Analysis of Facial Emotion Recognition for Image and Video Data using Convolution Neural Networks

*Abstract*—Facial Emotion Recognition is used to classify the emotional states of a human. The objective is to group every facial picture into one of the seven facial emotion classes like Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. Convolutional Neural Network (CNN) is used for classifying the emotion. Various gray scale pictures from the dataset and real-time videos are taken as input. Then feature extraction is done using the series of convolution and pooling layers of CNN and classification is done using softmax layer. To overcome the overfitting problem of the model, various procedures including dropout, cluster standardization and L2 regularization are used. Experiments are performed on FER2013 dataset and achieved better accuracy with our model than the existing works in predicting the individual emotions. In addition to this, the model is also used to predict the emotion of each image in the real time video data with good performance.

Keywords— Facial emotion recognition, Convolution Neural Networks, Classification

# Introduction

Emotion recognition is probably going to get the best outcome if applying various modalities including text, sound, image, and video to understand the behaviour/feelings of a particular human being. Recognition of emotion gives advantages to numerous establishments and parts of life. It is helpful and significant for security and medical services purposes. Additionally, it is essential for simple and straightforward recognition of human physiology at a particular instant without really asking them [1]. Emotions are communicated while collaborating and associating with others. Emotions are generally classified as happy, surprise, neural, fear, sad, anger.

Automatic Emotion recognition uses the Artificial Intelligence algorithms to identify the emotion of the human without interrupting. Every emotion is recognized as an Action Unit (AU) [2]. This implies that when we make a specific facial muscle development, it is connected to a particular inclination. For example, if both the AU "wrinkled forehead" and the AU "upside-down smile" are recognized together, the AI can infer that the individual being examined is miserable. Progressed feeling locators recognize complex sentiments by blending these fundamental groupings. Face Emotion Recognition is the innovation that examinations look from both static pictures and recordings to uncover data on one's enthusiastic state [3]. Facial Emotion Recognition is an innovation utilized for breaking down feelings by various sources, like pictures and recordings. It has a place with the group of advancements frequently alluded to as 'full of feeling processing', a multidisciplinary field of examination on PC's capacities to perceive and decipher human feelings and emotional states and it regularly expands on Artificial Intelligence innovations [4].

Facial Recognition checks the two appearances are same in the event or not. The utilization of facial acknowledgment is immense in security, bio-measurements, amusement, individual wellbeing, and so on [5]. The systems engaged with recognition of Face, two essential elements of passionate view of appearances have been recognized valence and arousal. Valence tends to the pleasantness of a face and is assessed on a straight scale with great/good inclinations toward one side, and offensive/melancholy sentiments at the other. Arousal shows how much a face carries an onlooker to a condition of more noteworthy readiness.

# Literature survey

In most of these non-intrusive techniques, uses video cameras to capture videos within intervals and some techniques uses stereo vision to capture the 3-dimensional images of the human face. Due to the fixed position of the camera, illumination (lighting condition) and pose variations may affect the accuracy/effectiveness of these techniques.

For the facial movement analysis, automatic classifiers for 30 facial actions (include blinking and yawn motions, as well as several other facial movements) from the Facial Action Coding system have been classified by using machine learning [6]. Even though features of the mouth are also used in assessing fatigue and drowsiness.

Qiang et al. used an infrared active sensor to detect both pupil and head motions in variable light conditions. By implementing the Kalman filter, facial features are tracked and a smoothening of the motion of the features is ensured. The Gabor wavelets are used for fast feature detection [7, 8]. Esra et al. detected the face and eyes by using the Adaboost classifier and are then passed through a Gabor filter. The output is normalized and passed to a data driven classifier, a support vector machines (SVM) [9]. In some existing works, the image analysis is realized in a top-down architecture. At first the human face is detected in a scheme based on a boosted cascade of Haar wavelets. Then the eyes are searched in the face and the occurring eye blinks are measured by analyzing the optical flows of the eye’s region.

