Facial Emotion Analysis using Deep Convolution Neural Network

Rajesh Kumar G A[1], Ravi Kant Kumar[2], Goutam Sanyal[3]
Department of Computer Science and Engineering
National Institute of Technology
Durgapur, India
E-mail: {rajuloki046, vit.ravikant, nitgsanyal}@gmail.com

Abstract: **Human emotions are mental states of feelings that arise spontaneously rather than through conscious effort and are accompanied by physiological changes in facial muscles which implies expressions on face. Some of critical emotions are happy, sad, anger, disgust, fear, surprise etc. Facial expressions play a key role in non-verbal communication which appears due to internal feelings of a person that reflects on the faces. In order to computer modeling of human's emotion, a plenty of research has been accomplished. But still it is far behind from human vision system. In this paper, we are providing better approach to predict human emotions (Frames by Frames) using deep Convolution Neural Network (CNN) and how emotion intensity changes on a face from low level to high level of emotion. In this algorithm, FERC-2013 database has been applied for training. The assessment through the proposed experiment confers quite good result and obtained accuracy may give encouragement to the researchers for future model of computer based emotion recognition system.**

Keywords— **Facial expressions, Facial Emotions, Non-Verbal Communication, Face Detection, Convolutional Neural Network (CNN), Deep Learning.**

I. INTRODUCTION

Emotions and related fluctuations in the facial muscles are together known as facial expressions [1]. It gives us clue about the state of a person and enables to make conversation with the other person based on their mood. Furthermore, facial expressions also support to judge the existing state of emotion and mood [2] of a person. Facial expression plays an important role in non-verbal communication between people. Diverse classification of facial expressions might be used in numerous applications like; Human Behavior Predictor [3], Surveillance System [4] and Medical Rehabilitation [5].

Seven elementary categories of human emotions [6] are unanimously predictable across different cultures and by numerous people are: anger, disgust, fear, happiness, sadness, surprise and neutral. Numerous scholars have used dissimilar methods for classifying facial expression. Identical bilateral amygdala impairment recognition of facial emotions [7], holistic template-matching to detect expression and geometric feature-based approach [8], the Active shape models: Assessment of a multi-resolution method [9], image preprocessing methods and descriptors based local binary patterns [10], Hidden Markov Model for expression detection [11], Many Hybrid approaches also has been hosted which are like view based Modular Eigen spaces and a hybrid approach of NN and HMM for facial emotion classification[12], Emotion based on joint visual and audio cues[13], Combining multiple kernel methods [14] and Convolution Neural Networks [15] etc. Robust face analysis using convolution neural networks gives the better and quick results.

The main aim of our proposed scheme is to find out the standardized parentages of several emotional states (happiness sadness, disgust, anger, surprise, and fear) in a face. The emotion having the maximum parentages is projected as its resulting emotion on a specified face. Likewise, founded on experimental outcomes, training and examination of various emotional phases (frame by frame) has also inspired us to develop a real-time facial expression recognition system. To attain such composite classification of images, an enormous and robust training is essential. Hence, in this proposed approach concept of deep learning using convolution neural network has been applied to train and test. The performance of a neural network mainly depends on numerous issues like initial random weights, activation function used, training data, and number of hidden layer and network structure of system. The convolutional neural networks use images directly as input. As a substitute of handcrafted intermediate features, convolutional neural networks are used to mechanically learn a pecking order of features which can further be used for classification.

Further, the paper is organized as: In section II, complete system architecture has been shown. The data set description are available in section III. Proposed technique algorithm is presented in section IV. In section V, we have explained complete results. Finally, section VI draws the concluding remarks.

II. SYSTEM ARCHITECTURE

Complete system architecture has been represented below. The main algorithm is divided into two parts, testing and training. First, we need to train the networks to classify the emotions of given face. The first step of our algorithm is to check whether the trained data are present or not. If not, then we need to train the system first and then we can perform testing for emotion classification.

