

Facial Expression Recognition Using Deep Learning

Advisor -

Dr. Snehasis Mukherjee

By -

Kalyan Babu

Kishan Kumar

Ashutosh Mishra

Shyam Nandan Rai

Abstract :

Facial expression recognition is one of the major challenging task in the field of computer vision. With the Introduction of convolutional neural network in the field of computer vision , the accuracy of performing facial expression recognition have improved significantly . In this project we have used FER-2013 dataset [1] for the checking model accuracies , which as result distinguishes between seven emotions present in the database. In experimentation part, we used pre trained model VGG Face[2] for extracting features and used SVM with different kernels . In addition , to this we used a unified model which fuses the CNN features and HOG feature which gave the highest accuracy than other models.

Introduction:

Human Facial Expression is another most important part of communication other than speech . This has a wide application in the field of computer vision and Human Computer Interaction . There are usually six types of basic expression which are widely accepted i.e anger , fear ,happiness , sadness and surprise [figure 1].

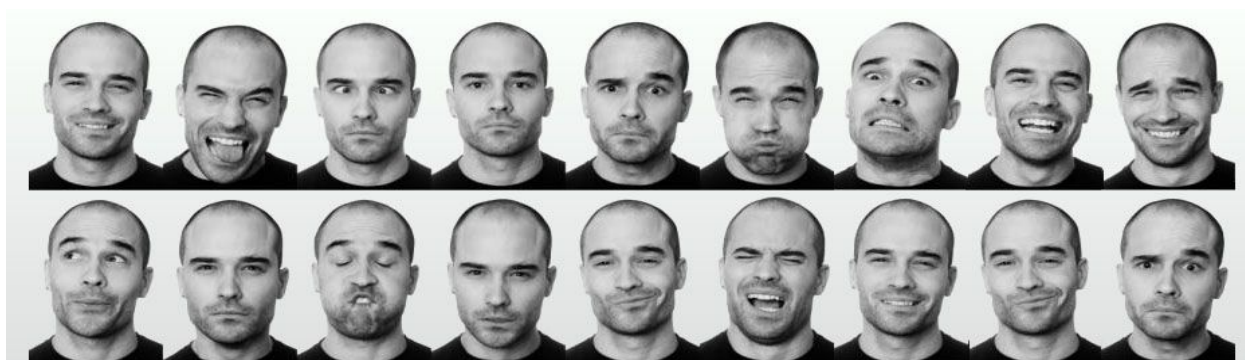



Figure 1. Images depicting anger, disgust, fear, happiness, sadness, surprise, as well as neutral.



But , the problem becomes hard when there is variation in the head pose , illumination , occlusion , and unposed expressions . With the success of Convolutional Neural Network (CNN) in Image Classification [3] , Recognition [4-5] leads us in solving various issues in Facial Expression Recognition (FER) [10]as these issues persists in hand crafted method LBP, SIFT Gabor filters which will be shown in experimentation section.

In this paper we have initially used Hand Crafted Methods and shown that their performance in FER task is poorer than the Deep Neural Methods. We have also used SVM and KNN along with various Kernels like - polynomial , RBM for classification. At Last, We used Pre Trained CNN Model and fused them with the HOG features which gave the highest accuracy among all methods.

State Of the Art

In the past few years vision researchers have developed various FER models [6-7] among all the models The best known psychological framework for describing nearly the entirety of facial movements is the Facial Action Coding System (FACS) [8]. FACS is a FER system to classify human facial movements by their appearance on the face using Action Units (AU). An AU is one of 46 atomic elements of visible facial movement or its associated deformation; an expression typically results from the accumulation of several AUs .

In FER -2013 dataset the highest accuracy[9] method used the primal objective of an SVM as the loss function for training. This loss function has been applied to neural networks before, but he additionally used the L2-SVM loss function, a new development that gave great results on the contest dataset and others.

Dataset:

In this project we used FER-2013 dataset, which consists of about 37,000 well structured 48×48 pixel gray-scale images of faces. The images are processed in such a way that the faces are almost centered and each face occupies about the same amount of space in each image. Each image has to be categorized into one of the seven classes that express different facial emotions. These facial emotions have been categorized as: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral. Figure 1 depicts one example for each facial expression category. In addition to the image class number (a number between 0 and 6), the given images are divided into three different sets which are training, validation, and test sets. There are about 29,000 training images, 4,000 validation images, and 4,000 images for testing.