The Audio/Visual Emotion Challenge and Workshop (AVEC 2019) proposed the algorithms to identify the depression or state-of-Mind, using Artificial Intelligence and Cross-Cultural Affect [10]. It comprises of 3 sub challenges 1. State of mind (SoMS), where level of mood is predicted from personal stories 2. Depression Detection, level of mood is predicted through virtual interviews by AI agent 3. Cross Cultural Emotion Recognition, here cross-cultural in-the-wild paradigm with Hungarian and German as Training and Testing material and Chinese as solely testing material. In SoMS, the level of mood was best predicted even though system trained in static scores and evaluated dynamically. In other approaches prediction of level of depression felt to be more challenging due to change in dimension calculation in noisy environment. Audio and visual features fused and applied to the ResNet and VGG architectures to classify the emotion

Smys et al., proposed a hybrid algorithm which has examined and proved a well recognition of the emotions efficiently. Besides, higher prediction accuracy has been achieved rather than a single classifier [11]. This paper highlights the capability of utilizing the Twitter dataset for measuring the depression of people from their posts and other activities in the web media forum. The analytical performance on the selected dataset, higher accuracy and sensitivity has been achieved to show the early prediction of depression phenomenon.

Pandey et al., proposed a combination deep learning and fuzzy inference system to identify the facial emotion. VGG16 is used to generate the index value with the predicted class for emotion classification. Class index and corresponding images are sent to the corresponding Fuzzy inference system for estimating the intensity level of detected emotion. This system identifies happy, sad, surprise, and angry emotions [12]. In this work, FER 2013, CK + and KDEF combined datasets are used for experimentation.

# Dataset

FER2013 Dataset comprises of 48x48 pixel grayscale pictures of appearances [12]. The countenances have been subsequently enlisted with the objective that the face is basically engaged. The task is to arrange each face considering the tendency showed in the investigate one of seven orders (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The dataset is splitted into train and test. The train set involves 28,709 models and test set includes 3,589 models as Shown in the underneath Table 1.

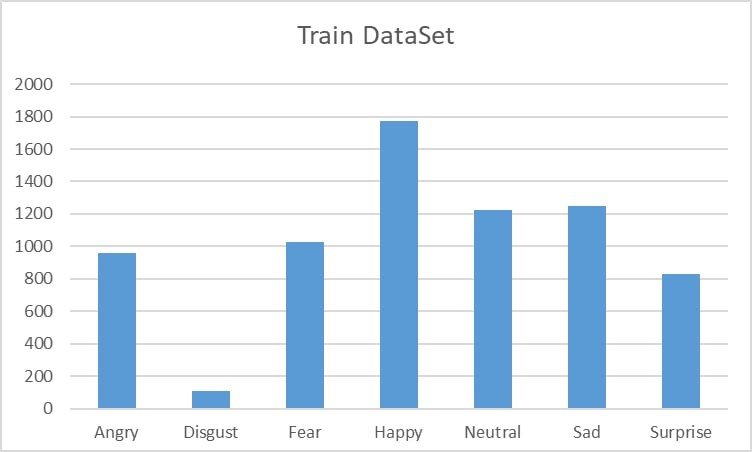
1. Emotions Count in Train and Test Dataset

|  |  |  |
| --- | --- | --- |
| Emotions | Train | Test |
| Angry | 958 | 3995 |
| Disgust | 111 | 436 |
| Fear | 1024 | 4097 |
| Happy | 1774 | 1174 |
| Neutral | 1223 | 4965 |
| Sad | 1247 | 4830 |
| Surprise | 831 | 3171 |

The facial features are extracted from this dataset are given as input to the classification algorithm and predicting the corresponding emotion of the image. Convolution Neural Network is used to train the model. The test tests are given as contribution to the model and anticipate the relating feeling. The include of various feelings in train and test information are displayed in figures1 and 2.