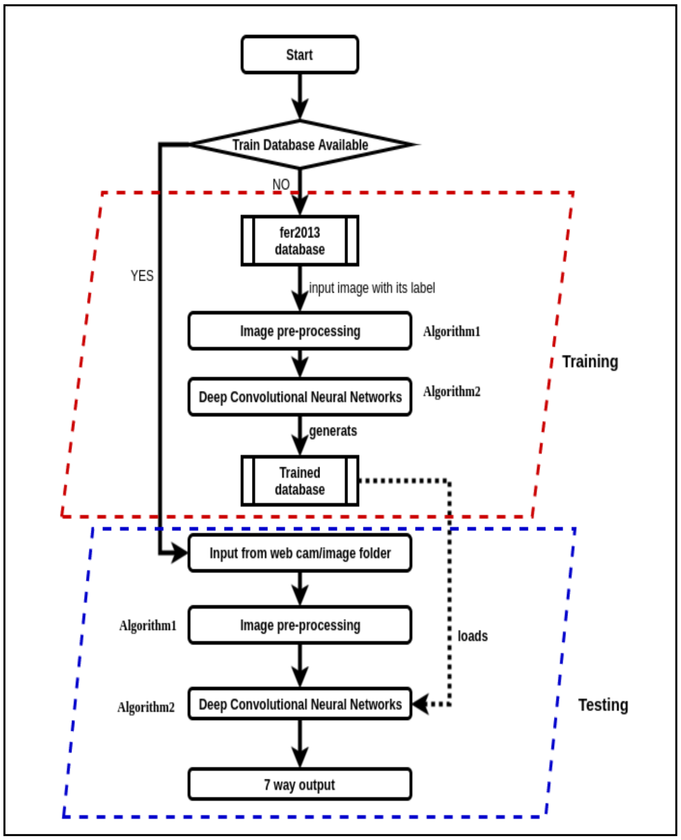


Figure 1. Complete System Workflow

III. DATASET DESCRIPTION

We are applying two databanks in this research namely FERC-2013[16] and Extended Cohn Kanade (CK+) database [17]. The datasets basically differs by image-quality, clearness and total number of images in database. In, FERC-2013 contains about 32000 low resolution face pictures of dissimilar age groups and having different degrees of angle are available. In adding to this, facial expressions have been exhibited very clearly in the CK+ database. (Because they are taken from similar distance and with high resolution images.), Whereas FERC-2013 database, demonstrates emotions in the wild (i.e. ’taken from random distance and are low resolution images’). Which made pictures from the FERC-2013 database are more tougher to interpret. We have trained our system on FERC-2013 database. Since images are ’very clear’ and have well define expressions, they effortlessly classified for different emotions on a face. Therefore, the convolutional networks are trained with the FERC-2013 database.

The database holds of [48x48] pixels of grayscale pictures of human faces. The faces are automatically processed, so that it holds up round a comparably equivalent volume of face space in all images. The prime task is to place each face in view of the emotions of one of seven classes (0: Happy, 1: Sad, 2: Surprise, 3:Angry, 4:Disgust, 5: Fear, 6:Neutral). Thus, database exists in the form of emotion and its matching pixels array. Some examples of FERC dataset are shown in Figure 2.

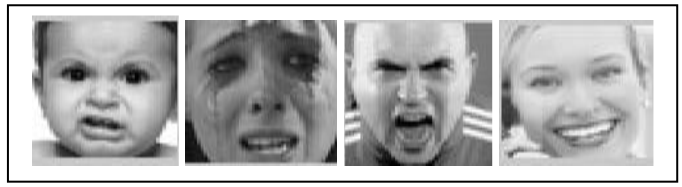


Figure 2. Some Valid Samples of FERC-2013 Database

IV. PROPOSED ALGORITHMS

Algorithm: Whole Algorithm (Figure 1)

step1: if (trained database is not available)
step2: run Algorithm1
step3: run Algorithm 2
step4: save trained database
step5: else (load trained database)
step6: Get input image from webcam or system folder
step7: run Algorithm1
step8: run Algorithm2
step9 :(result 1) display the emotions with percentage of each emotion.
step10 :(result2) analyses of emotions at different rate of intensity.

