2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) : March 17–18, 2021, Amity University Dubai, UAE

2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) | 978-1-6654-2921-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIKE51210.2021.9410768

Android based Emotion Detection Using Convolutions Neural Networks

Rabia Qayyum
Department of Software Engineering, MCS
National University of Sciences & Technology
Islamabad, Pakistan
rabiaqayyum7@gmail.com

Vishwesh Akre
Dubai Women's College
Higher Colleges of Technology
Dubai, UAE
vakre@hct.ac.ae

Talha Hafeez
Department of Computer Science
COMSATS University
Islamabad, Pakistan
talhahafeez666@gmail.com

Hasan Ali Khattak
Department of Computing, SEECS
National University of Sciences & Technology
Islamabad, Pakistan
hasan.alikhattak@seecs.edu.pk

Asif Nawaz
Department of Electrical Engineering
Higher College of Technology
Dubai, UAE
anawaz@hct.ac.ae

Sheeraz Ahmed
Department of Computer Science
Iqra National University
Peshawar, Pakistan
sheeraz.ahmad@inu.edu.pk

Pankaj Mohindru
Electronics & Communications Engineering
Punjabi University
Patiala, India
pankajmohindru@rediffmail.com

Doulat Khan
Department of Computer Science
Capital University of Science & Technology
Islamabad, Pakistan
doulatkhan77@gmail.com

Khalil ur Rahman (哈飞)
Deptt. of Comp. Sci. & Tech.
Northwest Normal University
Lanzhou, China
khalilkhan391@gmail.com

***Abstract*—With the advent of improved mobile processing capabilities we have seen many novel and useful applications. Among other usecases is the utilization of graphics capabilities of on-board computing capabilities. The study evaluates a new trend of functionality that has been considered in the emotion detection field. The proposed study uses Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to conduct a comparison of which deep learning technique works best for emotion recognition. Both neural network are trained using FER2013 dataset of Kaggle with seven emotion classes. The trained models are evaluated where CNN attains the accuracy of 65% and RNN lack behind with the accuracy of 41%. The trained models are then applied using music player based on one's facial expressions .The user gets the music according to the mood in two forms. Thus with the application user is provided with new and interactive way of getting the music that provides new and latest music and gets an entertaining music app. The Final Product has great scope as the end product can be modified and expanded where music recommendation can be exchanged with other recommendation systems like news, content etc. according to the emotion fetched.**

***Index Terms*—Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Convolutional Long Short Term Memory (ConvLSTM) Tensorflow, Emotion Detection, Android, Media Player.**

I. Introduction

People have different ways of communication they communicate through verbal or gestures or with written language but one thing that is common in all of this is emotion without emotion we cannot get the context of what they are saying or their clear meaning. Similarly, There are different ways of how emotion can be guessed or detected of the person like tone of voice, the punctuation's but the best one is the facial features and the body language they can tell the real emotions even if someone is lying. [1]

In the current era there is lot of development going on, on the following subjects where people are trying to make different methods on how to analyze emotions like text analysis, sound analysis etc. In our approach we are using facial feature extraction using CNN (Convolutions Neural Network) where different weights are assigned to different features and on the basis of those we get the result of the persons emotions as shown in Figure 1. This is just the basic sequence of events that follow in the process like image capture, image conversion using image processing, emotion detection through CNN. [2]

The proposed system basically scans a user photo/image and detects the face first and then the current emotion based on the detected mood. Convolutions neural networks are used to achieve the required functionality. CNN are trained using python scripts and thus the model is created on desktop and then finally transferred to mobile through tensor-flow lite. The neural networks trains in a way that it gets intelligent from the data-set provided e.g. data-set proposed system provide will be of facial images thus the network will be efficient enough to depict the mood of the user. The application is based on Artificial Intelligence and for android smart phones. The proposed system includes image processing. The system aims to approach teenagers and young adults who are fond of listening to music [3].

