Emotion Recognition Through Facial Gestures - A Deep Learning Approach

Shrija Mishra(✉), Geeta Ramani Bala Prasada(✉), Ravi Kant Kumar(✉), and Goutam Sanyal(✉)

Department of Computer Science and Engineering, National Institute of Technology Durgapur, Durgapur, India
shrija.mishra102@gmail.com, geetabala9@gmail.com, vit.ravikant@gmail.com, nitgsanyal@gmail.com

Abstract. As defined by some theorists, human emotions are discrete and consistent responses to internal or external events which have significance for an organism. They constitute a major part of our non-verbal communication. Among the human emotions, happy, sad, fear, anger, surprise, disgust and neutral are the seven basic emotions. Facial expressions are the best way to exhibit emotions. In this era of booming human-computer interaction, enabling the machines to recognize these emotions is a paramount task. There is an amalgamation of emotions in every facial expression. In this paper, we identified the different emotions and their intensity level in a human face by implementing deep learning approach through our proposed Convolution Neural Network (CNN). The architecture and the algorithm here yield appreciable results that can be used as a motivation for further research in computer based emotion recognition system.

Keywords: Face detection · Emotion recognition · Human-computer interaction Convolutional Neural Network (CNN) · Deep learning · Cross validation · SVM

1 Introduction

Communication plays a key role in our daily lives. In this era of technology, human-computer interaction (HCI) and automation, emotional recognition has become an indispensable field of study. Facial expressions and our actions are non-verbal means of communication which comprise of 93% human communication, of which facial gestures and human actions have 55% role [1]. Facial expressions are universal and important in establishing interpersonal relations. There are seven basic emotions [2]. These include happy, sad, anger, disgust, surprise, fear and neutral. All other emotions are a result of the heterogeneity of these emotions.

Some significant contributions made in this area are Facial expression recognition based on Local Binary Patterns [3]. Emotion recognition using binary decision tree [4], Facial Expression Recognition with Convolutional Neural Networks [5]. Modular Eigen spaces method for emotion classification using NN and HMM [6], Emotion analysis in visual and audio cues [7], Combining multiple kernel methods [8]. But these computational methods have far behind than human accuracy as their foundation is not based on the functioning of human deep learning and training.

A. Ghosh et al. (Eds.): MIKE 2017, LNAI 10682, pp. 11–21, 2017.
https://doi.org/10.1007/978-3-319-71928-3_2

The objective of our research is to examine the facial emotion in static images using various attempted Convolutional Neural Network (CNN). CNN [9] is a special kind of deep learning method that provides solutions to many problems in image recognition after huge training. Due to lack of large amount of training it is difficult even for the humans to detect an emotion in a face. For example, we cannot absolutely determine whether a person is surprised or happy. Thus, we try to delve into the matter and analyze different level of emotions present in a human face at an instance. FER-2013 [10] database present these emotions into 7 categories Neutral, Happy, Sad, Surprise, Disgust, Fear, Anger. Accuracy of 63.03% was obtained on absolute classifications. For the ambiguous emotions, considering the top 2 results as correct, we achieved an accuracy of 67%. To improve the performance, we applied regularizations, dropout, batch normalization using grid search and transfer learning.

Further, the paper is divided into 10 sections. Section 2 describes the dataset. In Sect. 3, pre-processing task has been applied on the dataset. Section 4 comprises SVM. Overall architecture of our proposed system has been mentioned in Sect. 5. CNN and our proposed network are discussed in Sects. 6 and 7 respectively. In Sect. 8, finally selected proposed network has been described. Section 9 comprises of emotions results and finally, Sect. 10 draws the concluding remarks.

2 Dataset

FER2013 dataset [10] has been used for the experiment. It consists of 37887 pre-cropped gray scale images with size of 48 * 48. The images are labeled in 7 emotions (0 = Anger, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral) (Fig. 1).

