Facial Emotion Detection Using Different CNN Architectures: Hybrid Vehicle Driving

by

Garimella Ramamurthy

Report No: IIIT/TR/2017/-1

Centre for Communications
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2017

Facial Emotion Detection using Different CNN Architectures: Hybrid Vehicle Driving

Siva Prasad Raju Bairaju
IIIT RKValley, RGUKT-AP
Email: sraju728@gmail.com

Sowmya A
IIIT RKValley, RGUKT-AP
Email: aarivalli23@gmail.com

Dr. Rama Murthy Garimella
IIIT Hyderabad,Gachibowli
Email: rammurthy@iiit.ac.in

***Abstract*—In this research paper, we train convolutional neural network(CNN) to be able to classify facial emotions/expressions. Using JAFFE(Japanese Female Facial Expressions) database of facial emotion images, we trained a CNN and are able to achieve good accuracy during training phase. We proposed the concept of hybrid vehicle employing a CNN for detecting drowsiness or alertness of the driver. We propose to be able to perform drowsiness detection in real-time.**

***Index Terms*—Convolutional Neural Networks, Data Augmentation, Drowsiness, Hybrid Vehicle, Emotion Detection**

I. Introduction

Facial emotions are important aspects in human communication that help us to understand the intentions of others. Facial expressions convey Non-Verbal Cues which play an important role to maintain interpersonal relations. According to different surveys verbal components(Speech) convey one-third of human communication and Non-Verbal components(Facial emotions, Gestures) convey two-third of human communication. Facial emotion detection became a well attempted research topic now a days due to its prospective accomplishments in many domains such as Medical engineering, Vehicles, Robotics and Forensic applications etc. For example,a robot could be developed to serve bed-ridden and disable people who can communicate through facial expressions . Humans Recognize facial emotions accurately without delay but for a machine it is a challenge.

II. Novel Application: Driver based and Driverless Vehicle Navigation

In this research paper we consider Drowsiness of a human being as an emotion.Thus in this section, we consider an application in which certain type of emotion recognition naturally arises. The application deals with detecting the "alertness" of a driver when navigating the vehicle. We can consider a camera capturing the facial expression of a driver. The emotion to be captured deals with being able to determine if the driver is DROWSY. Thus, the classification problem to be solved by a CNN could be BINARY i.e Drowsy, Non-Drowsy or TRINARY: Non-Drowsy, Partially Drowsy, totally drowsy and so on. Using a suitable database of such facial images, a CNN can be trained and the classification accuracy can be determined.

The vehicle navigation is switched to DRIVERLESS mode if the driver is classified to be DROWSY. For instance in the night, if the driver is detected to be drowsy and passing on a highway, the vehicle is switched to a driverless mode. After few hours if the daylight comes out, the vehicle is switched to the mode in which driver takes over control. Also, in the trinary classification of drowsiness by a CNN, an alarm can be given when the driver changes from partially drowsy to totally drowsy mode. Further the alarm can be linked to a Control Unit which brings the vehicle to a safe stopping point(with stop lights on) along side of the road. We thus introduced the following concepts.

- Hybrid vehicle(with respect to driver): A vehicle which is not fully driver less and which is not fully driver based is called a HYBRID VEHICLE.

A. New Ideas

1. Feature Extraction enabling Drowsiness detection or more finely Drowsiness/alertness degree determination.

- MLP based classification.

2. Comparing Computation time and classification accuracy with various interesting architectures i.e.

- Ordinary MLP: Emotion recognition time and accuracy.
- Convolutional Neural Network: Emotion recognition time and accuracy.

B. Originality of Contribution

Preprocessing of images by means of NOVEL FEATURES to detect Degree of Drowsiness.

For example

- Eyes are closing (Indicating that the person is tired) i.e feeling sleepy or Degree of Closure.
- The person is beginning to Yawn with open mouth.