I. STEP BY STEP DESCRIPTION OF ALGORITHM

A. Image pre-processing

Algorithm 1: Image pre-processing

step1: Get input from user.
step2: Face-detection using Viola Jones algorithm [18].
step3: Taking maximum area face among all faces.
step4: Crop the selected maximum area face from image.
step5: Resize the cropped face into 48x48 images.

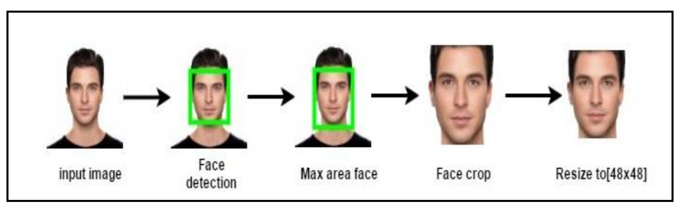


Figure 3. Image Pre-processing

The important elementary phase of algorithm is finding the faces in the picture. We used the existing well known face detector Viola-Jones algorithm. A face identifier must able to state whether a picture containing a human face or not. It is usually the preprocessing step for the face emotion detection system. The Viola-Jones algorithm[18,19], is the most robust face detection algorithm. The algorithm employed mostly three most important stages for fast and flawless face detection. For feature computation we use integral images, Adaboost method is used in feature selection from picture and to boost performance and an attentional cascade for resource allocation on pictures. The initial phase of Viola-Jones algorithm find out the Haar-Like Features [19], which are advance features and are exploited for object classification. The concept of integral image is used for the time reduction in the computational task. The value at pixel (x, y) is the totality of pixels above and to the left of (x, y), all-encompassing. Since the Haar function produces more than 160,000 features, out of which all are not related for face localizing, hence, the AdaBoost algorithm is used to remove the irrelevant features. A set of relevant features is named as weak classifier. The weak classifiers are linearly unified to make a strong classifier. The last step is producing a cascade classifier which is collection of phases. At each phase, it is determined whether the given sub-window contains a face or a non-face. If it fails, it is considered as a non-face.

A. Training Data

Before training, we pre-processed the FERC-2013 database images, in the pre-processing, we used the Viola-Jones algorithm [18,19] on the dataset, we used 28,709 samples for pre-processing and validation among them we got 11246 valid samples for training, Due to drawback of Viola-Jones algorithm, many samples fail in face detection task, some images are shown below:

Figure 4. Failed images of Viola-Jones algorithm

B. Convolutional Neural Networks

In recent times, convolutional neural networks (CNN) have confirmed inspiring performance in plentiful computer vision tasks. Though, excessive performance hardware is obviously very important for the use of CNN models due to the great computation difficulty, which forbids their additional extensions. Our prime objective in this paper is to utilize CNN architectures according to our classification requirement parameters to achieve better accuracy. To achieve this, we employ nine main layers while designing CNN architectures:

Algorithm 2: Deep Convolutional Neural Network

***Phase1*:** We initialize all filters and weights with random values.
***Phase2*:** The training image is input to the network and goes over the forward propagation phases (i.e. convolution layer, ReLU layer and pooling layer actions along with forward propagation in the Fully Connected layer) and detects the gives output probabilities for all class. Let's assume the output probabilities for the first given image are [0.5, 0.2, 0.3, 0.3, 0, 0, 0]. Since weights are randomly assigned for the first training image, therefore output probabilities are also random.
***Phase3*:** Calculating the entire error at the output layer is given as (Summation over all 7 classes).
Total Error = $\sum(\text{target probability} - \text{output probability})^2$
***Phase4*:** Using Back propagation we compute the gradients of the error for all weights in the network and use gradient descent to update all filter values / weights and parameter values to minimize the output error. The weights are updated in proportion to put their influence to reducing the total error. When the same image is imputed again, output probabilities might now be [0.1, 0.1, 0.7, 0.1, 0, 0, 0] which is closer to the target vector [0, 0, 1, 0, 0, 0, 0]. This implies now the network has learnt to categorize this particular picture correctly by altering its weights / filters, so that the output error is reduced. Factor like architecture of the network, number of filters used, filter sizes etc. have all been fixed before Step 1 and do not change during training process – only the values of the filter matrix and connection weights search out updated during the process.