Fig. 1. FER2013 dataset sample

3 Preprocessing

In general scenario, human vision system, first detects the faces, and then subsequently it recognizes the emotion associated with that face. In the same way, in this work, face detection is the pre-processing or prior work of the emotion recognition task. Face detection task has been done using Viola Jones algorithm [11] (Fig. 2).

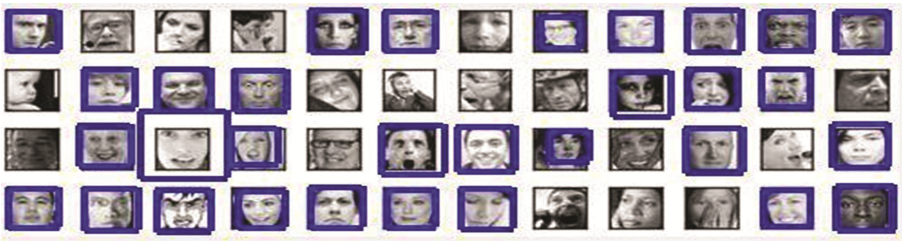

Fig. 2. Detected faces using Viola Jones algorithm

Haar Feature-Based Cascaded Classifier [11] is applied on all the images. This forms a bounding box around the face in the images. The area inside the bounding box is cropped and reshaped into 48 × 48 pixels. After pre-processing, the dataset consists of 11,246 images of the 7 emotions of which 1456 are angry, 240 are disgust, 1414 fear, 3235 happy, 1304 sad, 1362 surprise and 2235 are neutral. All the images are of frontal face. Non frontal faces (image of side face) and non-relevant images (images that were some random image or those with hands covering face, etc.) were removed (Fig. 3).

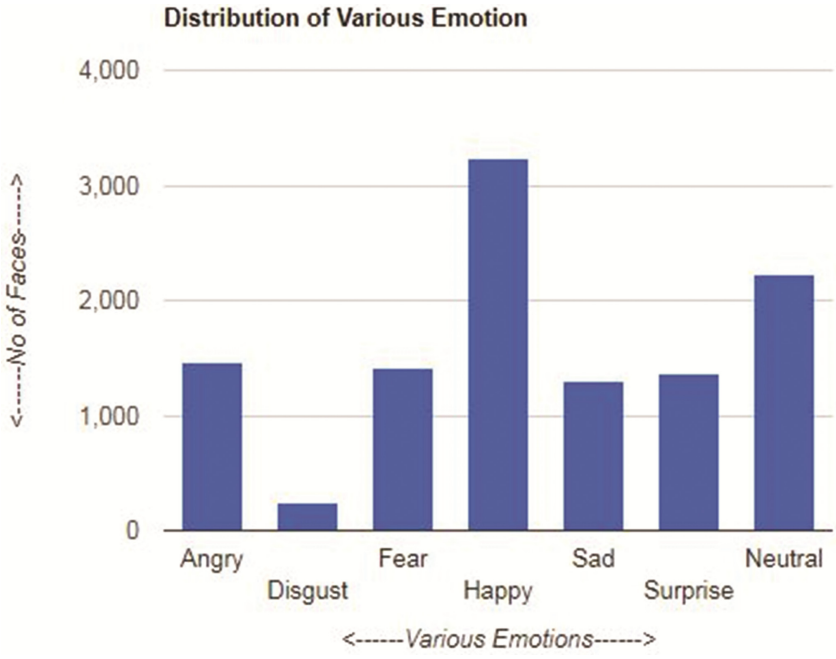

Fig. 3. Distribution of emotions after pre-processing

4 Emotion Prediction Using SVM

First, we applied SVM [12], previously the best-known image classification technique for testing its efficiency in the work of emotion detection. It is a supervised learning classification method that relies on results from statistical learning theory to guarantee high generalization performance. They are non-parametric models that need proper parameter tuning. The complexity and the computational cost grow with the number of training samples and the number of classes. The pre-processed images were used for training and testing. This multi class classification was carried out using the SVC function of scikit learn library. Training was performed on 70% of the data and the rest was used for testing. An accuracy of only 46.74% was attained on the test data.

