A CNN-LSTM BASED DEEP NEURAL NETWORKS FOR FACIAL EMOTION DETECTION IN VIDEOS

Arnold Sachith A Hans
Department of Computer Science and Engineering (Artificial Intelligence),
Presidency University Bengaluru, Karnataka, India
sachithhans@gmail.com

Smitha Rao
Department of Computer Science and Engineering (Artificial Intelligence),
Presidency University Bengaluru, Karnataka, India
smitharao@presidencyuniversity.in

Submitted: *Nov, 10, 2021* ***Revised:*** *Feb, 20, 2021* ***Accepted:*** *Feb, 28, 2021*

Abstract: Human beings while communicating use emotions as a medium to understand the other person. Face being the primary source of contact while communicating and being the most communicative component of the body for exhibiting emotions, facial emotion detection in videos has been a challenging and an interesting problem to be addressed. The Facial expressions fall under the category of non-verbal type of communication and understanding Emotional state of a person through Facial Expressions has many use cases such as in the field of marketing research – understanding the customers responses for various products, Virtual classroom – understanding the comprehension level of the students, Job Interview – in understanding the changes in emotional state of the Interviewee, etc. This research paper proposes a CNN- LSTM based Neural Network which has been trained on CREMA-D dataset and tested on RAVDEES dataset for six basic emotions i.e. Angry, Happy, Sad, Fear, Disgust, and Neutral. The Faces in the videos were masked using Open Face software which gets the attention on the Face ignoring the background, which were further fed to the Convolutional Neural Network. The research focuses on using LSTM networks which have the capability of using the series of data which will aid in the final prediction of emotions in a video. We achieved an accuracy of 78.52% on CREMA-D dataset and further also tested the model on RAVDEES dataset and achieved an accuracy of 63.35%. This research work will help in making machines understand emotions, can help systems make better decisions and respond accordingly to the user.

Keywords: Facial Emotion Detection, Masked Face, CNN, LSTM, Open Face.

I. INTRODUCTION

After COVID-19 has spread throughout the world, most of the activities have been happening in a virtual mode. The high-performance personal computers have become popular due to its varied applications and accordingly the interaction between the humans and these systems have been increasing on a daily basis. Therefore, making machines understand human emotions will help improve Human-machine interaction systems. The main goal of a human interface is to understand the user's emotions and give personalized media content accordingly

Emotions play a significant role in making life decisions. Emotions in humans are biological states, according to neuropsychology science, and these states have been linked to our nervous system, beliefs, feelings, and behavioral responses. Motivation, personality, character, and attitude are all intertwined with emotions. Emotions can be differentiated into positive and negative experience with respect to the different environmental activities the human is involved in.

Convolutional Neural Network (CNN) is a flavor of Deep Neural Networks, which focuses extracting the useful features from images. In their fully connected layers, CNN [4] uses a mathematical operation called convolution that is used instead of normal matrix multiplication to generate the features from the input data. The CNN [5] which convolves image data through many filters and finally produces a feature map, further the feature map produced is combined with the fully connected layers of neurons and finally the Emotions are recognized as belonging to a particular class based on output [6]. Apart from this CNN can be used for different applications involving images/videos such as medical image-based analysis, agricultural field (plant disease detection), Object Detection, Facial Recognition, Face detection etc.

LSTM is a type of Recurrent Neural Network (RNN) which is a part of the Deep Neural Networks. LSTM is extensively used where data varies according to time such as Speech recognition, Handwriting recognition, Video Analysis, Weather Forecasting, Stock Prediction etc [11]. LSTM is primarily made up of four basic units and they are self, forgot gate, input gate, output gate. CNN focuses on extracting the low-level features from an Image, while LSTM aids in establishing the temporal relationship between the frames in a video.

Emotion Detection from Facial Expressions can be used in a variety of ways. Emotions play an important role in deciding an individual's conduct, which in turn aids fellow humans in comprehending those around them [12]. The majority of people suffer from depression, which is readily identifiable from facial expressions. In a virtual classroom, determining the emotional state of the students helps in identifying comprehension level of the students involved in the class and many more.