Layer by Layer Explanation of CNN

- ***Layer 0*:** *Input layer* Input[48x48x1] contains the pixel values of the input image. In this case an image of width 48, height 48, and with one colour channel is considered.
- **Layer 1:** *Convolutional layer* calculates the output of

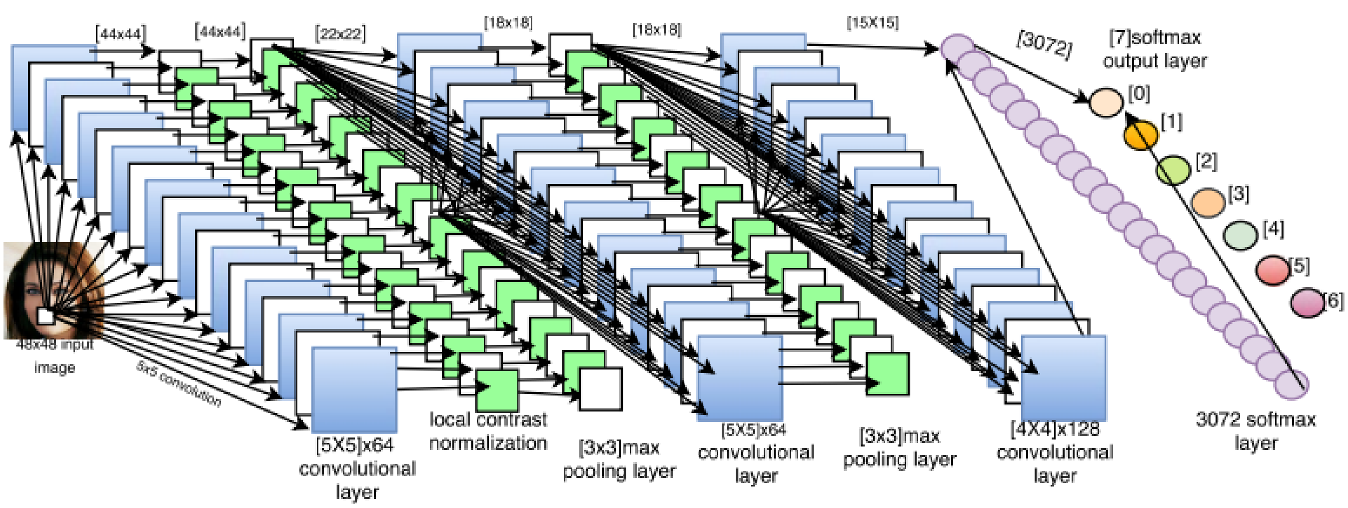


Figure 5. Architecture of Deep Convolutional Neural Network

all neurons that are associated to local regions in the input layer, each calculating a dot product among their weights and a small region they are associated to in

- the input volume. This might result in volume such as [44x44x64] if we decided to use 64 filters with 64 filters of size 5*5, stride 1and padding 0 .so now Total Size is [44 x 44 x 64] and (48-5)/1 + 1 = 44 is the size of the outcome and 64 depths because 1 set denotes 1 filter and there are 64 filters.
- **Layer 2:** *RELU layer* will apply on elementwise activation function, such as the *max* (0, *x*) zero. This leaves the size of the volume unchanged ([44x44x64]),and batch normalization is done.
- **Layer 3:** *Max pool layer* will perform a down sampling operation along the spatial dimensions (width, height), resulting in volume such as [22x22x64]. Max-Pooling with 3×3 filter and stride 2, gives size [22x22x64], i.e. (44-3)/2+1=22 is output size, depth is same as before, i.e. 64 because pooling is done independently on each layer.
- **Layer 4:** Convolution with 64 filters, size 5×5, stride 1, now size is [18x18x64],i.e. (22-5)/1+1=18; is size of output 64 depths because of 64 filters.
- **Layer 5:** Max Poling Layer with 64 filters, size 5×5, stride 1,now size is [18x18x64],i.e. (18+2*1-3)+1=18 original size is restored..
- **Layer 6:** Convolution with 128 filters of size 4x4 and stride 1and we used padding 0,therefore now size is given as [15x15x128], i.e. (18-4)/1+1=15, is size of output 64 and depths of 128 filters.
- **Layer 7:** Fully cconnected with 3072 neurons. In this layer, each of the 15x15x128=28800 pixels is fed into each of the 3072 neurons and weights determined by back-propagation.
- **Layer 8:** Fully-connected layer calculates the class scores, resultant volume of size [1x1x7], where each of the seven numbers correspond to a class score, such as among the seven classes of emotions. As with normal neural networks and as the name implies, each neuron in this layer will be linked to all the numbers in the previous volume and soft max layer with 3072 neurons.
- **Layer 9:** Soft max layer with 7 neurons to predict 7 classes output.