978-1-6654-2921-4/21/$31.00 © 2021 IEEE

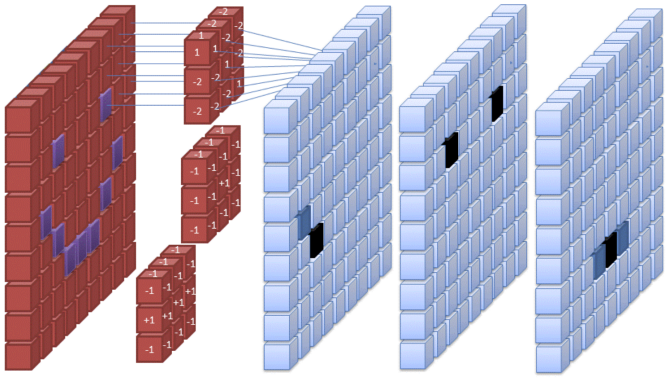

Fig. 1: Generic Overview of Convolution Neural Networks (CNN)

II. Background

There is a lot of work being done on CNN but its mostly on the side of desktop and cloud computing because of the heavy requirements of the neural network and such work requires good computational power as well thus this also caused a bit of motivation to explore a bit of this domain.

There is strict difference between the computational power of the cloud and desktop with respect to the android smart phone so we had to make certain adjustments and change the tools to fit the demand thus there are differences in the accuracy and speed of detection of emotion. In the desktop we can get the emotion in run time while in the android it takes about a few seconds. Also the neural networks can be used for different studies and nowadays artificial intelligence has a vast area of study, the proposed system cannot only be limited to detecting human emotion but using transfer learning the weights of the current model can be used to train on any other dataset and for any other or related problem and scope. Also the same emotion detection can also address other areas for example can provide news based on the emotion instead of limiting it to music player [4]. In this work we have used CNN in the beginning to get the required trained model and Linux is used for converting the model for android usage and prediction. The model is then used on android with including the libraries like Tensorflow lite for prediction from model [5].

III. Related Work

Emotion Detection and Recognition is an emerging field for research which is linked to Sentiment Analysis. Sentiment Analysis intends to detect various feelings from text such as positive, neutral, or negative, whereas Emotion Analysis is linked to the detection of types of feelings such as disgust, fear, anger, joy, and sadness [6].

Early studies [7]–[10] shows that to detect emotion you also need to be aware of the pitch of the voice in order to detect human emotion. The study gathered 3 different elements form the participants. 1)- Audio information, 2)- Video information 3)- both audio and video. This data is than combined and processed to give out the result. From this study we can safely assume that both video and audio contribute in human emotion. Hence our study reduced the initial data gathering to images which are easy to store and process and give accurate results.

The study and the analysis of computer-aided detection of facial expressions began in 1990s. Mase [8] further elaborated on the technique of optical flow for facial expressions recognition. Lanitis et al. [9] applied a different approach of using flexible appearance model in order to identify person facial expressions, genders and identities.

The parametrized models of image motion were used by Black and Yacoob [11]. These models were capable of tracking and calculating the non-rigid facial movement given to a rule-based classifier of diverse facial expressions.

The radial basis function networks and the optical flow were used by Rosenblum et al [12] to categorize facial expressions. Similarly Essa and Pentland [13] used an optical flow region-based approach. Otsuka and Ohya [14] used a hidden Markov model (HMM) and optical flow for expression recognition. Xam-Cognitive [2] detects the emotion of the user based on an image. It just tells the percentage of all emotions by using image of the user. This repo includes a Xamarin.Forms Demo (Droid and UWP) that is capable of integrating emotion, Face, and Vision Cognitive Services. This allows the real-time evaluation by processing the camera frame. Custom Rendered use is compulsory because Xamarin.Forms does not have a Camera Control.

Similarly like this there are cloud computing APIs that have same functions as they have the required computational power example of that being the Face API by Microsoft [3] this uses the cloud computation the user uploads its image and the model on the server detects the emotion and returns the confidence of the highest emotion calculated. The face api has a wider scope and this is just a part of it. The Face API now incorporates emotion recognition, returning the confidence for a set of emotions regarding each face in the image. These emotions are known to be globally and cross-culturally communicated with specific face expressions[1].

The third most used method for emotion recognition is from speech that uses speech and tone of wise using spectography to find the accurate emotion. This new technique acquires higher accuracy rate as compared to formally published results, limiting the latency at the same time. Speech inputs in this technique is processed in smaller segments which is not that much accurate as the sound segments last longer than that and may catch just the wrong fragment.