In this paper, we propose a method to recognize emotions from masked facial images derived from a sequence of images (video) by extracting the temporal relation between the images using Deep Neural Networks [13]. The organization of the paper is as follows: Section II discusses the related work carried out in emotion recognition, Section III includes the proposed Facial based emotion recognition method, and section IV presents the results obtained, and concludes with section V.

II. LITERATUREREVIEW

With the rapid growth of artificial intelligence techniques, the research in the field of Facial Emotion Detection has gotten a lot of attention in the recent decades [14]. Several feature-based methods for Facial Emotion Detection have been investigated. These approaches focus on the facial part in an image, detect the facial region in an image and extract the geometric features from the region [15] various different types of information on facial regions are attracted as appearance features [16]. Several research local regions and extracted region specific appearance features [17]. Further among these targeted local regions the important regions are identified which results in an improvement in the recognition accuracy [18]. There has been extensive development in the field of Deep Neural Networks, the Convolutional Neural Networks (CNN) and Recurrent Neural Networks [19] (RNN) has been applied to various applications involving computer vision [20]. CNN has achieved great results in various studies which involve face recognition, object recognition, and Facial Emotion Recognition [21].

In a Convolutional Neural Network [4] model with 8 layers addition of pooling and dropout layers has been proposed and was trained on 2 datasets i.e. RAVDEES[1] and Cohn-Kanade considering only 6 emotions surprised, happy, disgust, angry, neutral and sad, the results obtained were better for surprised and happy compared to all other emotions [2]. The RAVDEES Dataset [5] is a multimodal database of emotional speech and song. This dataset consists of 24 actors out of which 12 are male and 12 are female.

Emotions can be both positive and negative but recognizing the transition of the emotion shifting from positive state to negative state is the real challenge [3]. In[3]the research work has been carried on to determine the contribution of audio-visual cues in emotion recognition and how temporal relation between audio and video impacts the rate and speed on emotion recognition. In [6] work has been carried out on RAVDEES dataset [4]; the frames have been increased in the step of 5 starting from 5 frames/video till 65 frames/video. Two Neural Networks have been designed namely Spatial Temporal CNN [6] and GRU cell RNN. The articulation related cues and Facial features have been used to train the neural networks. For 35 Frames/video along with Facial features and accuracy of 47.9%hasbeenachieved, CNN [4] with Lip features on 65 Frames/video an accuracy of 56.9% and with RNN 48.5% were the benchmark results obtained. RNN with Lip features gave an accuracyof59.4%

III. PROPOSED SETUP

This section gives information about the dataset used, data pre-processing techniques carried out to extract the relevant features to perform the analysis and the proposed deep neural network model architecture.

A. Dataset Used–CREMA-D Dataset

The dataset consists of 7442 files in total from 91 actors. There were 48 male and 43 female actors involved between the ages from 20-74 who hailed from different races and ethnicities (African America, Asian, Caucasia n, Hispanic, and Unspecified) [13]. The data set includes videos having 6 different emotions they are: Neutral, Sad, Happy, Angry, Disgust and Fear. Each expression for an emotion is recorded at four different levels – high, medium, low and unspecified. These videos were rated by 2243 individuals who were crowd- sourced studies have divided the specific facial area into targeted.

B. Data pre-processing

The masked faces were extracted from the videos using Open Face [7] software at 30 frames/sec. There were totally 7442 videos in the CREMA-D corpus. All the videos had different duration in length. There were 3638 videos whose frames were between0-75, there were 3217 videos which had76- 100 frames, there were 565 videos which had 101 - 150 frames and around two videos which had more than 151 frames [8]. Due to the UN even distribution of video length in terms of duration. To carry out the experiment we fixed the total number of frames per video to be as 75 frames for one video. we deleted all the videos which had less than 35 frames per video, for videos with duration between 35 to 75 seconds, we used the formula [9].

$$n= (75-nf) \quad (1)$$

where nf is the total number of existing frames in a video, using the above-mentioned formula, we repeated the last frame n number of times this made sure there were 75 frames for each video. Similarly, forvideoshavingframesbetween75-110frames, the 75-110 frames were deleted but the initial 0-75 frames were retained along with its respective labeling. For150 frames/video it was divided into 75framespervideoandthesamelabellingwasused for both the set [10].