To try out a different method and for better performance, we went on to deep learning that is the most trending area of research and application in this era as it is known to give the best results to complex problems such as image classification, natural language processing, and speech recognition.

5 System Architecture

For better understanding, overall architecture of our proposed work has been divided into two phases. They have been termed as training and testing phase (Fig. 4).

Fig. 4. System architecture

5.1 Training Phase

The proposed network has been trained with about 7800 training images (70% of the images after pre-processing) taken from FER-2013 [10] database. This database contains seven standard categories of emotions for every subject. During training process deep convolutional neural network has been applied for feature extraction and training classification. It uses supervised learning approach over huge number of images. The proposed convolutional neural network (CNN) has been explained in Sect. 8 (Fig. 5).

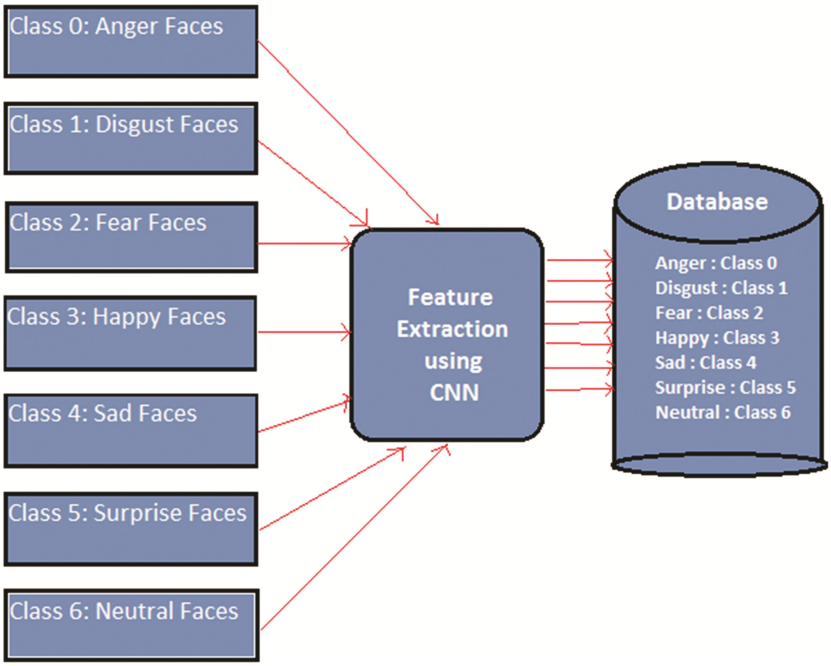

Fig. 5. Proposed training architecture

5.2 Testing Phase

The proposed network has been tested with about 3400 test images (30% of the images after pre-processing) taken from FER-2013 database. Like training phase, in the testing

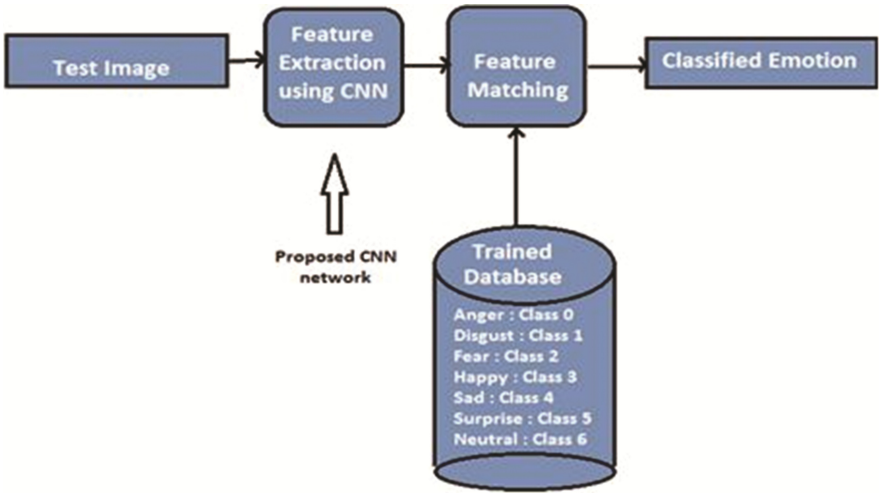

Fig. 6. Proposed testing architecture

phase, feature extraction is completed using proposed convolutional neural network. But classification of emotion is decided after matching of extracted features with trained features (Fig. 6).

