CNN) are used to categorize emotions viz verbal, non-verbal, speech and various facial expression etc.*

Date of Submission: 18-05-2022 Date of Acceptance: 02-06-2022

I. Introduction

Emotion Recognition refers to the procedure of recognising the emotions from speech, facial expressions etc. Several emotions are there which are interpreted by humans very efficiently. But in some conditions, humans can't act instantaneously after observing various emotions like in some medical conditions. Like various other fields, emotion recognition is also a current field where various researches and studies are going on for utilizing Artificial Intelligence. How do emotions occur in human bodies? As we can infer from 'abstract' paragraph above that, various changes take place for expressing various emotions. A very typical way of contraction of seven muscle groups is utilized for expressing anger. For the expression of happiness, greater zygomatic muscle is involved. Expressing sadness is done by Levator Labii superiosis. Anguli Oris express grief. Frowning is expressed by Corrugator Supercilli and Procerus. Platysma is used for showing the emotion of horror. Mentallis expresses doubt. Grinning is expressed by Risorius. Emotion of contempt by Zygomaticus minor. All these emotions can be recognised using Artificial intelligence by recording and analysing the different types of muscles and their different pattern of contractions by using audio visual methods.

II. Related Works

The advancement and utility of Computer software, networks and hardware has taken a fast pace. A considerable progress has been made in developing automatic expression classifiers by some researchers during recent years. Some expression recognition systems classify the face into a set of emotions such as emotion of sadness, emotion of happiness, emotion of anger and other types of emotions. Other types of recognition systems try to recognize the different movements of muscles that can be produced by face in order to present a formal description of face. An interesting psychological system for the interpretation of complete facial movements is known as FACS (Facial Action Coding Systems). This system uses Action Units (AU) on the basis of their appearance on face. One element out of 46 atomic elements is an AU. This AU is an element of facial movements of face or deformation of face. A number of AUs combined together, form an expression. Several advancements have taken place in expression recognition techniques. Neural Networks, Multilevel Hidden Markov Model (HMM) and Bayesian Networks. The drawbacks of timing or recognition rate are there. For acquiring a good level of recognition, two or more than two techniques can be combined. Then the desired features can be extracted. If the images are pre-processed based on feature extraction and illumination than the chances for a technique to be successful are high. The success of each technique is dependent on pre-processing of the images because of illumination and feature extraction.

III. Methods

CNNs with different depths were developed for evaluating the performance of the models for facial expression recognition. The network architecture considered in the investigation are as follows: (ReLU, Conv (SBN), Max-pool, and drop out, Affine (BN), Affine and Softmax.

M convolution layers are referred in the first part of the network. Batch Normalization (SBN), max pooling and Dropout are possessed by these convolution layers in addition to the ReLU nonlinearity and convolution layers. The fully connected layer N comes after these layers i.e., the M convolution layers.

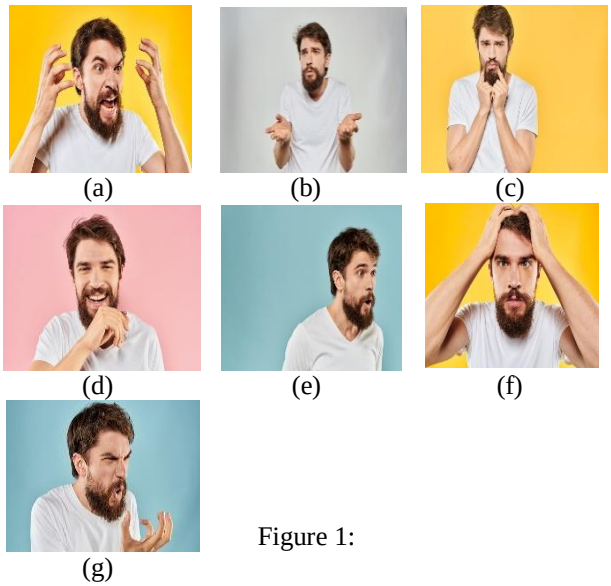
These layers always have ReLU nonlinearity and affine operation. Dropout and batch normalization (BN) can also be included by these layers. Scores and Softmax loss functions are computed by the affine layer. And the network follows this layer. Max pooling layers, dropout, existence of batch normalization and the number of fully connected and convolution layers are decided by the user because of a developed model.