The data was split into three parts namely train, validation and test data. The train data consisted of 5000 videos, validation – 1000 videos and the rest 1704 videos were present in the test dataset. Figure 1 explains the data distribution in train, validation and test set with respect to emotions.

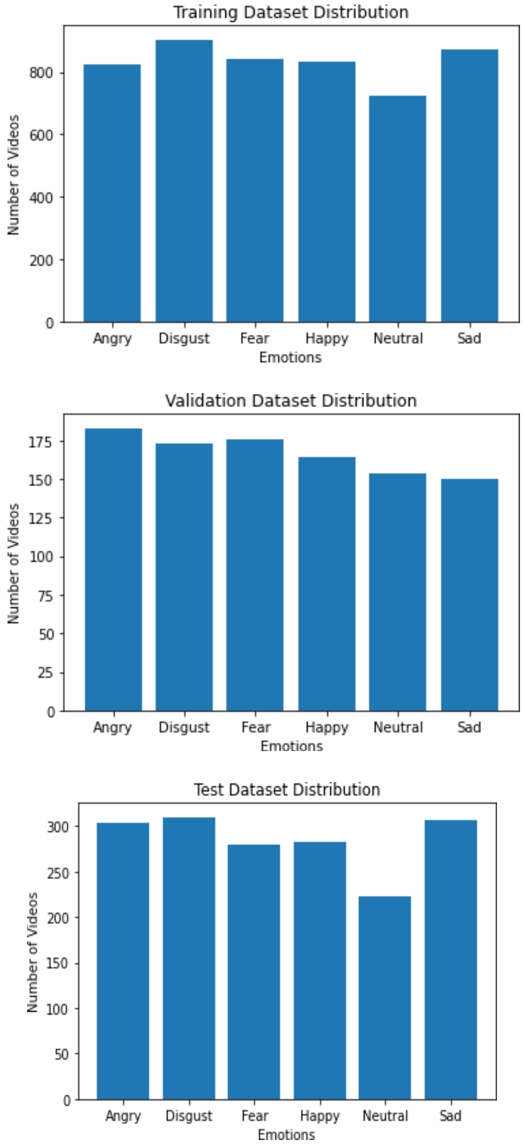

Fig. 1 Data split –train, validation and test set

The original frames in which the faces were masked obtained from the Open Face tool had a size of 112X112 further we have resized the images to 28X28 to carry out the analysis in Figure 2.

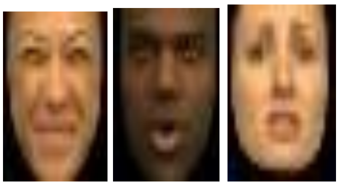

Fig. 2 Masked face images in RGB format, size– 28X2

The masked face images were further converted into array format and were normalized by dividing the values by 255.

C. Proposed Architecture

This section describes the proposed Deep Neural Network architecture. The architecture consists of 5, 6, 7 layered neural network architecture. It consists of CNN layers at the initial stages followed by LSTM layers and finally soft max layer.

The Input shape of the images is 75x28x28x3 for RGB format. The array input of the particular image is fed into 3D convolutional block which consists of convolutional filters, L2 kernel regularization, activation function – Leaky ReLU, Batch Normalized and Spatial Dropout 3D. For 5-layer no further 3D Convolution block was added and the output of 3D Convolution block was directly fed in to LSTM block with 64 units and drop outs. Further it was fed into dense block with 32 units and activation function – Leaky ReLU. Finally, it was fed into the output layer with Soft Max activation. For 6-layer one more extra 3D convolutional block was added and consecutively one extra layer was added to carry out the analysis and we stopped our analysis at 7- layer. It was fed into the output layer with Soft max activation in Figure 3.