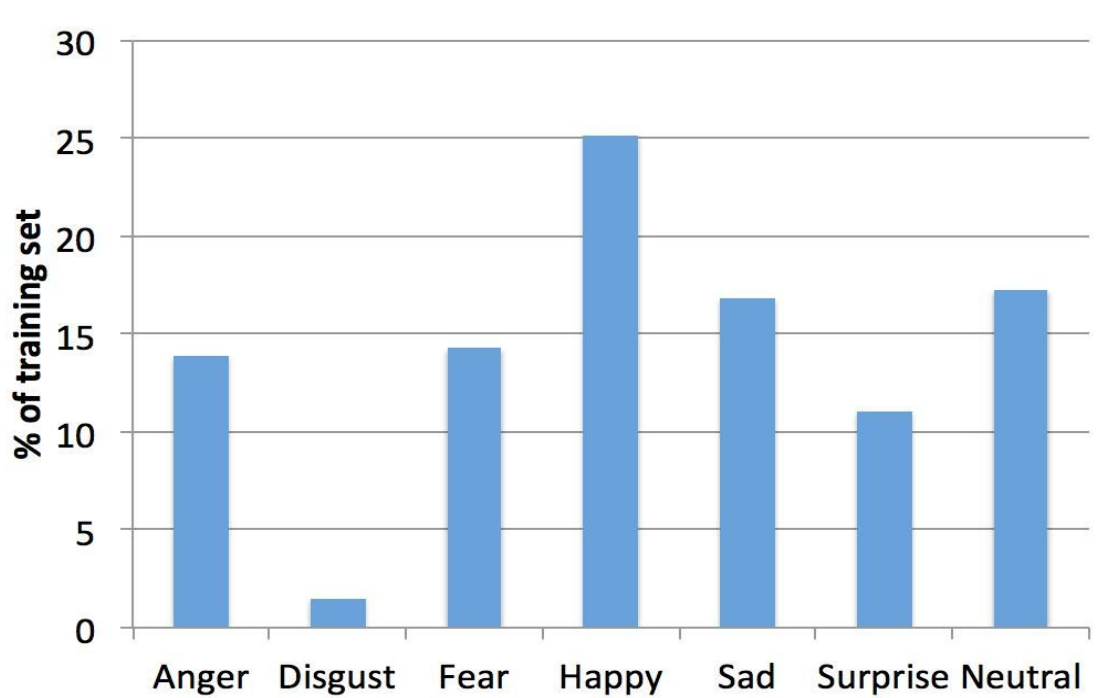


Figure 2: Distribution of different emotions across the FER-2013 dataset

Experimentation & Results

Hand Crafted Methods

Local Binary Patterns

In this method from each image we extract the image in the form local binary histograms . The LBP codes of an image were computed using 8 sampling points defined in (8×2) matrix on a circle of radius 3. But , Local Binary Pattern gave the least accuracy in FER database when used with KNN and SVM (polynomial kernel) classifiers .

Dense Scale Invariant Feature Transform (DSIFT)

DSIFT can quickly compute descriptors for densely sampled keypoints with identical size and orientation. It can be reused for multiple images of the same size. While experimentation, the size of a SIFT spatial bin is 8 and magnification of 3 . The DSIFT feature almost gave double the accuracy than LBP , but they did not perform well on SVM classifier.

Histogram of oriented gradients

This feature gave the best accuracy among the all hand crafted methods .The HOG features was calculated for the whole image using cell sizes 8 in pixels, block sizes 8 in cells, and number of gradient bins 8. This leads to a 64 dimensional feature vector for each cell. The HOG features dimensions were reduced using PCA and as a result the reduced feature complemented the SVM classifier giving higher accuracy than HOG alone.

DEEP CONVOLUTION AND AGGREGATED METHODS

VGG FACE (Pre Trained Model)

It comprises 11 blocks, each containing a linear operator followed by one or more non-linearities such as ReLU and max pooling. The first eight such blocks are said to be convolutional as the linear operator is a bank of linear filters (linear convolution). The last three blocks are instead called Fully Connected (FC); they are the same as a convolutional layer, but the size of the filters matches the size of the input data, such that each filter "senses" data from the entire image. All the convolutional layers are followed by a rectification layer (ReLU) but they do not include the Local Response Normalisation operator.