L2 regularization, drop out as well as batch normalization was included in the implementation. Quantities of strides, zero-padding, filters can be given by user. Default values are considered if these are not given. An idea to combine features extracted by convolutional layers with HOG feature through raw pixel data was proposed.

Till here same architecture was utilized with a difference of added HOG features to the feature exiting the last convolution layer. Score and loss calculation were done in the fully connected layers by the hybrid feature set.

Emotion recognition through AI is an active area in the present days' ongoing research of computer vision. Automatic interpretation of visual inputs is done for assessing sentiments and detecting emotions.

An algorithm inspired from the brain of human beings or animals, is ANN i.e., Artificial Neural Network. The network where convolution as the mathematical operation is used is called Convolution Neural Network. A CNN is a neural network which have some convolution layers. These convolution layers are used for the purpose of image processing, segmentation, classification and also other types of automatic related data. The network uses two dimensional CNN for the purpose of recognising a data having images. One another commonly used neural network is DCNN i.e., Deep Convolution Neural Network where video and images with different patterns can be recognised. This network is trained to differentiate several facial emotions with good accuracy. This training is performed through the set of data collected with the help of camera of cell phone.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 1:

Photos of facial expressions like smile, sad, surprise, anger, happiness, disgust, neutral etc. Here the model mentioned above is implemented in Torch and an advantage of GPU accelerated deep learning features is taken so as to process the model training faster.

IV. Data Set And Features

Kaggle website had provided the dataset of images of faces. These images of faces were consisted of approximately 37000 pixel-grey images of 48 x 48 size. The processing of these images is such that each face occupies the equal space and also each face has got a centre position in the image. Each image has got one category out of different categories, for a particular facial emotion. These different emotions are: Angry, Disgust, Fear, Happy, Sad Surprise and Neutral. In the figure above, there is a depiction of one example for each emotion. There are 3 sets viz, validation, training and test sets into which the different images are divided. The number of images for validation, training and testing are 4000, 29000 and 4000 respectively. After reading the raw pixel data, these were normalized by subtracting the mean of the training images from each image. Here validation and test set images were also included. Mirrored images were produces by flipping them horizontally in the training set for data augmentation purpose.

For the purpose of classification of expressions, the features generated by convolution layers using the raw pixel data were used. For more explorations, HOG features generated by convolution layers were linked with learning models. And these were given as input into FC i.e., fully connected layers.

Parameter	Value
Learning Rate	0.001
Regularization	1e-6
Hidden Neurons	512

Table 1: showing the hyper-parameters obtained by cross validation for the shallow model

V. Analysis

A shallow CNN was built to carry out this project. There were two convolutional layers in this network. This network also had one FC layer. The first convolution layer had filters of size 323x3 with 1 as stride size. These had dropout and batch normalization with no max-pooling. The second convolutional layer had 643×3 filters with 1 as the stride size. These had dropout and batch normalization also. These had the feature of max-pooling also, of filter size2x2. There was a hidden layer with 512 neurons and Softmax as the loss function in the FC layer. Rectified Linear Unit (ReLU) as the activation function was used in all the layers. Some sanity checks were done to make sure that the implementation of the network was correct, before training the model. The initial loss was computed with the absence of regularization for carrying out first sanity check. Because the classifier had 7 distinct classes, it was expected to get an approximate value of 1.95. In order to do a second sanity check, it was tried to fix the model using a small portion of the training set. This shallow model was successful in both of these sanity checks. Then, scratch was used to train the model. Deep learning facilities accelerated by GPU were exploited on torch for faster processing of the training. In order to process the training, all of the images were used in the training set with 128 as batch size and 30 epochs and hyper parameters of models, with different number of hidden neurons, learning rate and distinct values for regularization were cross-validated. The validation set was used for validating the model in each iteration. The test set was used for evaluating the performance of the model. A shallow model considered best, had given an accuracy of 54% on the test set and 55% on the validation set. After cross validating for the shallow model, the hyper parameters are shown in table1.

For the purpose of observation of the effect of adding convolutional layers and FC layers to the network, a deeper CNN with two FC layers and four convolution layers were trained. In the first convolution layer, there were 643x 3 filters. The number of filters in the second convolution layer were 1285x5. The third convolution layer had 5123x3 filters. The number of filters in the last convolution layer were 5123x3.There was a stride of size 1 in all the convolution layers. The activation functions were drop out, batch normalization, ReLU and max-pooling. There were 256 neurons in the hidden layer of first FC layer and 512 neurons in the second FC layer. Dropout, ReLU and batch normalization was used both in FC layer and convolution layer. As a lost function, Softmax was used.

