2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Emotion Detection using Visual Information with Deep Auto-Encoders

Siva Prasad Raju Bairaju
IIIT RKValley, RGUKT-AP
Email: sraju728@gmail.com

Sowmya Ari
IIIT RKValley, RGUKT-AP
Email: aarivalli23@gmail.com

Dr. Rama Murthy Garimella
Mahindra Ecole Centrale
Email: rama.murthy@mechyd.ac.in

***Abstract*—Facial Emotion Detection is a important task for the machines to understand the emotional changes in human beings. In this research paper, we utilize combination of Convolution Neural Networks and Auto Encoders to extract features for Facial Emotion Detection. We proposed three architectures which are trained on JAFFE (Japanese Female Facial Expressions) database and got considerable classification accuracy.**

***Index Terms*—Emotion, Auto Encoders (AE), Convolution Neural Networks (CNN), Data Augmentation , VGG Net-16.**

I. Introduction

Face of a human being is an important body part which plays a significant role in the human to human or human to machine interaction. Now a days Emotion Detection is receiving a lot of attention from researchers due to its potentials in improving human-Machine interaction such as social - welfare robots to monitor disabled people, human behaviour predictor and in ADAS (Advanced Driver Assistance Systems). Different surveys telling that facial expressions contribute 55% of the speaker information, among which verbal part contribution is 7% whereas vocal part is 38%. Even though plenty of dissimilar methods have studied this issue, recognizing facial emotions with considerable accuracy is still difficult.

In this paper, we introduced some architectures for facial emotion detection. Our models are implemented by taking different combinations of Auto Encoder (like traditional or convolution Auto-Encoder) and Convolution-Pooling layers , and also Auto Encoder and VGG Net-16 for feature extraction where each combination output is given to a fully connected layers for facial emotion classification.

In Neural Networks point of view, capability of a network to extract features from raw data to acquire their own knowledge is called Learning or learning is our means of attaining the ability to perform task. To avoid over-fitting in our models we applied techniques like dropout and Data Augmentation. ReLU (Rectified Linear Unit) layer is used to introduce non-linearity in our models and Soft Max function is used as the output of a classifier, to represent the probability distribution over '7' different facial emotions such as Happy, Surprise, Sad, Angry, Disgust, Fear, Neutral.

We discussed following sections in our research paper: In section-II, Related research work about facial emotion detection was discussed. Introduction about Auto-Encoders and different variations of Auto-Encoders was discussed in section-III . Section-IV contains our proposed models to detect facial emotions and Experimental results are discussed in section-V. Finally future Work is explained in section-VI whereas the paper concludes in section-VII.

II. Related Work

Recently, researchers have made considerable advancement in human facial emotion detection with Artificial Intelligence and Computer vision techniques. In twentieth century, research on facial emotions has began. In early 1970s, Ekman and Friesen, American Psychologists did an extraordinary work on facial emotions and they entrenched seven universal facial expressions: Happy, Sad, Surprise, Angry, Disgust, Fear, Neutral. And they implemented Facial Action Coding Systems(FACS) which was further used to categorize human facial movements by their appearance with the help of Action Units(AU). From this a new Facial Emotion Recognition era has began. In 2003, Ira Cohen and Nicu Sebe et al presented an architecture of Hidden Markov models(HMMs) for classifying expressions from video. Shan et al proposed a method for emotion detection using Boosted LBP(Linear Binary Patterns) descriptors in 2009. In Later research Pyramid Histograms Of Gradients(PHOG) are also used for Emotion Detection. In present days, Deep learning architectures like Convolution Neural Networks and Auto Encoders are used for feature extraction from an image. Firstly Liu et Al used 3D-CNN and a deformable facial action part model to locate facial action parts and learn part-based features for emotion categorization. In the year of 2016, Ali et Al, proposed a model which is a collection of boosted neural networks for Multi Ethenic facial emotion recognition. The results of any Deep Learning architectures mainly dependent on how the pre processing was done, appropriate Feature selection by the model and amount of data provided to train the network. There are enormous approaches which have been developed to design effective facial emotion detection system. In this research paper, we utilize both Auto Encoders and Convolution Neural Networks for Facial Emotion Detection.

III. Auto Encoders

The typical example of a representation learning algorithm is AUTO-ENCODER. An auto encoder acquires knowledge to compress data from the input layer into a short code, and then decompress that code into another representation that closely matches the original data. Simply an auto encoder is a feed forward and recurrent neural network that is trained to attempt to copy its input to its output. Internally, it has a hidden layer 'h' that describes a code used to represent the input 'x'. So, Auto-Encoders are successfully applied to dimensionality reduction, information retrieval tasks and Image Recognition etc.

$$Encoder function : h = f(x) \tag{1}$$

$$Decoder function : r = g(h) = g(f(x)) = x \tag{2}$$

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

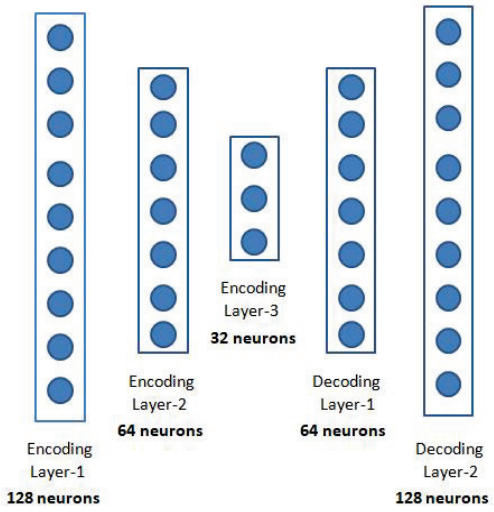

Fig. 1. Traditional Auto Encoder

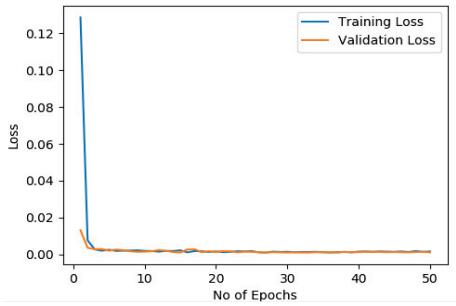

Fig. 2. TAE: Loss vs Epochs

Auto Encoder is an example of unsupervised learning technique. Since explicit labels are not giving for training the auto encoder. There are different variations of auto-encoders. Here, we are discussing two of them.