The EEG signals from DEAP dataset is used to detect emotion and proves to be a benchmark for emotion recognition by Tripathi et al. [17]. It uses two neural networks i.e. Deep Neural Network and Convolutional Neural Network.The study proves that neural networks can outperform in recogizing emotion using brain signals.

[1]https://www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning/

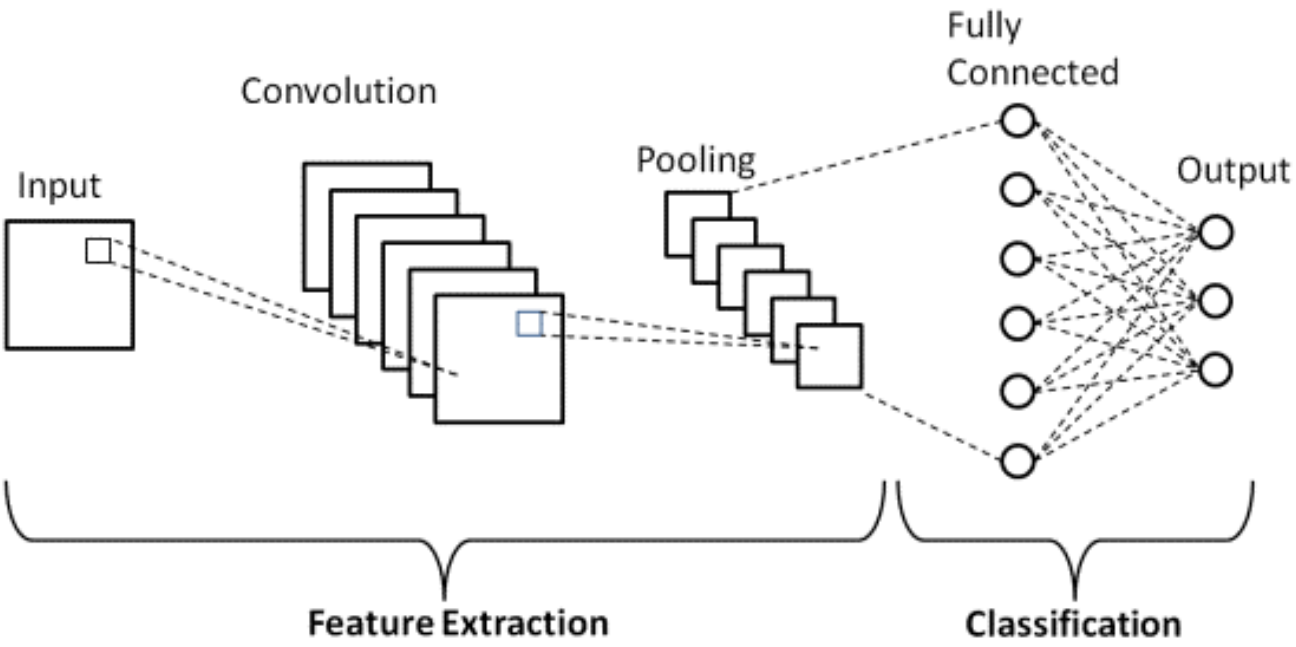

Fig. 2: General Architecture of CNN

Ensemble models using CNN are used by Knyazev et al [18] which achieves 60.03% classification accuracy. the models are trained using video dataset.Audio features are also used to complement the models with an additional modality.

Speech recognition is also helpful in emotion recognition where RNN are proving its success. Mirsamadi et al [19] proposed a solution using RNN that is evaluated using IEMOCAP corpus. The study shows RNN provide more accurate predictions compared to other existing emotion recognition algorithms in deep learning.

IV. Emotion detection for user-end applications

A. *Methods:*

Two methods are used for emotion detection i.e. Convolutional neural network and Recurrent neural network. Given below are the details of the algorithms:

1. Convolutional neural network: One of the most widely used deep neural network is the Convolutional Neural Network (CNN). As the name specifies that it performs mathematical linear operation between matrixes called convolution. CNN have multiple layers; convolutional layer, non-linear layers, pooling layer and fully-connected layer shown in Fig. 2. The convolutional and fully-connected layers have parameters but pooling don't have parameters to learn. The CNN has proven to be an excellent neural network in performance for deep learning problems. It is mostly used in applications that deal with image data, such as largest image classification data set (Image Net), computer vision, and in natural language processing (NLP) [20]. The results achieved were very amazing as compared to traditional machine learning algorithms.