The structure of the deep network is shown is figure 2. In the shallow model, before network training, checking the initial loss and examination of the capability of the network to be fit, was performed where a small portion of the training set was used. It could be concluded from the sanity checks that network implementation was correct.

The network containing all images in the training set, was trained. Here a batch size of 128 and 35 epochs were used. Now, an accuracy of 64% on the test set and 65% on the validation set was achieved. Values for the hyper parameter having the highest accuracy in the model is shown in figure 2.

Networks with 5 and 6 convolution layers were also trained for exploring the deeper CNNs, but classification accuracy was not increased. So, the model with 2 FC layers and 4 convolution layers was considered the best network for the dataset. Features generated by the convolution layers, which used the raw pixel data as the main feature for the performing classification work were only examined, in both the deep as well as shallow models.

Because of being sensitive to edges, usually HOG features are used for facial expression recognition. Researchers wanted to explore the performance of the model when it has a combination of HOG features along with raw pixels. So, a model containing two neural networks was built. One network was containing the convolution layers and the other one was the network containing fully connected layers. The features received from the first network were linked to the HOG features and the resulting hybrid features were linked into the second network. For assessing the performance of network with hybrid features, two networks, one deep network and the other shallow network were trained. The characteristics of these networks were the same as the deep and shallow networks in the previous experiment. Now, the shallow model had an accuracy, very close to the accuracy of the previous shallow model which used only raw pixels. The deep model also got an accuracy similar to accuracy of the previous deep model which used raw pixels.

Parameter	Value
Learning Rate	0.01
Regularization	1e-7
Hidden Neurons	256, 512

Table 2: The hyper- parameters obtained by cross validation for the deep model

VI. Results

For comparing the performance of deep and shallow model, loss history and accuracy obtained in these models were plotted. The results are shown in figure 3 and 4. It can be seen in fig.4 that validation frequency was increased by 18.46% because of deep network. Over fitting behaviour of learning model was reduced because of addition of more non-linearity and hierarchical usage of anti over fitting techniques such as batch normalization and dropout in addition to L2 regularization. We can see from figure 3 that highest value was quickly achieved by the training accuracy and the shallow network converged at fast pace.

Confusion matrices were also computed. Visualization of the confusion matrices can be seen in figure 5 and figure 6. It is depicted in these figures that the deep network causes higher true predictions for most of the labels. It was concluded by the models that it is easier to interpret the emotion of a glad face as compared to other expressions. The grid shows the category which have a possibility to be misguided by the trained networks. As an example, the trained network can have a confusion to distinguish between sorrows, furious or scared expressions.

Expression	**Shallow Model**	**Deep model**
furious	41%	53%
Disgust	32%	70%
Scared	54%	46%
Glad	75%	80.5%
Sorrow	32%	63%
Surprise	67.5%	62.5%
Neutral	39.9%	51.5%

Table 3: Depiction of accuracy of different expression in the shallow and deep models

As a human also it is difficult to distinguish between a sorrow and furious emotion. This may be due to the reason that different people express same emotion differently.

In table 3, conclusion of computation of accuracy for many emotions is shown. Accuracy for several expressions has increased because of the use of deep networks. Many expressions like scared emotion and surprise have got a good degree of prediction as well as decreased accuracy. It can be inferred from this that going deeper does not always provide better features.

For investigating the effect of utilizing different features in the CNN model, learning models that concatenate the HOG features with those generated by the convolutional layers were developed. And these were used as input feature to the FC layers. So, one shallow and one deep network were trained.

Accuracy obtained in different repetitions for the shallow model is shown in figure seven and for deep model, it is shown in figure eight. Here accuracy received, is very close to the accuracy from the model which is not having HOG characteristic.

So, it can be concluded that CNN is sufficiently powerful to pull out enough information which includes information with raw pixel taken out of HOG characteristic.

Imagination of activation maps of different layers during the forward pass was done to observe the characteristics that the trained network pulls out at every layer. This visualization is depicted in figure nine.