6 Convolutional Neural Network

Convolutional neural networks have the most influential innovations in the field of computer vision. It is biologically inspired from visual cortex and imitates the working of human brain for visual analysis.

All the networks described here are programmed using Keras, a deep learning python library on Tensorflow in the backend. This facilitated faster and easier experimentation. ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. The image is fed into the network and then the network analyses the features of the image. A brief description of the layers used in ConvNet is:

Input Layer

- It has the raw pixel values of the image as ($w \times h \times c$) where w and h are the width and height of the image and c is the number of colour channels. In our case, it is ($48 \times 48 \times 1$) where 1 is for the gray scale images.
- Since the dimensions are fixed, pre-processing needs to be done before feeding the pixels in the input layer.

Convolutional Layer

- This layer computes the dot product between the weights and a small region to which the neurons are connected to in the input layer. The number of filters is passed as one of the hyper parameters which are unique with randomly generated weights. The filter also called a kernel, is convolved with image (i.e. element wise multiplications between filter values and the input pixel values).
- This generates a feature map that acts as feature identifiers sensitive to the edges and the orientations that represent how the pixel values are enhanced. This results in ($w \times h \times f$), where f is the number of filters used.
- Convolutional layers are followed by a pooling layer that down samples the dimensions along the width and the height to reduce the computational time due to a large number of convolutional layers. MaxPooling is used that reduces the dimensions of the map by a factor of window size and only the maximum pixel value in the original feature map window is retained.

Dense Layer (Fully Connected Layer)

- This layer is fully connected with the output of the previous layer. These are typically used in the last stages of the CNN to connect to the output layer and construct the desired number of outputs. It transforms the features through layers connected with trainable weights.

- This layer identifies the sophisticated features in the image that brings out the entire image.
- Sometimes it becomes prone to over fitting. This is reduced by adding a dropout layer that randomly selects a portion (usually less than 50%) of nodes to set their weights to zero during training.

Output Layer

- This layer is connected to the previous fully connected layer and outputs the required classes or their probabilities.
- Since, in human some emotions are generally an amalgamation of emotions which is computed by probability of each emotion. This is achieved by using softmax layer in the network.

7 Various Attempted Networks, Their Comparisons and the Selection of Proposed Network

The following models have been tried and cross validation is done for the model selection. The results of cross validation are given in Table 1. A pool size of (2, 2), kernel size of (3, 3) and (5, 5) are used. L2 regularization, dropout of 0.3, batch normalization and 'Uniform Kernel Initializer' has been used for more accurate results. The models are trained for 60 epochs.

(A) The first architecture we used is inspired from Lenet architecture by Yann LeCun [13]. Lenet is a small network consisting of 2 convolutional layers followed by a dense layer. We modified it by adding an extra dense layer to it. Thus, the network comprises of 2 convolutional layers followed by a MaxPooling layer. 2 dense layers with number of filters 200 and 100 follow. Last layer is a softmax layer that gives the probability of different classes. The hyperparameters are tested and chosen such that the performance metrics are maximised.
(B) The above network is modified by replicating the convolutional layers and the MaxPooling layers to identify the finer edges and patterns more specifically. Thus, this network consists of 2 convolutional layers followed by a MaxPooling layer which is followed by the similar pattern of 2 convolutional layers and MaxPooling layer twice. The dense layers of 64 and 32 filters are subsequently added followed by the final softmax layer. The number of filters in dense layers is reduced to compensate the increased computation time due to the addition of convolutional layers.
(C) Convolutional layers are now added into the second network while the dense layer is kept as before. This is done to check the performance of the addition of the convolutional layers in our model. Hence, this network consists of three blocks of 3 convolutional layers and a MaxPooling layer followed by the 2-dense layer and the final softmax layer. Here, we see that the addition of convolutional layers has a positive impact on the performance metric.
(D) To study the impact of dense layers, we made a network consisting of convolutional layer followed by a MaxPooling layer which is again followed by a