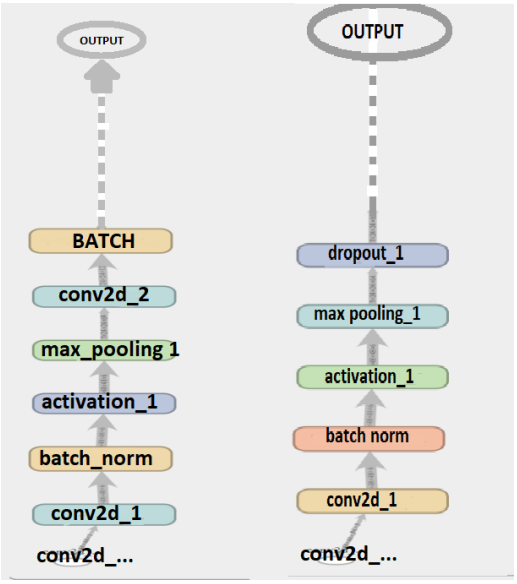

Fig. 3: CNN and RNN work flow

The model constructed for emotion detection using this type of network comprise of four convolutional layers, two fully connected, and four max pooling layers Fig. 3.

Convolutional layers: Convolutional layers perform a convolution over the input passed from the previous layer or new input from the dataset. The first convolutional layer constructs 64 filter of kernel size of 3x3 with the activation function as ReLU (rectified linear unit). The second convolutional layer has 128 filters with the kernel size of 5x5 and activated using

Classification Report

	precision	recall	f1-score	support
0	0.57918	0.55165	0.56508	484
1	0.82051	0.57143	0.67368	56
2	0.54353	0.46016	0.49838	502
3	0.83207	0.83478	0.83342	920
4	0.50556	0.60768	0.55193	599
5	0.80046	0.78054	0.79038	442
6	0.58983	0.59386	0.59184	586
avg / total	0.65949	0.65617	0.65615	3589

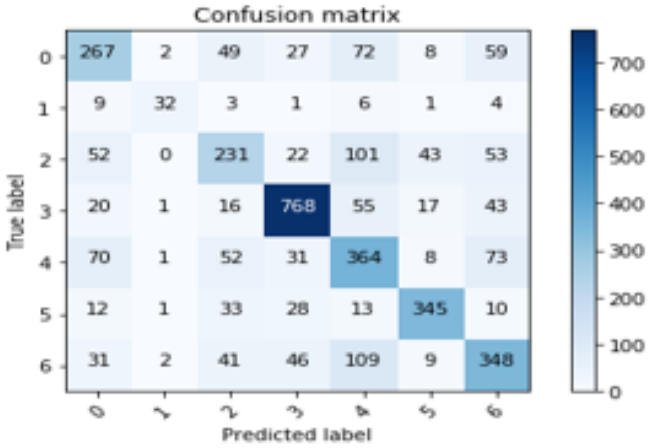

Fig. 4: CNN Report and Model

ReLU. The third convolutional layer has 512 filters with the kernel size of 3x3 and activated using ReLU. The fourth convolutional layer has 512 filters with the kernel size of 3x3 and activated using ReLU. More diverse representation of the input can be calculated by using various types of filters applied to the input. Each convolutional layer consist of dropout layers which deactivates 25% of total neurons in the layer so that the problem of overfitting during the training can be reduced.

Max Pooling Layers: The input is diminished by applying a maximum function on the max pooling layer. This layer presents translational invariance in accordance with the overall size of the filter used. Four max pooling layers are used and with the pool size of filter as 2x2.

Fully connected layers: This type of layer links all the neurons of former layers to each and every neuron of its own layers. The network architecture contains two fully connected layers. The first layer is created with the dimensionality of output space as 256 with activation function as ReLU and dropout layer which drops 25% of the neurons. The second fully connected layer contains the dimensionality of the output space as 512.