III. Related work

Recently, researchers have made considerable advancement in human facial expression recognition with Artificial Intelligence and Computer vision techniques. In twentieth century, research on facial expressions has began. In early 1970s, Ekman and Friesen, American Psycologists did an extraordinary work on facial emotions and they entrenched six universal facial expression: Happy,Sad,Surprise,Angry,Disgust,Neutral. And they implemented Facial Action Coding Systems(FACS) which was further used to categorize human facial movements by their appearance with the help of Action Units(AU). From

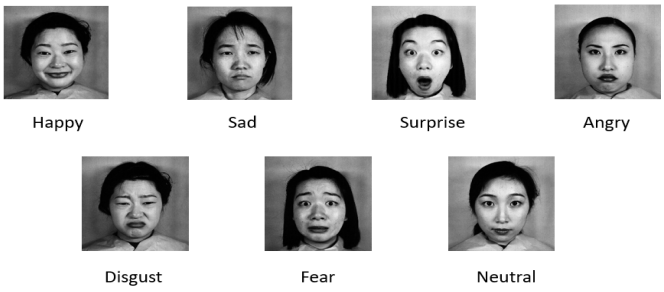


Fig. 1. Seven basic facial expressions in JAFFE Dataset

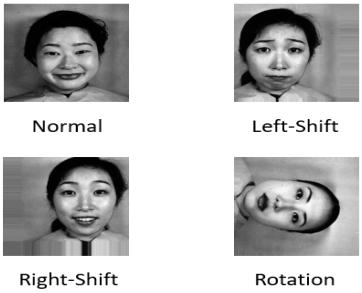


Fig. 2. Images after Data Augmentation

this a new Facial Emotion Recognition era has began. In 2003, Ira Cohen and Nicu Sebe et al presented an architecture of Hidden Markov models(HMMs) for classifying expressions from video. Shan et al proposed a method for emotion detection using Boosted LBP(Linear Binary Patterns) descriptors in 2009. In Later research Pyramid Histograms Of Gradients(PHOG) are also used for Emotion Detection. In present days, Deep learning architectures like Convolutional Neural Networks and Auto Encoders are used for feature extraction from an image. Firstly Liu et al used 3D-CNN and a deformable facial action part model to locate facial action parts and learn part-based features for emotion categorization. In the year of 2016, Ali et al, proposed a model which is a collection of boosted neural networks for multiethenic facial emotion recognition. The results of any Deep Learning architectures mainly dependent on how the preprocessing was done, appropriate Feature selection by the model and amount of data provided to train the network.

IV. Dataset Preparation

In this paper, we use JAFFE(Japanese Female Facial Expressions) database to train our network. This database contains 213 gray images of 7 facial expressions posed by 10 Japanese female models.And these 213 images are static images with 256 X 256 pixels. After doing lot of experiments with Deep learning architectures the primary thing that everyone realize is the data which is used during training plays the most important role. To achieve best classification accuracy with Deep learning architectures network should be trained with large amount of dataset. To overcome problem of limited quantity and limited diversity of data, we generate our own data with the existing data. This methodology is known as Data Augmentation. For Data Augmentation some in-built packages are available in different Frameworks. In this paper we are using Keras Framework which has ImageDataGenerator to do Data Augmentation and our models are implemented using TensorFlow software library.

Some of the Data Augmentation techniques are: Scaling, Translation, Rotation(at 90 degrees), Rotation(at finer angles), Flipping, adding Salt and Pepper Noise, Lighting condition and Perspective transform.

V. Proposed Architectures

In this paper, we proposed two CNN architectures(Fig.3 and Fig.4) which are trained using JAFFE database and Finally we compare the accuracy of results before performing Data Augmentation and after performing Data Augmentation. In Fig.3 we built a architecture with 3-convolutional layers, 3-max pooling layers, 2-Fully connected layers and Softmax function as output layer. In Fig.2 CNN architecture includes 5-convolutional layers, 5-max pooling layers, 2-Fully connected layers and Softmax function as output layer. In both the architectures 10 filters are taken in each convolutional layer with 3x3 pixels and we use max pooling with a pool width of two and a stride between pools of two. In both the architectures first fully connected layer has 256 neurons whereas second fully connected layer has 128 neurons. The fully connected layers contain dropout, a mechanism which reduces the risk of the network overfitting and The Rectified Linear Unit(ReLU) was used as Activation function. Generally the system operates in two categories: Training and Testing.