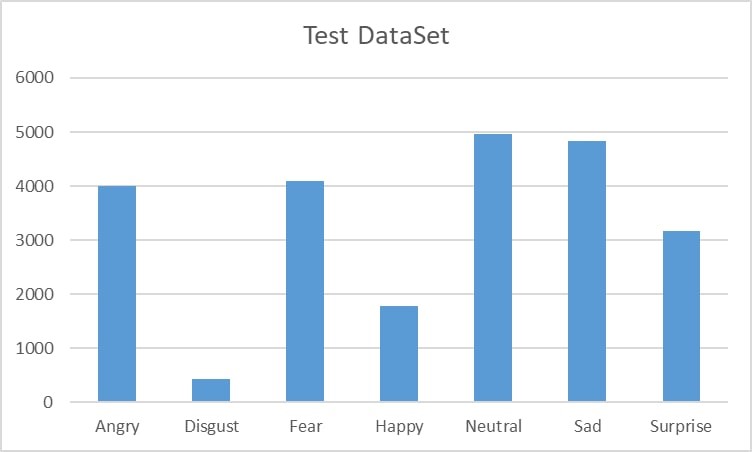
## Video Dataset

Video is about reaction of social occasion while they are noticing some interesting thing on the screen. It is around 29 seconds video which presents different responses of people changing quickly together activities going with respect to the screen. It comprises 5 people, 2 are men and 3 are women. There are giving enunciations as per thing happening front of

them. 

1. Count of Different Emotions Used in Train Dataset

This is how the input will give to move further process of max pooling and batch normalisation. Activation function will be going to participate in this which is a forward move to the previous pooling and reshape it then get a result to passed input.



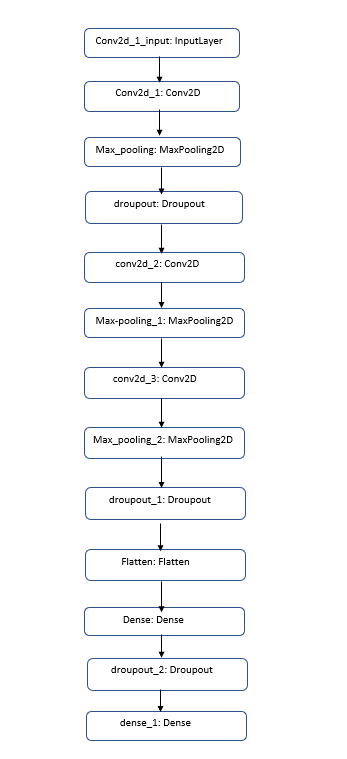
1. Count of Different Emotions Used in Test Dataset

# Methodology

In deep learning, Convolutional Neural Networks (CNN) is a type of artificial neural network, which is used for image recognition and classification. It automatically detects the important features. CNN is mainly used for image processing, computer vision tasks like localization, segmentation, video analysis, recognition of objects, etc. Convolutional neutral network is class which manages the picture include extractions and related elements of symbolism information. So decided to include a Convolutional Neural Network to deal with this face affirmation issue. Without a doubt this kind of Neural Network (NN) is incredible for eliminating the features of pictures and is extensively used for picture examination subjects like picture portrayal [3].

CNNs have been used for understanding in Facial Emotion. A Neural Network is a learning framework that involves in various layers of phony neurons. Each center point gets weighted data, passes it into an inception limit and results the delayed consequence of the limit. A data layer that will get the data. The size of the data layer depends upon the size of the data. A couple of mystery layers that will allow the NN to learn complex interchanges inside the data.

CNN is constructed with multiple layers of artificial neurons, each layers generates several activation functions that will be passed to the next layer. The first layer extracts basic features such as horizontal or diagonal edges, this output is passed to next layer which extracts more complex features like corner or combinational edges. As we move into deeper layers in the networks it can identify more complex features such as faces, objects etc.