V. RESULTS

The detected emotions and their percentage have been shown below.

C. *Successfully detected emotions*

Figure 6. Row wise: (1) is happy face, (2) is surprised face, (3) is angry face, (4) is Fearful face, (5) is sad face, (6) is disgusted face.

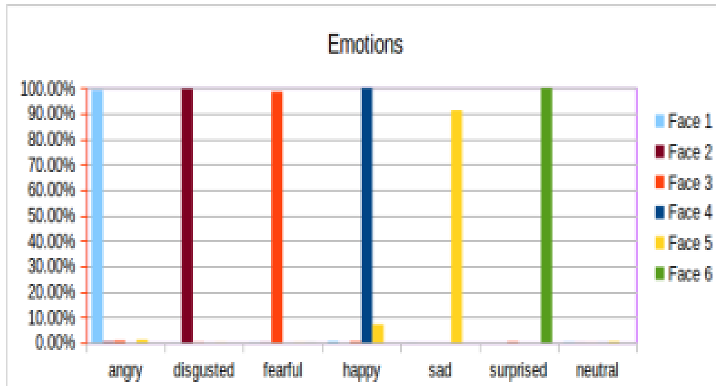
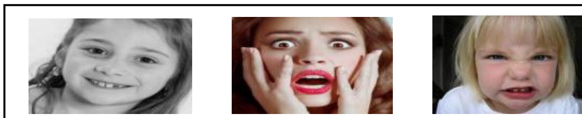


Figure 7. Emotion percentages of successfully detected faces

Some failure cases have also been found. In the majority of the failure cases, the dominant expression is not well defined in the input image itself.

A. *Some failure test cases*



(1) (2) (3)

Figure 8. (1) Happy detected as neutral, (2) Surprised detected as neutral, (3) angry detected as sad

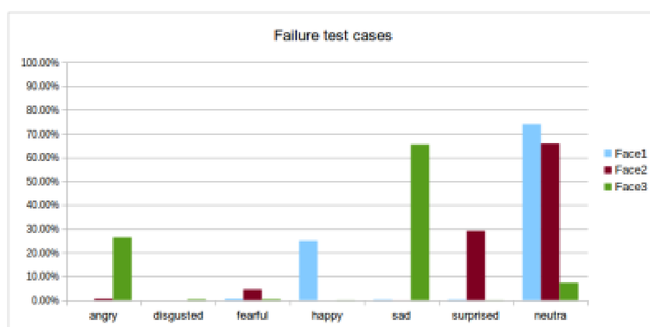


Figure 9. Emotion percentages of above failure test images

The above failures may be due to the dataset imbalance, the (FERC-2013) data set contains non-uniform number of images to different emotions in training set is shown in figure 10. Among 28,709 samples after pre-processing and validation among them we got 11246 valid samples for training Due to drawback of Viola-Jones algorithm [5], [6] most the samples fail during validation