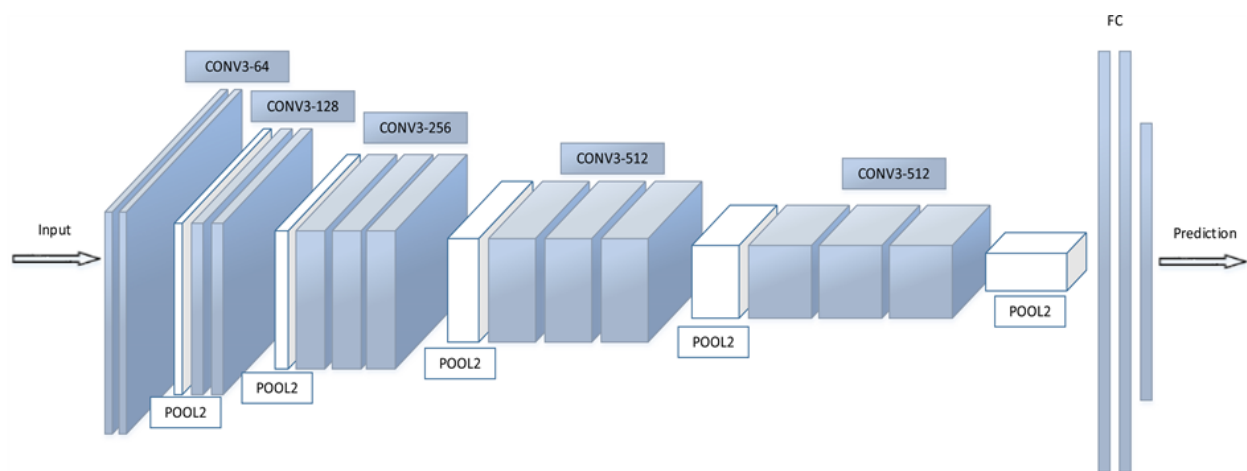


Figure : Visualization of the VGG Face Network

In Our experimentation, we removed the last fully connected layer and softmax layer . Then , we passed the images through the modified network to get 4k features . Now ,these features we applied L2 normalization before going through KNN classifier . Due large dimension of the output features the SVM did not gave much higher accuracy than KNN. So, We applied PCA on the 4k features to reduce the chances of overfitting . As a result , using SVM on reduced dimension feature gave high accuracy.

VGG FACE Fine Tuning

In this method , we removed the last two fully connected layers of VGG face and added fully connected layers along with the dropout layer , relu layers and softmax layer that classifies the facial expression into seven categories .

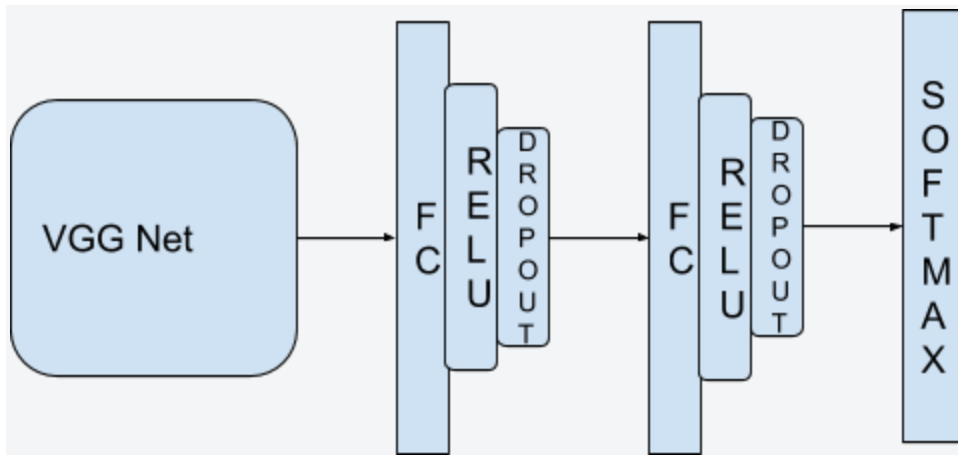


Figure 3: The figure shows the VGG face model fine tuned for FER Database

During training the network , the modified VGG learning rate was kept zero and the rest of network had a high learning rate of order $10e-3$ and learning rate decay of 0.95 . The input to the network was given into batches of 32 images and the whole network ran for 40 epochs.