Output layer: The output layer works as one hot vector that denotes the class of the given image. The dimensionality of the layer is equivalent to the number of classes. The dimensionality of the output layer is 7 as the number of classes worked with is 7. The activation function used is Sigmoid. The sigmoid activation function ensures that the output neuron produces result as 1 if the output is greater than or equal to 0.5;if the output is less than 0.5 than it outputs 0.

2. Recurrent neural network:

For this method convolutional Long Short Term Memory neural net is used and it is hybrid of convolutional and recurrent neural network. Convolutional neural networks take into account previous training examples, same as the Time-Delay Neural Network, for context. The convolutionalLstm neural network is given images with no direct relation with each other with the pattern differences between previous images and their labels. The architecture for recurrent neural network (Fig.2) is as follows: Convolutional LSTM layers: The architecture contains three convolutional LSTM layers. All the convLSTM layers contains 10 filters with the kernel size 4x4 and time delay of 1 second. For all convLSTM layers the dropout layer is added which drops 25% of the actual neurons in each layer. The activation function used is sigmoid for each layer.

Convolutional Layers: The architecture contains only one convolutional layer with a single feature map as number of filters is 1 with the kernel size of 4x4 and activation function used as sigmoid.

Output Layer: This layer works same as one hot vector for representation of the class of the image as used in the architecture of convolutional neural network.

B. Data-set:

The data set used for emotion detection is used from Kaggle's facial expression recognition competition named fer2013. This data-set has been introduced by Pierre-Luc Carrier and Aaron Courville. The annotations consist of certain actions and emotions. The data-set consists of pixels of 35,887 facial images. The images are gray scale 48x48 in size. It is categorized into seven basic emotions categories i.e. 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. The emotions collected are of people of different age groups and mainly of same ethical background.

V. Results

A. Convolutional Neural Network:

The accuracy of model trained using convolutional neural network is 65%. To determine if the model architecture is efficient enough, it is tested using the fer2013 dataset's test images which are total of 3589 samples. The classification report and confusion matrix for the trained model are presented in fig.2 and fig.4.

Accuracy : 0.43466146558930063

Classification Report

	precision	recall	f1-score	support
0	0.33494	0.28719	0.30923	484
1	0.75000	0.16071	0.26471	56
2	0.27861	0.11155	0.15932	502
3	0.54147	0.75217	0.62966	920
4	0.28412	0.21202	0.24283	599
5	0.64543	0.52715	0.58032	442
6	0.34743	0.51877	0.41615	586
avg / total	0.41827	0.43466	0.40947	3589

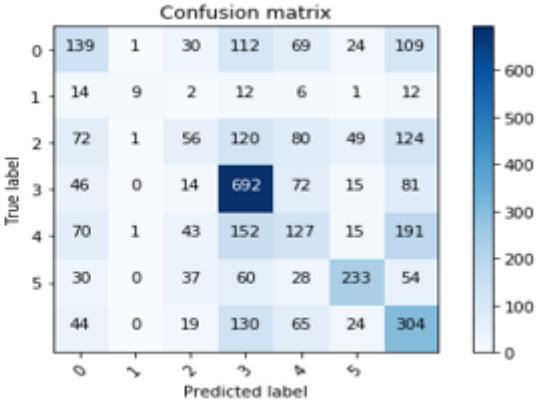

Fig. 5: RNN Report and Model

B. Recurrent neural network:

The calculated accuracy of the trained model using recurrent neural network is 41% using the test images in the fer2013 dataset which are 3589 samples. The classification report and confusion matrix of the trained model for the specific architecture are revealed in fig. 3 and fig. 5.

C. Comparison of Convolutional neural network and Recurrent neural network:

The architecture defined in both cases are trained and the models are overfitting that means the validation or test accuracy is less than training accuracy and the model works less accurate on new images. The model trained using convolutional neural network is more accurate on new images than the model trained using recurrent neural network. The table provides the validation accuracy for both the networks.