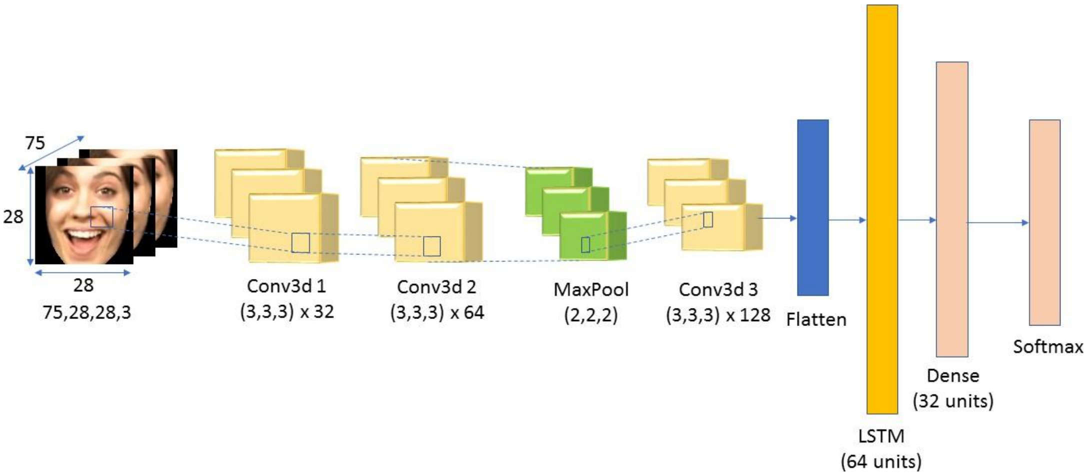

Fig. 3 Proposed 6-layer CNN-LSTM architecture diagram

D. Model Training

The dataset was trained on the above-mentioned architecture. The architecture was trained with 50,75 epochs, batch size of 32, optimizer–Adam [Beta_1= 0.99, Beta_2 = 0.999] with a learning rate of 1-e3, 1-e4,1-e5, loss function of category cal cross entropy, metrics accuracy and the model check point was created to save the model with the best validation accuracy.

IV. RESULTS

This section gives an overview on the experimental results obtained while carrying out the analysis. Table 1 describes the accuracies obtained for the proposed architecture with different number of epochs, learning rate and layers varied accordingly.

TABLE 1 Train, validation and test accuracies on CREMA-D dataset

Model	Epochs	Learning Rate	Train Loss	Test loss	Val_ loss	Train Acc	Validation Accuracy	Test Acc
5 Layer	50	0.0001	0.193	0.729	0.713	0.943	0.771	0.754
6 Layer	50	0.0001	0.227	0.721	0.736	0.950	0.789	0.785
7 Layer	50	0.0001	0.333	0.767	0.730	0.933	0.80	0.781
5 Layer	50	0.001	1.317	1.331	1.346	0.455	0.459	0.454
6 Layer	50	0.001	1.257	1.25	1.242	0.535	0.544	0.537
7 Layer	50	0.001	1.312	1.332	1.314	0.546	0.56	0.547
5 Layer	50	0.00001	0.623	0.746	0.771	0.795	0.734	0.73
5 Layer	75	0.00001	0.462	0.701	0.718	0.858	0.764	0.74
6 Layer	50	0.00001	1.244	1.224	1.234	0.543	0.555	0.567
6 Layer	75	0.00001	1.215	1.26	1.250	0.568	0.569	0.554
7 Layer	50	0.00001	1.081	1.02	1.016	0.646	0.672	0.660
7 Layer	75	0.00001	0.827	0.85	0.848	0.751	0.745	0.73

From Table 1 based on test accuracy we can infer the top three models they are firstly 6 layers CNN- LSTM with learning rate of 0.0001 which obtained a Test accuracy of 78.52% and Validation accuracy of 78.9%, secondly 7 layers CNN-LSTM with learning rate of 0.0001 performed well with Test accuracy of 78.11% and validation accuracy of 80% and thirdly 5 layers CNN-LSTM with learning rate of 0.0001 performed well with Test accuracy of 75.47% and validation accuracy of 77.1%**.** Figure 4 depicts the change of training and validation accuracy with respect to epochs. Figure 5 shows the confusion matrix for CREMA-D dataset based on the results from the 6-layer CNN-LSTM model.