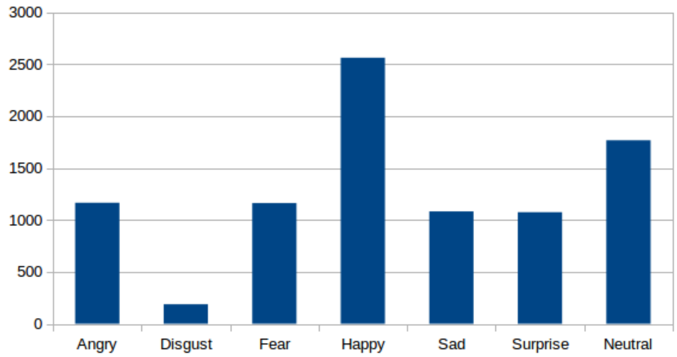


Figure 10. number of sample images for each emotion in FERC-2013 database

.

B. *Emotions at different rate of intensity*

There are numerous diverse models about the nature of emotion and the way that it is characterized in the brain and body. The novelty of the work lies in determining the different degree of the emotions. By this proposed method, we are not only finding the dominant emotion but also finding the percentages of all the presented emotions in the face. Here we are giving the new method for analyzing the degree of emotion while it is changing from one stage of emotion to the other higher state. There are some other emotions which are getting influenced with changes within a time interval, are clearly shown in graphs below. For some basic emotions percentage variation with different time interval, are also depicted. Our approach is very useful to explore micro expressions.

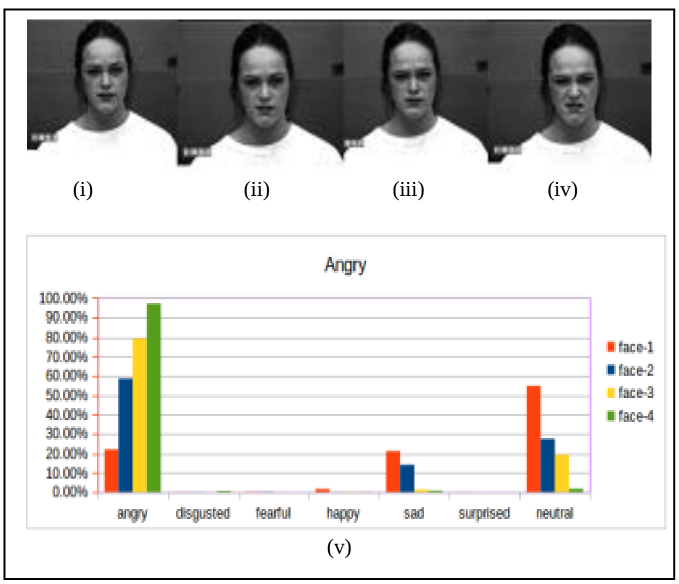


Figure 11. (i)(ii)(iii)(iv) are showing angry face from low level to its extreme level of angry emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level.

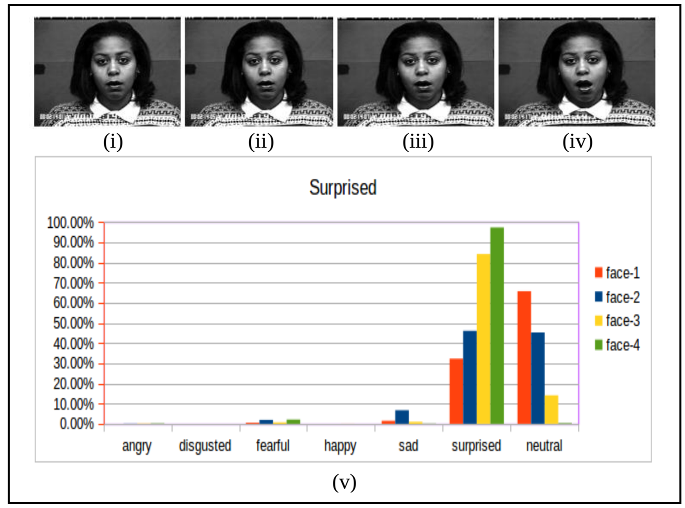


Figure 12. (i)(ii)(iii)(iv) are showing Surprised face from low level to its extreme level of disgust emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level.