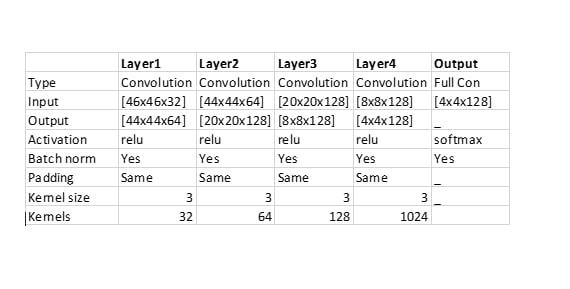
1. Workflow Of Convolutional Neural Networks For Facial Emotion Classification

A Neural Network with a lot of hid away layers is known as a Deep Neural Network [3]. The current composition centers around seven fundamental look classes revealed, which are fear, sad, disgust, angry, neutral, surprise, happy. The CNN calculation introduced in this original copy focuses on expressional assessment and to describe the given picture into these seven fundamental emotion classes [7]. Emotion recognition is the most common way of distinguishing human emotion [4]. Individuals change generally in their exactness at perceiving the feelings of others. Utilization of innovation to assist individuals with emotion recognition is a somewhat early examination region.

Many methodologies have been proposed and a lot more are springing up. We can effectively prepare a simple neural network to perform classification. In any case, it may not perform well with images. Contrast with this CNNs are totally related feed forward networks.

CNNs are outstandingly convincing in diminishing the number of boundaries without losing on the idea of models. Pictures have high dimensionality (as each pixel is considered as a component) which suits the above portrayed limits of CNNs. This review scarcely started to expose CNNs here however gives a fundamental instinct. The workflow of CNN used in our work is given in Fig 3.

The convolution Neural Network Layer specifications are shown in Figure 4.



1. Convolution Neural Network used for Facial Emotion Recognition

Dropout: diminishes overfitting by arbitrarily not refreshing the loads of certain hubs. This holds the NN back from relying upon one hub in the layer unnecessarily. In this work, softmax activation function is used for classification work as it is usually utilized for multi-label classification.