A. Convolutional layer

The purpose of the convolutional layer is to extract features from the input data. It learns image features using small squares of input image and creates a feature map by maintaining spatial relation between pixels. After giving input to the convolutional layer, convolution is performed between input and the features learned by the network. Convolution is a linear process which is element wise matrix multiplication and addition. And output convolutional layer has same resolution as input. Since JAFFE database images are with 256*256 pixels, convolutional layer output is also have same pixels. If h_k is a filter with kernel size axb and is supposed to convolute with image x, output of this can be calculated as:

$$C([x_{u,v}]) = \sum_{i=-a/2}^{a/2} \sum_{j=-b/2}^{b/2} x(i,j).h(u-i, v-j)$$

B. Rectified Linear Unit

Each convolutional layer output(Feature map) is passed through ReLU layer. ReLU is a Non-linear operation used to normalize the output of convolutional layer. It is applied per pixel and replaces all negative pixel values in the feature map

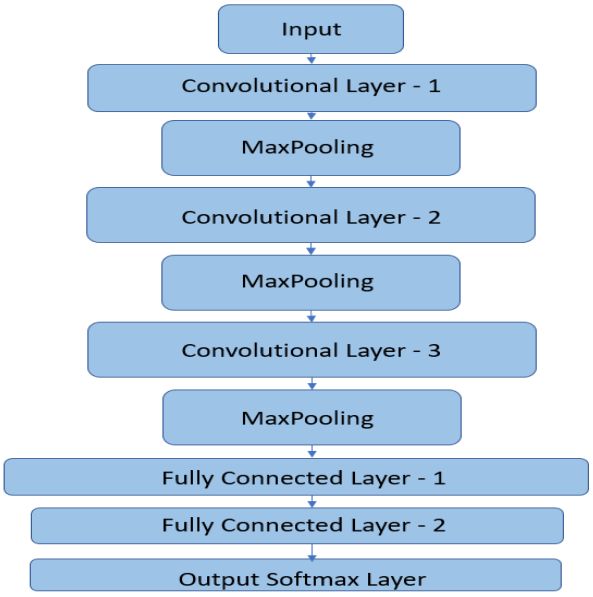


Fig. 3.

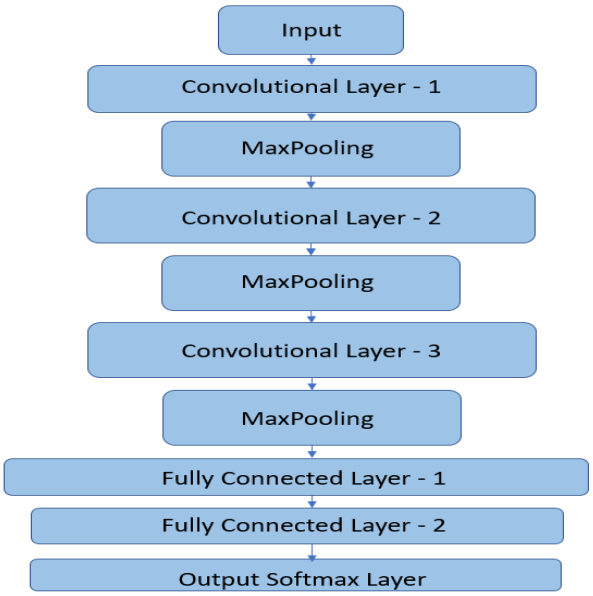


Fig. 4.

by Zero. Output feature map of ReLU layer also have same resolution as input image and ReLU is:

$$R(x) = max(0, x)$$

C. Pooling or Subsampling

Pooling reduces the dimensionality of each feature map but retains the most wanted information. In our architecture we use Max pooling in which the largest pixel value from the rectified feature map within the selected window is taken. Here we stack all the output feature maps of pooling layer and give as input to Fully connected layers.

D. Fully connected layer

The output of the final pooling layer acts as an input to the fully connected layer. Conventionally Fully connected layer resembles a multi layer perceptron. In our architecture we have taken two fully connected layers one with 256 input neurons and another with 128 input neurons. Fully connected layer uses SoftMax as the final classification layer to predict the given input category. Output of fully connected layer with 'l' no of neurons with input x will be as follows:

$$F(x) = A(\sum_{i=1}^{l} W.x)$$

Where 'A' is a Activation function,'W' is a Weight matrix.