convolutional and a MaxPooling layer followed by a convolutional layer. Dense layer with 3072 filters is then added followed by the output layer (softmax). We observe that increasing the number of filters in the dense layer has a positive influence on the accuracy.
(E) To improve the performance of the previous architecture, we added convolutional layers. The network consists of 2 blocks of 2 convolutional layers and a MaxPooling Layer followed by a convolutional layer and a dense layer with 3072 filters. This is our final model. There is still a scope for improvement. More combinations can be tried, and a proper grid search can be performed with different parameters, which have great computational overhead, but will give much better recognitions.

The emotion recognition accuracy using SVM and all the attempted CNN networks (i.e. A, B, C, D and E) have been depicted in Table 1.

Table 1. Attempted network accuracy

Network	SVM	A	B	C	D	E
Accuracy (%)	46.74	58	50	58	62.32	**63.03**

8 Proposed Network Architecture

Amongst all attempted CNN networks, we have got maximum accuracy with Network E (Accuracy in Table 1). The algorithm steps are described below. The architecture of Network E has been shown in Fig. 7. The proposed architecture and its working approach to determine the emotion from a facial gesture involves following steps:

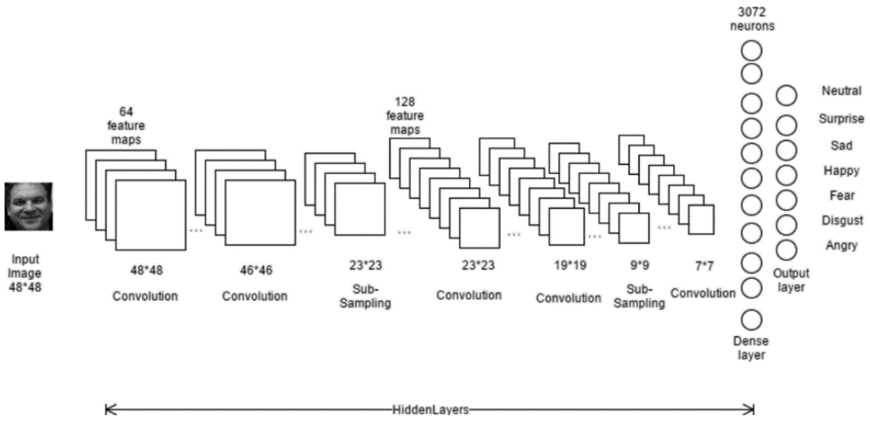

Fig. 7. Proposed network architecture (i.e. Network 'E')

Algorithm 1. Emotion Recognition using Facial Gestures

Input: An Image
Output: Emotion of the faces in image.
Prerequisites: A large dataset with different facial emotions.
1. The dataset is split into training and test sets (70%-30%).
2. Viola Jones algorithm is applied on all the images in training and test sets for pre-processing.
3. A simple model is prepared based on some existing model.
4. Model is then trained on the training set and tested for accuracy on test set.
5. Grid search is applied for hyper parameter tuning.
6. Dense and convolution layers are added, and the model is tested for performance metrics.
7. In case of over fitting, cross validation is done.
8. Several models are tried with appropriate addition of convolution and dense layers, for better performance.
9. The best model is then taken as the final model.

9 Results

Our proposed model yielded a promising accuracy of 63.03% which is considerably good with less training data. The results for 5 test images are shown in Table 2. The two topmost emotion percentages are highlighted in bold.