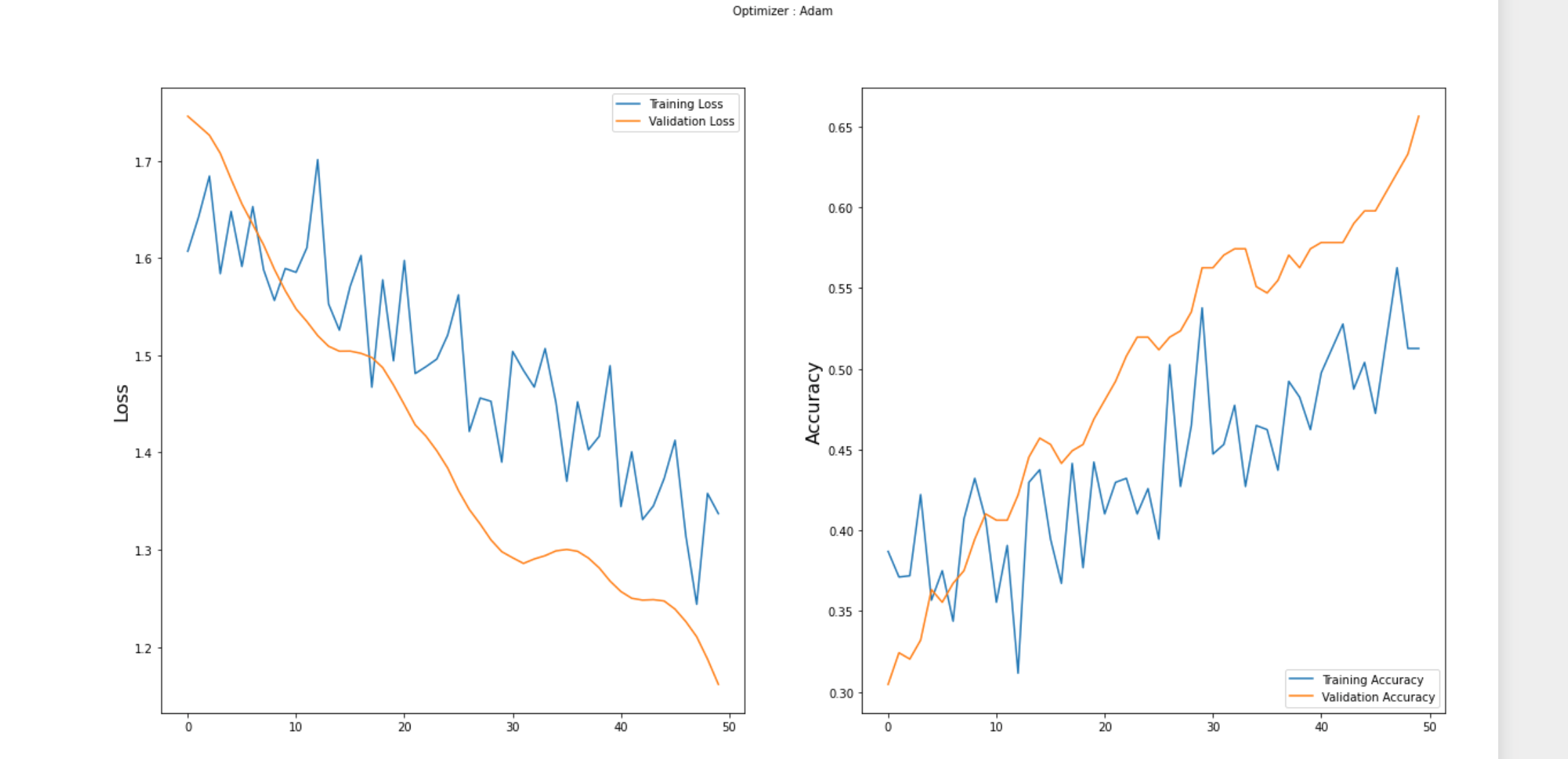
For this project we used CNN architecture for model building. Initially we imported all the required libraries for our model construction. Then we initialized the training and validation generators, which meant rescaling all the images we needed to train our model and converting them to grayscale images. We used FER2013 dataset for model building. Here we trained our network on all the images we have i.e., FER2013 dataset and then saving the weights in model for the future predictions. Finally, using OpenCV harassed xml, we detected the bounding boxes of face in webcam and predict the emotions.