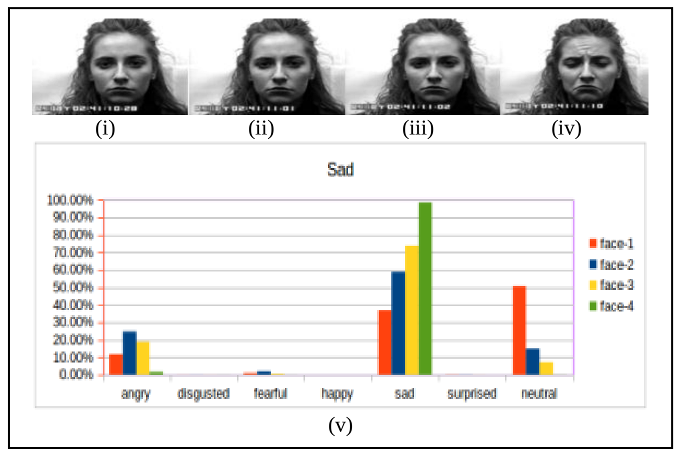


Figure 13. (i)(ii)(iii)(iv) are showing Sad face from low level to its extreme level of happy emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level.

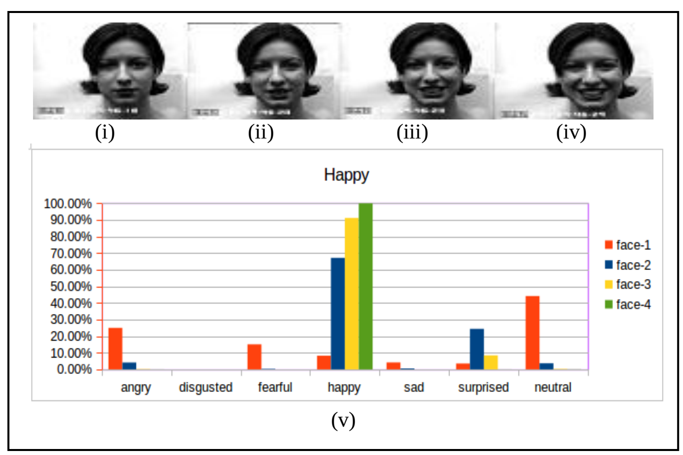


Figure 14. (i)(ii)(iii)(iv) are showing Happy face from low level to its extreme level of sad emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level.

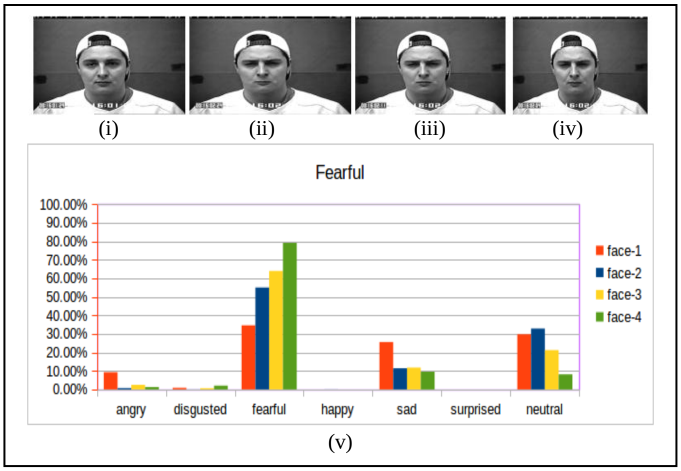


Figure 15. (i)(ii)(iii)(iv) are showing Fearful face from low level to its extreme level of Fearful emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level