Table 2. Emotions results in percentage

Test Image	1	2	3	4	5
Emotions					
Anger	0.5	3.4	2.3	5.75	2.6
Disgust	2.3	2.6	1.2	**31.95**	2.2
Fear	0.2	9.4	7.1	4.93	8.84
Happy	**93.9**	0.9	2.3	1.3	**33.48**
Sad	0.7	**35.7**	14.3	10.1	9.45
Surprise	1.7	3.2	6.1	2.12	4.8
Neutral	0.45	**44.5**	**66.5**	4.63	**38.56**

We see that our model has correctly identified the emotions in Test Images 1, 3 and 4 in the above table as the highest emotion percent (highlighted in bold) is the true emotion of the face. The accuracy obtained in this way by considering the highest emotion percent as the result in every face and then comparing it with its true emotion

was 63.03%. It was also observed that incorporating the second-best emotion, we can achieve emotions with an accuracy of 67%. For Test Images 2 and 5, we see that the second highest emotion percent (highlighted in bold) is the true emotion of the test face rather than the highest emotion percent. This is analogous to how a human perceives an emotion. It is sometimes difficult to decipher the emotion from a face. As in test image 2, the child appears to be sad to some and neutral to others. While in test image 5, we confuse between happy or neutral.

10 Conclusion

Emotion recognition is still a difficult and a complex problem in computer science because every expression is a mix of emotions. This work tries to address the problem of emotion recognition with deep learning approach using convolutional neural network. First, we have implemented SVM and then attempted 5 different CNN networks (namely A, B, C, D and E) and tested the accuracy. Network E gives the maximum accuracy among all including SVM too. Training and testing of these networks have been performed on FER-2013 database. The system is independent of factors like gender, age, ethnic group, beard, backgrounds and birthmarks. The proposed system is very promising and provides better accuracy in emotion recognition than SVM. The proposed architecture and the algorithm here yield noticeable results. Hence it can motivate the researchers to design the better 'Deep Learning CNN Architecture' to enhance the emotion recognition system.

References

1. Carton, J.S., Kessler, E.A., Pape, C.L.: Nonverbal decoding skills and relationship well-being in adults. J. Nonverbal Behav. **23**(1), 91–100 (1999)
2. Izard, C.E.: Human Emotions. Springer, New York (2013)
3. Happy, S.L., George, A., Routray, A.: A real time facial expression classification system using Local Binary Patterns. In: Intelligent Human Computer Interaction (IHCI), 4th International Conference, pp. 1–5. IEEE (2012)
4. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9), 1162–1171 (2011)
5. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with Convolutional Neural Networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
6. Hu, T., De Silva, L.C., Sengupta, K.: A hybrid approach of NN and HMM for facial emotion classification. Pattern Recogn. Lett. **23**(11), 1303–1310 (2002)
7. Sebe, N., Cohen, I., Gevers, T., Huang, T.S.: Emotion recognition based on joint visual and audio cues. In: 18th International Conference on Pattern Recognition, ICPR, vol. 1, pp. 1136–1139. IEEE, August 2006
8. Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., Chen, X.: Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th ACM International Conference on Multimodal Interaction, pp. 494–501 (2014)

9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
10. FERC 2013: Form 714 – Annual Electric Balancing Authority Area and Planning Area Report (Part 3 Schedule 2) 2006–2012 Form 714 Database, Federal Energy Regulatory Commission (2013)
11. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
12. Weston, J., Watkins, C.: Multi-class support vector machines. Technical report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London (1998)
13. LeCun, Y.: LeNet-5, Convolutional Neural Networks (2015). http://yann.lecun.Com/exdb/lenet

http://www.springer.com/978-3-319-71927-6

Mining Intelligence and Knowledge Exploration
5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings
Ghosh, A.; Pal, R.; Prasath, R. (Eds.)
2017, XX, 438 p. 125 illus., Softcover
ISBN: 978-3-319-71927-6