# Eperimental Results

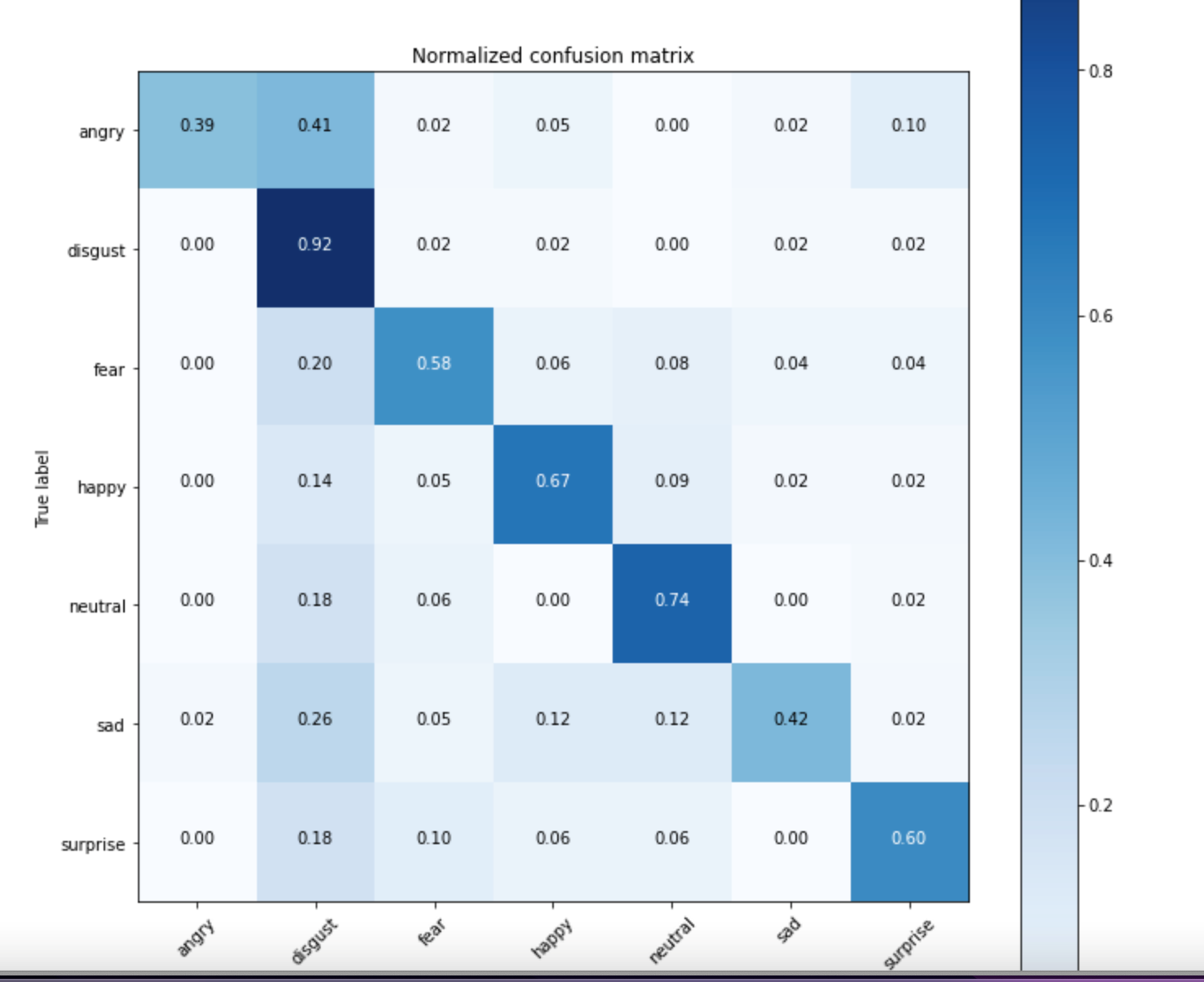
An output layer that will give the possible result, for instance a class assumption. The size of the output layer depends upon the number of classes really want to make. Convolutional Neural Networks similarly have Convolutional layers that gathers pixels that are near each other. Thusly those developments have a predominant understanding of models that can find in pictures. In this work, used a couple of ordinary strategies for each layer. Result Analysis for FER2013

In this cycle, the information will go through the 4 convolutional layers and 2 completely associated layers. Proposed work utilized the soft max and Relu activation functions in CNN for flattening process. As shown in Fig. 5 the mentioned graphs are presenting the loss and accuracy of both training and testing process. In graph1 it is all about the loss of training and validation and in graph2 shows accuracy in training and testing. Overall, testing the several images of emotions it exposes 75% accuracy of several emotions.

Figure 6 Confusion matrix addresses the exactness of every feeling through preparing dataset of pictures with different feelings. Emotions referenced are like 1. angry, 2. disgust, 3. fear, 4. happy, 5. neutral, 6. sad, 7. Surprise. It is perceiving feeling in each picture that applied in approval process then the resultant network as displayed in underneath Fig. 6.