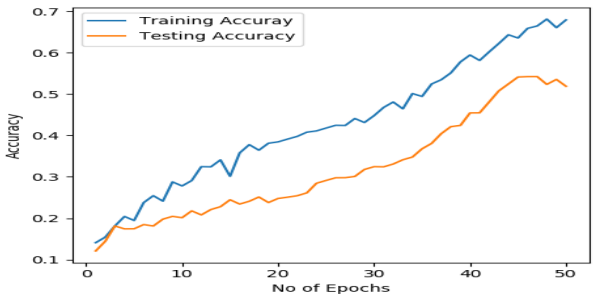
The output of the Convolutional layer and Pooling layer constitute features of the input image. The requirement of the fully connected layer in a network is to utilize these features for classifying the input image into several categories based on the training dataset. After all this analysis we conclude that Convolutional layer and Pooling layer acts as Feature Extractor from the input image while Fully connected layer acts as classifier.

VI. Experimental Results and Discussions

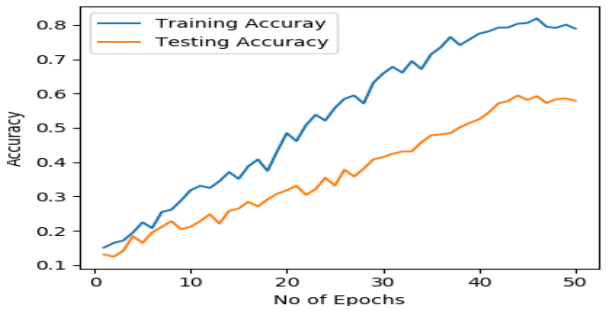
We implemented two Convolutional Neural Network Architectures with JAFFE dataset with Data Augmentation and Without Data Augmentation. And the models were trained for 50 epochs with learning rate of 0.001. We applied Batch Normalization(Mean=0,Standard deviation=1) to the Dataset to get best classification results.

First architecture in Fig-1 is a shallow network and It is trained with JAFFE dataset before and after Augmentation. Results are tabulated in Table-1, Graphs are drawn between Epochs vs Accuracy with Before Augmentation[Fig-5] Results and After Augmentation[Fig-6]results.

Table - 1	Without Augmentation	With Augmentation
Dataset	213	2000
Training Data	142	1472
Testing Data	71	528
No of Epochs	50	50
Accuracy	54 %	59 %



$Fig - 5$

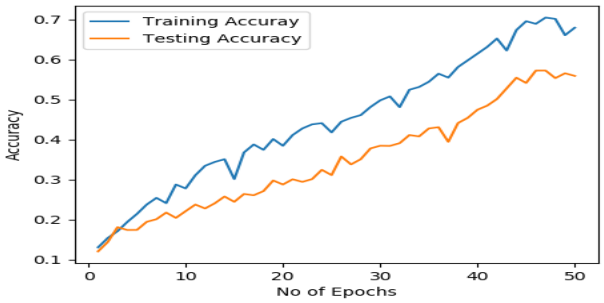


$Fig - 6$

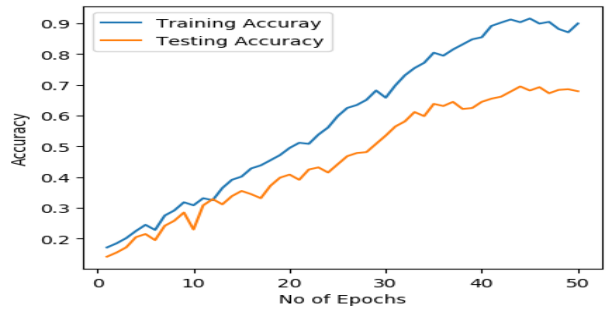
From these results we can conclude that Classification accuracy mainly dependent on amount of training data. Here we may also conclude that network trained with large number of Epochs can better classify the data.

Second architecture in Fig-4 is a Deep CNN architecture and it also trained with JAFFE database without Data Augmentation[Fig-7] and with Data Augmentation[Fig-8]. And got Results as below.

