Emotion Recognition using Convolution Neural Network
Supreet U Sugur, Sai Prasad K, Prajwal Kulkarni, Shreyas Deshpande
Undergraduate Students
supreet2004@gmail.com, saiprasadk463@gmail.com, prajwalkulkarni100@gmail.com, shreyaspande97@gmail.com
Umadevi V, Associate Professor, umadevi.cse@bmsce.ac.in
Department of Computer Science and Engineering
B.M.S. COLLEGE OF ENGINEERING, BENGALURU 560 019, India

Abstract - Human computer interaction has been an important field of study. If computer could understand the feelings of humans, it can cater the services based on the feedback received. In this model we give the overview of the work done in the past related to Emotion Recognition using Facial expressions along with our approach towards solving the problem. The approaches used for facial expression include classifiers like Support Vector Machine (SVM), Convolution Neural Network (CNN) are used to classify emotions based on certain regions of interest on the face like lips, lower jaw, eyebrows, cheeks and many more.

Keywords - Convolution Neural Network (CNN), Face Emotion Recognition (FER), Region of Interest (ROI)

I INTRODUCTION

Emotion recognition is a process of identifying human emotions most likely from facial expressions as well as from speech. The application of emotion recognition system is that it promotes emotion translation between cultures that can be used in multi-cultural communication systems. Human beings have capability to recognize emotions easily, but it is difficult for the computers to do the same. If computers could recognize these emotional inputs, they could give specific and appropriate help to users in ways that are more in tune with the user’s needs and preferences. Facial expressions help computers in detecting emotions. This paper deals with helping computers to recognize human emotions in real-time.

II PREVIOUSLY PUBLISHED ARTICLES

In a paper [1], a hybrid approach in which multi modal information for facial emotion recognition is used. In the experiment conducted by authors, they chose two different speakers using two different languages. The evaluation is carried out with three different media clips, (1) audio information of the emotions only, (2) video information of the emotions only, (3) both audio and video information (original video clip). Video and audio dominance of each type of emotion is recorded and compared. The results of audio and facial recognition are provided as input to the weighing matrix. Inside the weighing matrix computations are made and the expression whose computed value is maximum is the result.

According to a paper [2], the problem that was solved is about Emotion recognition using facial expression. Microsoft Kinect was used for 3D
modelling of the face. Microsoft Kinect has 2 cameras. One works with visible light and the other one works with infrared light. It gives three-dimensional co-ordinates of specific face muscles. Facial Action Coding System (FACS) was used to return special coefficients called Action Units (AU). There are 6 Action Units. These Action Units (AU) represent different region of face. Six men of the age group 26-50 participated and tried to mimic the emotions specified to them. Each person had 2 sessions and each session had 3 trials. 3-NN had an accuracy of 96%. MLP had an accuracy of 90%.

According to the paper [3], CERT can detect 19 different facial actions, 6 different prototypical emotions and 3D head orientation using Facial Action Unit Coding System (FACS) and three emotion modules. It follows 6 stages: (1) Face Detection using Gentle Boost as boosting algorithm, (2) Facial Feature Detection – Specific location estimates are estimated by combining log likelihood ratio and feature specific prior at that location, and these location estimates are refined using Linear regressor, (3) Face Registration – affine wrap is made and L2 Norm is minimized between wrapped facial feature position and canonical position from GENKI dataset, (4) Feature Extraction – feature vector is obtained using Gabor filter on face patch from previous patch, (5) Action Unit Recognition – feature vector is fed to Support Vector machine to obtain Action Unit Intensities, (6) Expression Intensity and Dynamics – Empirically CERT outputs significantly correlates with facial actions.

In a paper [4], Psychological theories state that all human emotions can be classified into six basic emotions: sadness, happiness, fear, anger, neutral and surprise. Three systems were built- one with audio, another with face recognition and one more with both. The performances of all the systems were compared. Features used for speech-global prosodic features, for face-data from 102 markers on face. Both feature level and decision level integration were implemented. The result proved that performance of both the systems was similar. However, recognition rate for specific emotions presented significant errors. The type of integration to be used is dependent on the nature of the application.

III PROPOSED SYSTEM

Video of a person is recorded using web camera. The video is converted to frames and provided as an input to a classifier to get the desired emotion.

This system has three modules. They are

i. Pre-processing : Real-time video is captured using the camera at the rate of 30 frames per second (fps). The frames are in BGR (Blue Green Red) format. It is converted into greyscale format which makes computing easy.

ii. Face detection : A pre-trained classifier called Haar-cascade provided by OpenCV is used for face detection. It also returns face co-ordinates. These co-ordinates are used to crop the image and obtain only the face.

iii. Classifier : Previously built Convolution Neural Network (CNN) model is used and the license to use this CNN model is provided in [6]. The code available in the link [6] was used to train the model. The trained data is used to predict emotion. A list with probabilities of all 7 emotions is obtained as an output. The required output is the maximum of these values and the corresponding emotion is predicted as the final output.
Facial Expression Detection subsystem uses FER2013 [5] dataset which is open-source dataset consisting 48x48 pixel grayscale images. This dataset has 35,887 images, with 7 emotions labelled as (1) happy, (2) neutral, (3) angry, (4) sad, (5) surprise, (6) disgust, (7) fear. In this dataset 32,298 images are used for training and the rest 3,589 are used for testing. The training dataset is a csv file consisting of image in pixels (where each pixel varies from 0 - 255) along with a label indicating the emotion of the person in image. The first 32,298 images are used for training the model. The rest 3587 images are used for testing. The confusion matrix is obtained as shown in the Table 1.

In Table 1, the confusion matrix for various emotions has been shown. The matrix shows the comparison of actual values with the predicted values for each emotion. The matrix shows that angry and sad are confused with a probability of 0.14. Similarly surprise and fear are also confused with a probability of 0.13. The diagonal values give us the accuracy of predicting various emotions. The highest accuracy is that of happy with an accuracy of 87%. The least accurately detected emotions is fear with an accuracy of 42%.

Overall Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}

\[
= \frac{286+35+223+766+300+309+431}{3589} \times 100
\]

\[
= 65.48\%
\]
In this paper, procedure to predict emotions of a person by processing the frames of video through various stages, such as pre-processing, face detection, and classifier using CNN is showcased. The overall accuracy of the model is 65.48%. The emotion with highest accuracy is happy i.e. 87%.

CONCLUSION

Several emotions such as happiness, anger, surprise and fear are better classified using facial recognition. Some others are classified better using speech input such as sadness and fear. So in order to achieve better accuracy, we need to integrate both audio and facial based systems into a single system.

FUTURE WORK

Several emotions such as happiness, anger, surprise and fear are better classified using facial recognition. Some others are classified better using speech input such as sadness and fear. So in order to achieve better accuracy, we need to integrate both audio and facial based systems into a single system.

ACKNOWLEDGMENT

The work reported in this paper is supported by the college (B.M.S. College of Engineering) through TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-III] of the MHRD, Government of India.

REFERENCES

[6] License Link, URL:
   https://github.com/oarriaga/face_classi
   fication/blob/master/LICENSE