1. Representation of Loss and Accuracy of the system

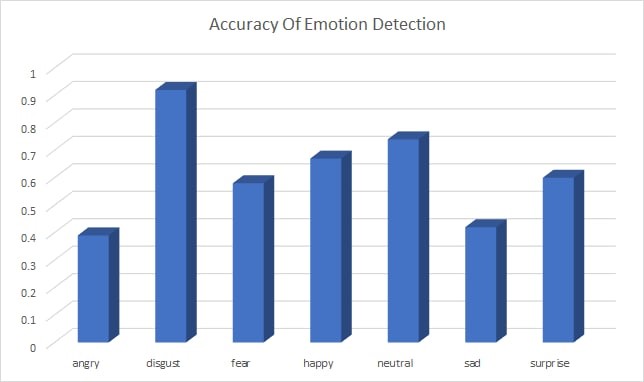


1. Dataset Representation of Confusion matrix for individual Emotions



1. Representation of various emotions during model training

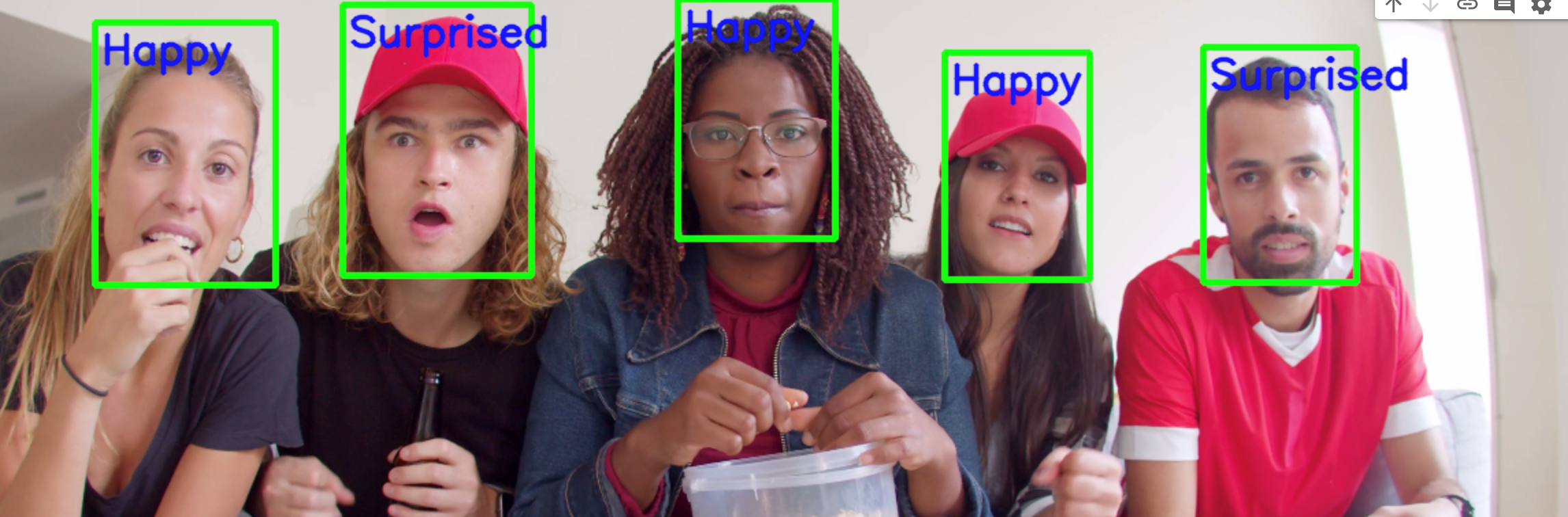
As Fig. 7 shows visual chart portrays the exactness of various feelings by following above technique utilizing convolutional neural network model. These shows every feeling expectation rate while going through above utilized model and at first the model prepared with some dataset containing different enthusiastic pictures and tried with pictures of feelings found above precision rate.



1. Accuracy of Individual Emotions Using Bar Graph

## Result Analysis For Video

CNN model is used to perceive the feeling on picture premise. At first, model prepared with picture dataset in preparing area. This model created with 2 convolutional layers and 2 completely associated layers and toss the information through this layer. Finally, this is going under leveling process utilizing softmax enactment layer. After this it utilizes the Haar cascade which can give powerful output in expression recognition [9]. Then passed a video with multiple persons, it considers whole video as several frames and identify the emotion of multiple persons in that frame. Displayed result is that recognition of emotion in that frame of video.