Table - 2	Without Augmentation	With Augmentation
Dataset	213	2000
Training Data	142	1472
Testing Data	71	528
No of Epochs	50	50
Accuracy	57 %	68 %



$Fig - 7$



$Fig - 8$

From Second Architecture we can conclude that Classification Accuracy also depends on how deep the Networks are designed i.e in Fig-3 we have taken 3-Convolutional layers and 3-Pooling layers while in Fig-4 contains 5-Convolutional layers and 5-Pooling layers. When we observe, the results of both architectures are comparatively different i.e Deep Network got more classification accuracy than the Shallow Network.

In this paper we trained our CNN architectures for 50 epochs and we got better accuracy results than the previous works for JAFFE database.Then we increased the number of epochs which is giving impressive accuracy results.

VII. Conclusion and Future Work

Any research work acquire its significance only when it is used in a real-time applications. In this paper we proposed an application called 'Hybrid Vehicle Driving' which can be developed using Convolutional Neural Networks to detect driver drowsiness i.e. discussed in Section-II. In Section-V we proposed two new CNN architectures which are trained by JAFFE database before applying Data Augmentation and after applying Data Augmentation technique. From the acquired results, For two architectures accuracy vs epochs graphs are drawn and we made following conclusions:

- Deep Learning architectures accomplish considerable accuracy when it is trained with large amount of data and this is proved in our paper through Data Augmentation technique.
- Deep neural networks get more accuracy than the Shallow neural networks.
- a Network achieve better training accuracy if it is trained with larger number of epochs.

In future we want to develop a Hybrid Vehicle which can be controlled by detecting driver drowsiness through our proposed Deep Learning architectures. By detecting the drowsiness of a driver we alert him through alarm which comes under ACTIVE SAFETY in ADAS(Advanced Driver-Assistance Systems) applications.

References

[1] Shima Alizedeh, Azar Fazel, *Convolutional Neural Networks for facial expression recognition*, cs231n.stanford.edu/reports/2016.

[2] Dan Duncan, Gautam Shine, Chris English *Facial Emotion recognition in Real-time*, cs231n.stanford.edu/reports/2016.

[3] Siyue Xie and Haifeng Hu *Facial Expression recogntion with FRR-CNN*, ELECTRONICS LETTERS 16th February, Vol.53, No.4, pp. 235-237.

[4] Julio Cesar Batista, Vitor Albiero, Olga R.P. Bellon and Luciano Silva *AUMPNet: simultaneous Action Units detection and intensity estimation on multipose facial images using a single convolutional neural network*, 2017 IEEE 12th International Conference on Automatic Face Gesture Recognition.

[5] Ariel Ruiz-Garcia, Mark Elshaw, Abdulrahman Altahhan, Vasile Palade *Stacked deep convolutional Autoencoders for emotion recognition from facial expressions*, 2017 International Joint Conference on Neural Networks(IJCNN).

[6] Yize Liu and Yixiang Chen *Recognition of facial expression based on CNN-CBP features*, 2017 29th Chinese Control and Decision Conference (CCDC).

[7] Xiaoguang Chen, Xuan Yang, Maosen Wang and Jiancheng Zou *Convolution Neural Network for Automatic Facial Expression Recognition*, Proceedings of the 2017 IEEE international Conference on Applied system innovation.

[8] Aysegul Ucar *Deep Convolutional Neural Networks for facial expression recognition*, 2017 IEEE International Conference on Innovation in Intelligent System and Applications(INISTA).

[9] Gerard Pons and David Masip *Supervised Committee of Convolutional Neural Networks in Automated facial Expression Analysis*, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

[10] Vedat TUMEN, Omer Faruk SOYLEMEZ and Burhan ERGEN *Facial Emotion Recognition on a dataset using convolutional neural Network*, 2017 International Artificial Intelligence and Data Processing Symposium(IDAP).

[11] Pengfei Dou, Shishir K. Shah and Ioannis A. Kakadiaris *End-to-end 3D face reconstruction with deep neural networks*, 2017 IEEE Conference on Computer vision and pattern recognition.

[12] Elad Richardson, Matan Sela Roy Or-Sela and Aon Kimmel *Learning Detailed Face Reconstruction from a Single Image*, 2017 IEEE Conference on Computer vision and pattern recognition.

[13] Nishiki Katayama and Satoshi Yamane *Similarity Calculation for Verification with Convolutional Neural Network*, Proceedings of the SICE Annual Conference 2017, September 19-22, 2017, Kanazawa University, Kanazawa, Japan.