Neural Network Model	Validation Accuracy
Convolutional neural network	65%
Recurrent neural network	41%

TABLE I: Validation Accuracy For RNN and CNN

VI. Conclusion

The proposed work uses the power of a simple android phone and tries to attach it with good software structure using Convolutional neural networks combined with tensor flow lite to do the work being done on larger scale. The system tries to detect the emotion of the user by capturing its image on the runtime and based on that recommend and play music to the user from the two modes available online and offline. The research was increased in scope and compared with Retro neural network on android but the model only gave accuracy of about 45% in comparison to the CNN which gave an accuracy of about 67%.The product leads the way for many innovation and useful products. This work acts as a proof of concept and shows that there is still room for improvement.

The system uses a trained model using CNN (convolutional neural network) to detect mood of the user through taking the picture from camera of the device and music playlist is recommended using the detected mood of the user. The CNN is constructed using the library of google, Tensorflow which further simplified using Keras Library. The final product is of great scope as the end product can be changed and expanded according to user demand and also the system can be made more accurate using the user feedback. Similarly this method can also be used in the cars where a camera be placed in the rear view mirror and that can scan the emotion of the people in the car and play music automatically. Other Future applications of this systems can also be used for safety of users, thus detecting whether the drivers are asleep. More Dynamic application can be selection of news, music suggestions and in some cases food suggestions.

References

[1] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Deep android malware detection and classification." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[2] Khattak, H.A., Arshad, H., ul Islam, S., Ahmed, G., Jabbar, S., Sharif, A.M. and Khalid, S., 2019. Utilization and load balancing in fog servers for health applications. EURASIP Journal on Wireless Communications and Networking, 2019(1), p.91.

[3] GitHub. (2019). francedot/Xam-Cognitive-Demo. [online] Available at: https://github.com/francedot/Xam-Cognitive-Demo [Accessed 8 Jan. 2019].

[4] Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, Mujtaba G, Chiroma H, Khattak HA, Gani A. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access. 2019 May 22;7:70701-18.

[5] Azure.microsoft.com. (2019). Face API - Facial Recognition Software — Microsoft Azure. [online] Available at: https://azure.microsoft.com/en-us/services/cognitive-services/face/ [Accessed 4 Jan. 2020].

[6] Satt, Aharon et al. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." INTERSPEECH (2017).

[7] Ney, H., 1982. Automatic speaker recognition using time alignment of spectrograms. Speech Communication, 1(2), pp.135-149.

[8] Sahoo, S. (2018). Emotion Recognition from Text. International Journal for Research in Applied Science and Engineering Technology, 6(3), pp.237-243.

[9] Developer Blog. (2019). Emotion Detection and Recognition from Text Using Deep Learning - Developer Blog. [online] Available at: https://www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning/ [Accessed 2 Jan. 2019].

[10] Mase, K., 1991. Recognition of facial expression from optical flow. IEICE TRANSACTIONS on Information and Systems, 74(10), pp.3474-3483.

[11] Lanitis, A., Taylor, C. and Cootes, T. (1995. " A Unified Approach to Coding and Interpreting Face Images." in Proc. International Conf. on Computer Vision, 368-373 .

[12] De Silva, L.C., Miyasato, T. and Nakatsu, R., 1997, September. Facial emotion recognition using multi-modal information. In Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. (Vol. 1, pp. 397-401). IEEE.

[13] Black, M. and Yacoob, Y.(1995) "Tracking and Recognizing Rigid and Non-rigid Facial Motions Using Local Parametric Models of Image Motion." in Proc. Int. Conf. on Computer Vision, 374-381

[14] Rosenblum, M., Yacoob, Y., and Davis, L. (1996) "Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture" IEEE Trans. On Neural Network, 7(5):1121-1138.

[15] Essa, I. and Pentland, A. (1997) "Coding, Analysis, Interpretation, and Recognition of Facial Expressions, IEEE Trans. On Pattern Analysis and Machine Intelligence" 19(7): 757-767 .

[16] Otsuka, T. and Ohya, J. (1997) "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences" Proc. Int. Conf. on Image Processing, 546-549.

[17] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshi Mittal, and Samit Bhattacharya. 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 4746–4752.

[18] Knyazev, Boris, et al. "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video." arXiv preprint arXiv:1711.04598 (2017).

[19] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

[20] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp.1-6, doi: 10.1109/ICEngTechnol.2017.8308186.