1. Recognition of Emotion While Playing the Video

Fig. 8 shows the result of video emotion recognition system where the input is video. It is process of identification of emotions of persons through on-going video. Identification of the emotions by taking frames for every second. On screen moments are collected for every person in the video. Our model is efficient in classifying the emotion of every person in the video.

# Conclusion

The proposed work developed a facial emotion recognition system using CNN presenting different expressions of people by considering images of them and video mode dataset which is running of person’s emotion on basis of trained emotional images. It classified the emotions in seven categories such as happy, sad, neutral, surprise, disgust, fear, angry. It visualizes the emotion of person image.

##### References

1. Dalal, Navneet, and Bill Triggs (2005). Histograms of oriented gradients for human detec- tion. Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. Vol. 1
2. Lonare,Ashish,andShwetaV.Jain(2013).ASur- vey on Facial Expression Analysis for Emotion .Recognition. International Journal of Advanced Research in Computer and Communication Engi- neering 2.12
3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep con- volutional neural networks.” Advances in neural information processing systems. 2012.
4. Simonyan,Karen,andAndrewZisserman.”Very deep convolutional networks for large-scale im- age recognition.” arXiv preprint arXiv:1409.1556 (2014).
5. Divya Meena Ravisharan, “An approach to face detection and recognition.” 2016 IEEE international Emotion Recognition Using Convolutional Neural Network (CNN) July 2021Journal of Physics Conference Series (https://www.researchgate.net/journal/Journal-of-Physics-Conference-Series-1742-6596).
6. G. Bhardwaj, S. V. Singh and V. Kumar, "An Empirical Study of Artificial Intelligence and its Impact on Human Resource Functions," 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2020, pp. 47-51, doi: 10.1109/ICCAKM46823.2020.9051544.https://developers.google.com/machine-learning/crash-course/first-steps-with-tensorflow/toolkit .
7. Kukla E., Nowak P. (2015) Facial Emotion Recognition Based on Cascade of Neural Neworks. In: Zgrzywa A., Choroś K., Siemiński A. (eds) New Research in Multimedia and Internet Systems. Advances in Intelligent Systems and Computing, vol 314. Springer, Cham. doi.org/10.1007/978- 3-319-10383-9\_7.
8. P. Qiang Ji; Zhiwei Zhu; Lan, "Real Time Non-intrusive Monitoring and Prediction of Driver Fatigue," *Vehicular Technology, IEEE Transactions,* vol. 53, pp. 1052 - 1068, 2004.
9. M. C. Esra Vural, Aytul Ercil, Gwen Littlewort, Marian Bartlett and Javier Movellan, "Drowsy Driver Detection Through Facial Movement Analysis " in Human–Computer Interaction. vol. 4796, ed: Springer Berlin / Heidelberg, 2007, pp. 6-18
10. Ringeval, Fabien & Schuller, Björn & Valstar, Michel & Cummins, Nicholas & Cowie, Roddy & Tavabi, Leili & Schmitt, Maximilian & Alisamir, Sina & Amiriparian, Shahin & Messner, Eva & Song, Siyang & Liu, Shuo & Zhao, Ziping & Mallol-Ragolta, Adria & Ren, Zhao & Soleymani, Mohammad & Pantic, Maja. (2019). AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition.
11. Dr. S. Smys, Dr. Jennifer S. Raj, “Analysis of Deep Learning Techniques for Early Detection of Depression on Social Media Network - A Comparative Study”, Journal of trends in Computer Science and Smart technology (TCSST) (2021) Vol.03/ No. 01 Pages: 24-39
12. Pandey, A., Kumar, A. Facial Emotion Intensity: A Fusion Way. SN COMPUT. SCI. 3, 162 (2022). https://doi.org/10.1007/s42979-022-01049-5