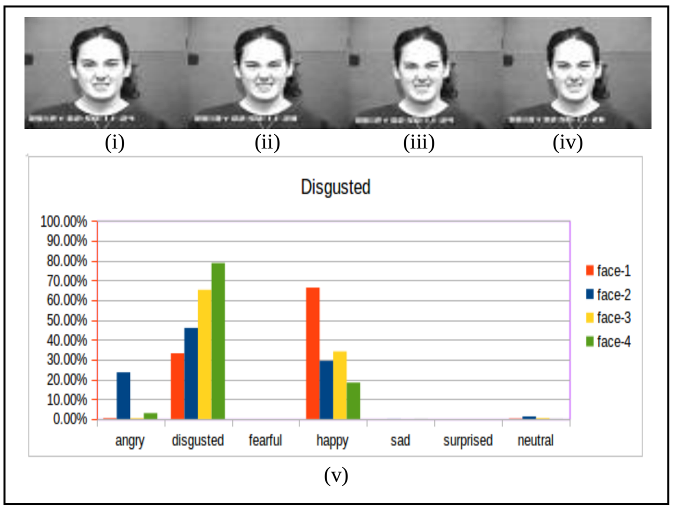


Figure 16. (i)(ii)(iii)(iv) are showing Disgusted face from low level to its extreme level of surprised emotion, and (v) graphical representation of emotion percentages and how other emotions are influencing while emotion level changing from low level to high level.

VI. CONCLUSION

A facial expression of emotions determines the state, mood and current feeling of a person through nonverbal communication. We can understand a person emotion if we analyze it in various stages. In different stages the percentages of emotions are significantly varying. In this paper, we have used convolution neural network with 9 layers, for training and classification of 7 types of standard emotions. For better analysis and interpretation of micro expressions, percentages of emotions in various stages have also been measured with our proposed method. FERC-2013 and Extended Cohn Kanade (CK+) databases have been used in this experiment. For detecting the faces Viola Jones algorithm has been applied prior to recognizing emotion. The normal accuracy rates for people prior to training in Matsumoto & Hwang's (Studied Based on American Physiological Association) study were 48%. A real-time emotion recognition system using face data is proposed and developed using convolution neural networks and the accuracy of the system we are getting around 90+ %.

REFERENCES

[1] Ekman, Paul, and Harriet Oster. "Facial expressions of emotion." *Annual review of psychology* 30.1 (1979): 527-554.

[2] P Forgas, Joseph., and H Gordon. Bower. "Mood effects on person-perception judgments." *Journal of personality and social psychology* 53.1 (1987): 53.

[3] Kagan, Jerome, et al. *Galen's prophecy: Temperament in human nature*. Basic Books, 1994.

[4] Aggarwal, K Jake., and Quin Cai. "Human motion analysis: A review." *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*. IEEE, 1997.

[5] C Borod, Joan. *The neuropsychology of emotion*. Oxford University Press, 2000.

[6] Izard, E Carroll . *Human emotions*. Springer Science & Business Media, 2013.

[7] Adolphs, Ralph, et al. "Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala." *Nature* 372.6507 (1994): 669.

[8] Zhang, Zhengyou, et al. "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron." *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998.

[9] C. J .TECootes., and A. Lanitis. "Active shape models: Evaluation of a multi-resolution method for improving image search." *Proc. British Machine Vision Conference*. 1994.

[10] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. "Facial expression recognition based on local binary patterns: A comprehensive study." *Image and Vision Computing* 27.6 (2009): 803-816.

[11] Lien, James Jenn-Jier, et al. "Detection, tracking, and classification of action units in facial expression." *Robotics and Autonomous Systems* 31.3 (2000): 131-146.

[12] Mian, Ajmal, Mohammed Bennamoun, and Robyn Owens. "An efficient multimodal 2D-3D hybrid approach to automatic face recognition." *IEEE transactions on pattern analysis and machine intelligence* 29.11 (2007).

[13] Sebe, Nicu, et al. "Emotion recognition based on joint visual and audio cues." *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 1. IEEE, 2006.

[14] Liu, Mengyi, et al. "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild." *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014.

[15] Matsugu, Masakazu, et al. "Subject independent facial expression recognition with robust face detection using a convolutional neural network." *Neural Networks* 16.5 (2003): 555-559.

[16] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." *International journal of computer vision* 57.2 (2004): 137-154.

[17] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE, 2001.

[18] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010.

[19] Lienhart, Rainer, and Jochen Maydt. "An extended set of haar-like features for rapid object detection." *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 1. IEEE, 2002.