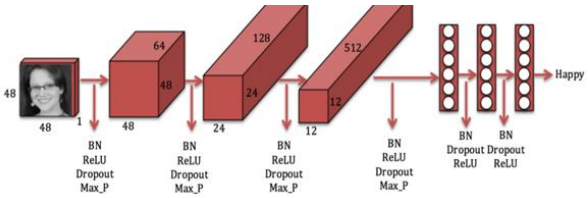


Figure 2: The Architecture of Deep network. 4 Conventional layers and 2 connected layers.

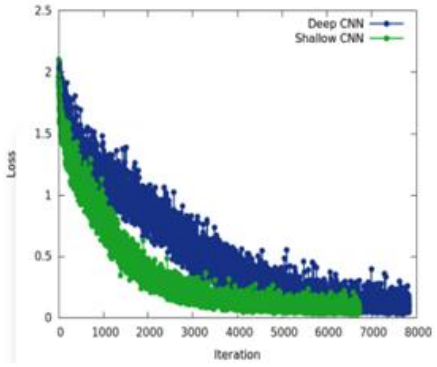


Figure 3: The loss history of the shallow and deep models

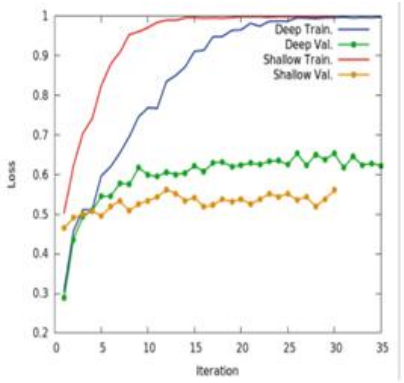


Figure 4: The accuracy of the shallow and deep models for different numbers of iterations

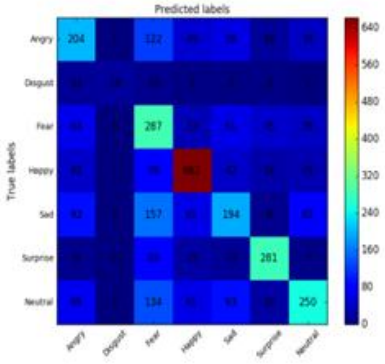


Figure 5: The confusion matrix for the shallow model

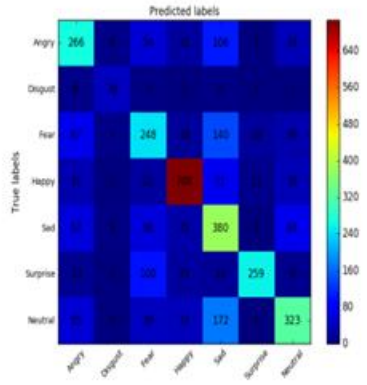


Figure 6: The confusion matrix for the deep model

For finding qualification of the trained network, weights of the first network were visualised.

It can be seen from figure 10, that there are smooth filters without any noisy pattern.

It can be concluded that network has been trained adequately long enough and regularization strength is also sufficient

Deep Dream technique for the best predictive model for finding enhanced pattern in the images, was applied. An example for each expression along with its Deep Dream output is displayed by figure 11.

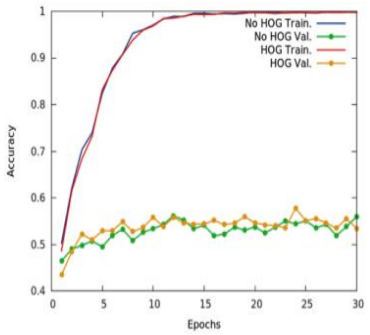


Figure 7: The accuracy of the shallow model with hybrid features for different numbers of iterations

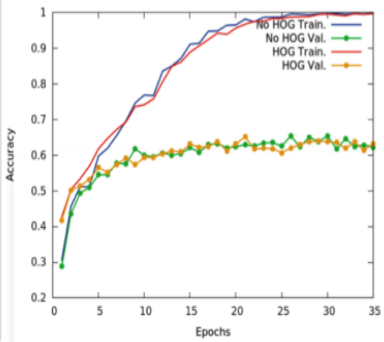


Figure 8: The accuracy of the deep model with hybrid features for different numbers of iterations