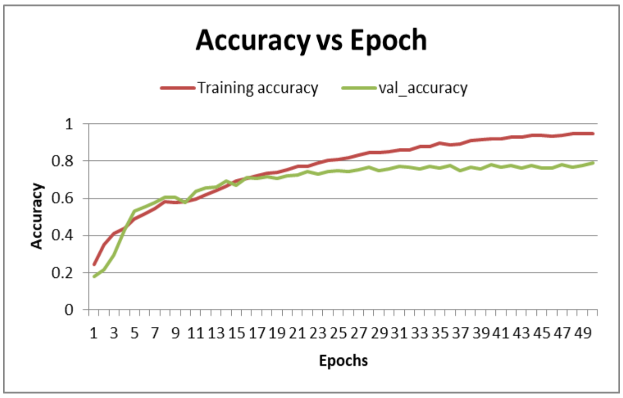

Fig. 4 Accuracy vs epochs graph for 6-layer CNN- LSTM model learning rate = 0.0001

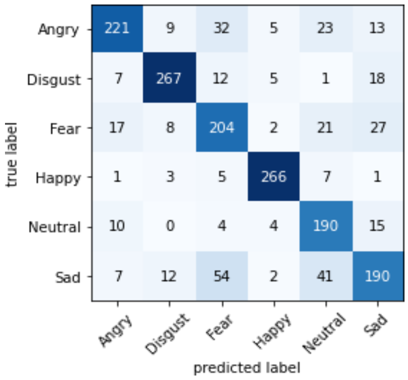

Fig. 5 Confusion matrix based on test data of CREMA-D dataset

It is observed that after 30 epochs the validation accuracy has been trending between 76-78% while the training accuracy reaches 95% at the end of 50 epochs. It can be inferred from Figure 5 that emotion Happy having 18 misclassifications achieved top accuracy of 93.6% while Fear and Neutral had 107, 93 misclassifications respectively and achieved accuracy of 65.5% and 67.13% respectively. It can also be observed that highest misclassification was 54 where emotion sad was detected as fear by the model for CREMA-D Dataset.

TABLE 2 Top 3 model accuracy on RAVDEES dataset

Model	Accuracy on RAVDEES Dataset
6 Layer CNN-LSTM, lr = 0.0001, Epochs = 50	63.35%
7 Layer CNN-LSTM, lr = 0.0001, Epochs = 50	60%
5 Layer CNN-LSTM, lr = 0.0001, Epochs = 50	57.86%

The best model that is 5, 6, 7 layers CNN-LSTM model with the learning rate set at 0.0001, was also tested on RAVDEES [5] dataset. The videos with the Label calm and surprised were deleted from the RAVDEES dataset since our model was trained only on 6 emotions – Angry, Disgust, Fear, Happy, Neutral and Sad. Table 2 summarizes the accuracy which was obtained testing on RAVDEES dataset as per the observations the 6 Layer CNN-LSTM performed well when compared with all other models and gave an accuracy of 63.35%. Figure 6 depicts the confusion matrix for RAVDEES dataset.

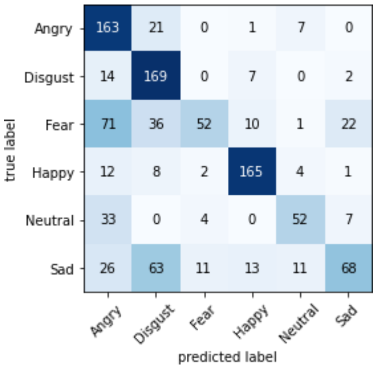

Fig. 6 Confusion matrix based on testing performed on RAVDEES dataset

The overall accuracy obtained is 63.35% as seen in Table 2. The highest individual accuracy 84.1% was obtained by happy emotion and there were only 31 misclassifications. Total of 156 misclassifications was found for angry emotion and the lowest individual accuracy of 51.1% was obtained. Most of the angry emotions were misclassified as Fear which can be seen in Figure 6.

V. CONCLUSION

The research work carried out proposes a CNN-LSTM based Deep Neural Network architecture which takes into consideration the temporal relation between the masked facial images at 75 frames for each prediction. The 6-layer CNN-LSTM with learning rate set to 0.0001 which was trained on CREMA-D dataset gave the test accuracy of 78.53% and the same model was tested on RAVDEES dataset and achieved an accuracy of 63.35%. Further the model can be improved by adding speech features, Facial Action Units to the same architecture and these features can be further stacked together before sending them to the LSTM layer which might help in improving the accuracy of the model.