Aggregated VGG FACE Fine Tuning

As ,In the Hand Crafted features the HOG features complemented the SVM classifiers and relatively performed well on KNN classifier. So, it boils down to be best feature suited for aggregation with VGG face model.

Firstly , HOG features were extracted for train and test images , then the feature dimension were reduced using PCA to prevent overfitting . Now , we create a parallel neural network consisting of two input nodes where on one node we put the raw image through the modified VGG Face and on another we feed the extracted HOG features . Then parallely these features were passed through fully connected layer and a softmax layer. Then , these feature are aggregated and passed through fully connected layer and at last , passed through

softmax classifier . In this learning rate of the VGG net is kept zero and the rest of network learns at the rate of $10e-3$.

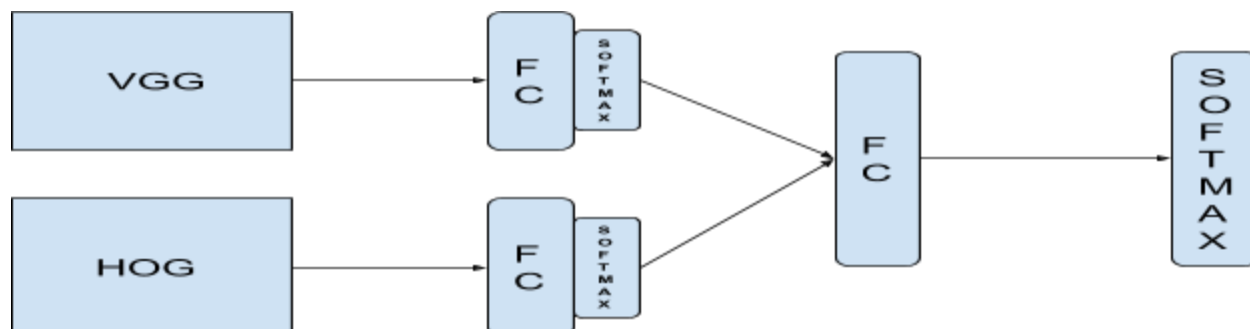
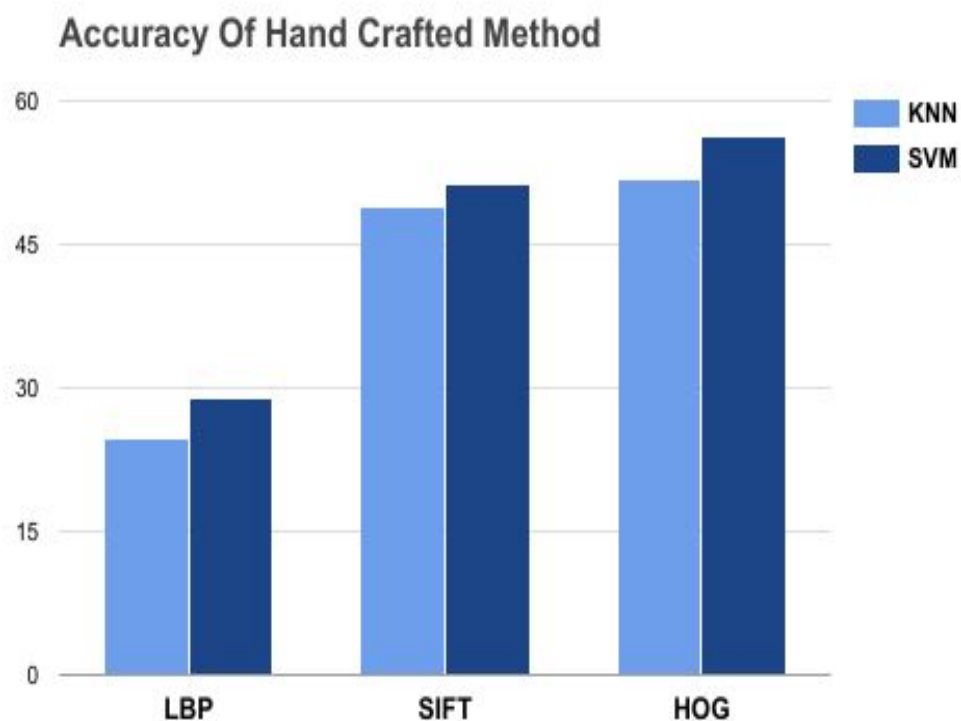
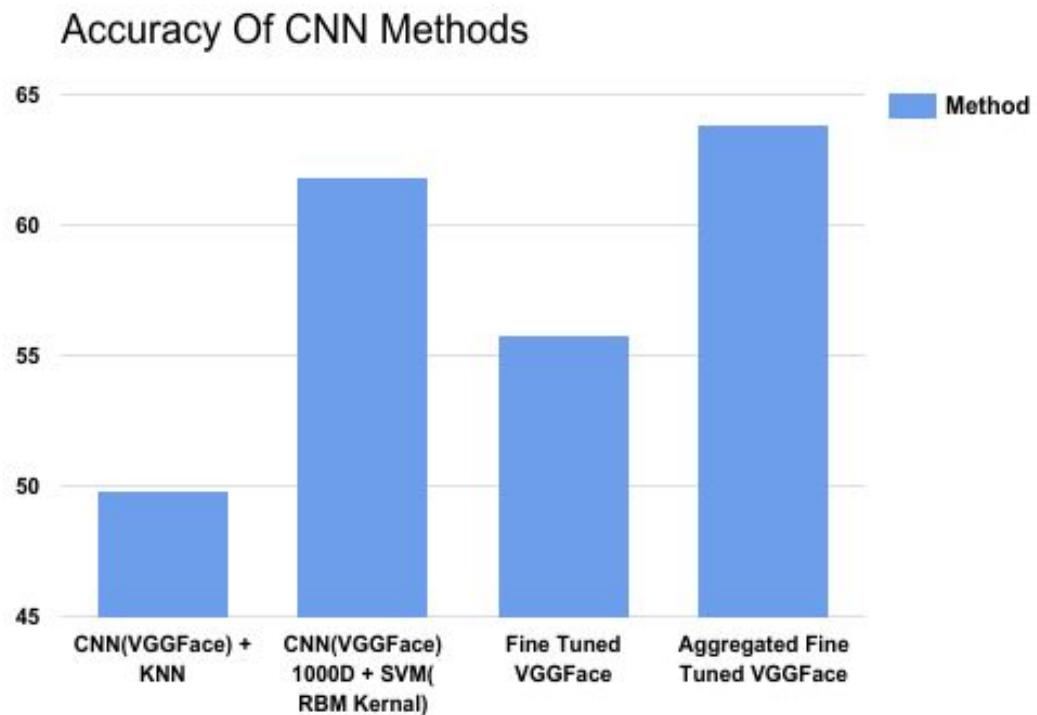