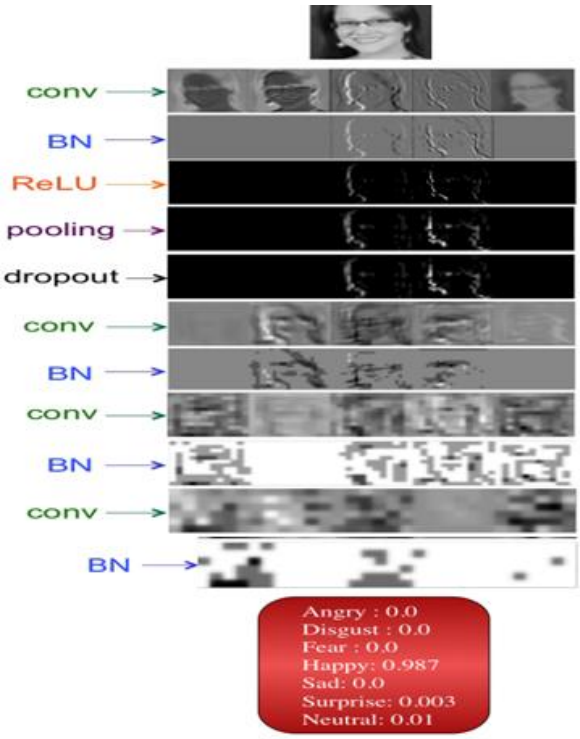


Figure 9: Visualisation of activation maps for different layers in our CNN.

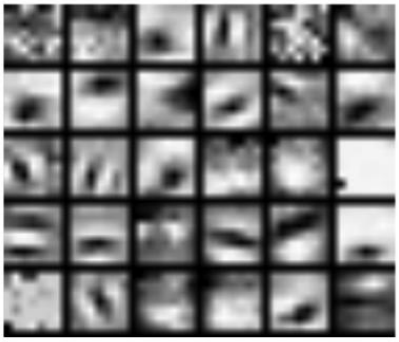


Figure 10: Visualization of the weights for the first layer in our CNN

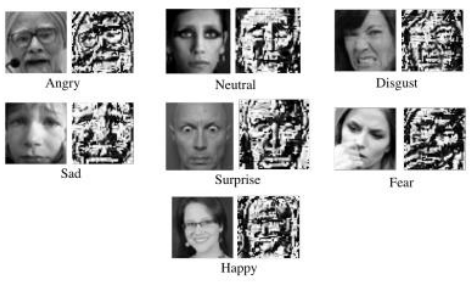


Figure 11: Examples of applying DeepDream on our dataset

[The pictures are taken from "Convolution Neural Network for Facial Expression Recognition by Shima Alidade, Stanford University]

VII. Summary

7.1 Conclusion

Several CNNs for a facial expression recognition purpose were developed. Their performance by utilizing various visualization and post processing techniques were evaluated. From all the results it was concluded that deep CNNs are efficient in learning facial characteristics and these can also improve facial emotion detection.

Hybrid feature sets were not helpful for the improvement of model accuracy. So, it can be inferred that convolutional networks can intrinsically learn the features of the face only by utilizing raw pixel data.

7.2 Future Work

In this project, models from scratch using CNN packages in Torch were trained. Future work can include, models extended to colour images. In this way efficacy of pre trained models such as Alex Net [18] or VGGNet [19] for facial emotion recognition can be investigated.

An added feature that can be implemented will be the process of emotion estimation following the procedure of face detection.

References:

[1]. Shima Alizadeh (Standford University)
[2]. Azar Fazel (Standford University) on "Convolutional Neural Networks for Facial Expression Recognition"
[3]. Umadevi V, Associate Professor, Supreet U Sugur, Sai Prasad K, Prajwal Kulkarni, Shreyas Deshpande on " Emotion Recognition using Convolution Neural Network".
[4]. Pranav E (School of Engineering), Suraj Kamal (Department of Electronics, Cochin University), Satheesh Chandran C. (Department of Electronics, Cochin University), Supriya MH (Department of Electronics, Cochin University) On " Facial Emotion Recognition using Deep Convolutional Neural Network"
[5]. Anvita Saxena, Ashish Khanna, Deepak Gupta on "Emotion Recognition and Detection Methods".
[6]. Eriko Kuramoto, Saori Yashinaga, Seiji Nemoto on "Characteristic of Facial Muscle Activity during voluntary Facial Expressions".
[7]. Article by Varun Pandula on " What are the muscles responsible for Facial Expressions".

Gulrukh Shahab, et. al. "Emotion Recognition Using Artificial Intelligence - A Review." *IOSR Journal of Computer Engineering (IOSR-JCE),* 24(3), 2022, pp. 36-43.