REFERENCES

[1]. A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam and I. Zaman, "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames," International Conference on Robotics, Electrical and Signal Processing Techniques, 2019, pp. 281-285.

[2]. M. G. de Pinto, M. Polignano, P. Lops and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks

and Mel-Frequency Cepstral Coefficients," IEEE Conference on Evolving and Adaptive Intelligent Systems, 2020, pp.1-5.

[3]. E. Ghaleb, M. Popa and S. Asteriadis, "Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition," 8th International Conference on Affective Computing and Intelligent Interaction, 2019, pp. 552 -558.

[4]. Z. Rzayeva and E. Alasgarov, "Facial Emotion Recognition using Convolutional Neural Networks," IEEE 13th International Conference on Application of Information and Communication Technologies 2019, pp. 1 - 5.

[5]. S. R. Livingstone, F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):" A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE Vol. 13 No. 5, 2018, e0196391.

[6]. S. Bursic, G. Boccignone, A. Ferrara, A. D'Amelio, R. Lanzarotti, "Improving the Accuracy of Automatic Facial Expression Recognition in Speaking Subjects with Deep Learning." Appl. Sci. Vol. 10, No. 11, 2020, 4002.

[7]. Open Face 2.0: "Facial Behavior Analysis Toolkit Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency," IEEE International Conference on Automatic Face and Gesture Recognition, 2018.

[8]. M. F. H. Siddiqui, A. Y. Javaid, "A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images." Multi modal Technol. Interact, Vol. 4, No. 3, 2020, pp.46.

[9]. S. W. Byun, S. P. Lee, "Human emotion recognition based on the weighted integration method using image sequences and acoustic features." Multimed Tools Applications, 2020, pp. 1-15.

[10]. X. Wang, X. Chen, C. Cao "Human emotion recognition by optimally fusing facial expression and speech feature". Signal Process Image Communication, Vol. 84, 2020, pp. 115831.

[11]. Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach." Information Fusion Vol. 46, 2019, pp. 184–192.

[12]. M. S. Hossain, G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data." Information Fusion Vol. 49, 2019, pp. 69–78.

[13]. H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset." IEEE Trans Affect Computer. Vol. 5, No. 4, 2014, pp. 377-390.

[14]. G. Deepak, L. Joonwhoan, "Geometric feature-based facial expression recognition in image sequences using multi-class." Ada Boost and support vector machines. Sensors Vol. 13, No. 6, 2013, pp. 7714 –7734.

[15]. J. Mira, K. Byoung Chul, N. Jae Yeal, "Facial landmark detection based on an ensemble of local weighted regressors during real driving situation." International Conference on Pattern Recognition (ICPR), 2016, pp. 2198-2203.

[16]. J. Mira, K. Byoung Chul, K. Sooyeong, N. JaeYeal, "Driver facial landmark detection in real driving situations. " IEEE Trans Circuits System Video Technology, Vol. 28, No. 10, 2018, pp. 2753-2767. https://doi.org/10.1109/TCSVT.2017.2769096

[17]. R. A. Khan, A. Meyer, H. Konik, S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images." Pattern Recognition Letter, Vol. 34, No. 10, 2013, pp. 1159-1168. https://doi.org/10.1016/j.patrec.2013.03.022

[18]. M. H. Siddiqi, R. Ali, A. M. Khan, Y. T. Park, S. Lee, “Human facial expression recognition using step wise linear discriminant analysis and hidden conditional random fields.” IEEE Transactions on Image Processing, Vol. 24, No. 4, 2015, pp. 1386–1398. https://doi.org/10.1109/TIP.2015.2405346

[19]. D. Ghimire, S. Jeong, J. Lee, S. H. Park, “Facial expression recognition based on local region specific features and support vector machines." Multimedia Tools Applications, Vol. 76, No. 6, 2017, pp. 7803–7821. https://doi.org/10.1007/s11042-016-3418-y

[20]. S. L. Happy, A. George, A. Routray, “A real time facial expression classification system using local binary patterns.” IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, 2012, pp. 1-5.

[21]. B. Hasani, M. H. Mahoor, “Facial expression recognition using enhanced deep 3D convolutional neural networks.” IEEE Conference, 2017, pp. 30-40.