Figure 4 : Aggregated VGG Face Fine Tuned Network model

In the results we see that the deep methods are clear winner over hand crafted methods . Though HOG features with SVM give comparable accuracy than CNN Methods.





Conclusion

In this project, we aimed to classify facial images into seven emotional categories. We experimented with various hand-crafted techniques and various CNN techniques, such as fine-tuning and fractional max-pooling which helped in achieving the highest accuracy of ~64% on FER dataset. The results demonstrated that deep CNNs are capable of learning facial characteristics and improving facial emotion detection. Also, the hybrid feature sets did help in improving the model accuracy. Given more time, we would have liked to combat overfitting and approach state-of-the-art accuracies.

References:

- [1] Goodfellow, Ian J., et al. "Challenges in representation learning: A report on three machine learning contests." *International Conference on Neural Information Processing*. Springer Berlin Heidelberg, 2013.
- [2] Parkhi, O. M. and Vedaldi, A. and Zisserman, A., Deep Face Recognition, British Machine Vision Conference, 2015.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.
- [4] K. He, X. Zhang, shaoqing Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. 1512, 2015.
- [5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015
- [6] Nicu Sebe, Michael S. Lew, Ira Cohen, Yafei Sun, Theo Gevers, Thomas S. Huang (2007) Authentic Facial Expression Analysis. *Image and Vision Computing* 25.12: 1856-1863
- [7] 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.
- [8] P. Ekman, W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, 1978
- [9] Yichuan Tang. Deep learning using linear support vector machines. *Workshop on Challenges in Representation Learning, ICML*, 2013.
- [10] <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data>