A. Traditional Auto Encoder

An auto encoder, in which Fully connected layers perform ENCODING and DECODING is defined as 'Traditional Auto Encoder (TAE)'. Architecture of traditional auto encoder is represented in Fig-1. In this paper our traditional auto encoder contains 3-encoding layers and 2-decoding layers whereas encoding layer-3 represents the feature of the given input image. We trained this traditional auto encoder with JAFFE database and it retrieve the input image with negligible loss which is represented in Fig-2.

B. Convolution Auto Encoder

Convolution auto encoder(CAE), one of the variation of an auto encoder, which contains convolution layers along with pooling layers acts as encoder whereas de convolution layers along with pooling layers acts as decoder and Fig-3 is the architecture of CAE.

We trained this auto encoder also, with JAFFE database by taking 3-Convolution layers and 2-Deconvolution layers. Conv-3 contains the feature which represents the input image and our proposed Convolution Auto Encoder retrieve the input image

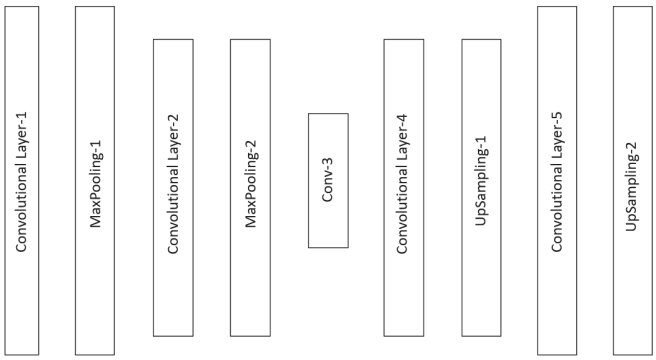

Fig. 3. Convolution Auto Encoder

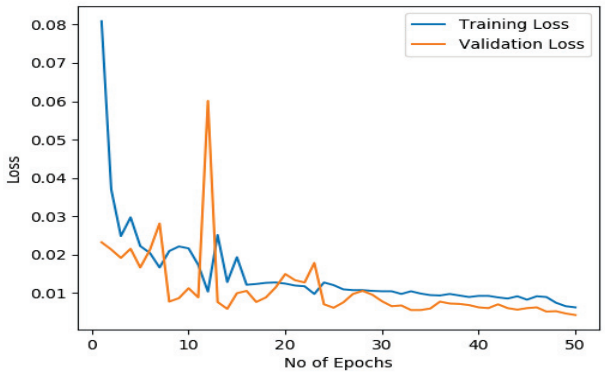

Fig. 4. CAE: Loss vs Epochs

with negligible loss. Fig-4 represents the graph between loss vs epochs.

C. Data set Preparation

We trained our models with JAFFE (Japanese Female Facial Expressions) data set which contains 213 gray images of 7 facial expressions posed by 10 Japanese female models. These JAFFE database images are static with 256 X 256 pixels. We applied Data Augmentation technique to generate extra amount of data with the existing data since classification accuracy of Deep learning architectures is mainly depends on amount of training data. Fig-5 contains 7 facial expressions of JAFFE database whereas Fig-6 is images after Data Augmentation.

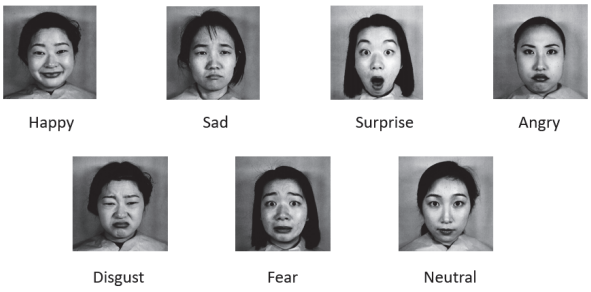

Fig. 5. 7 Universal facial expressions

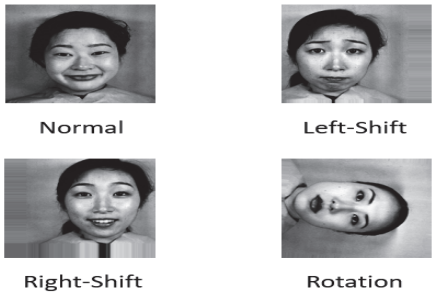

Fig. 6. Images after Data Augmentation

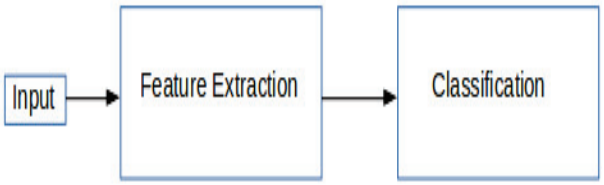

Fig. 7. Basic work flow of Neural Network classifier

IV. PROPOSED ARCHITECTURES

In this research paper, we proposed three architectures for emotion detection which are discussed below.

Basic work flow of neural network classifier was explained in Fig-7.

A. Series combination of Auto Encoder and Convolution Neural Networks

In this architecture, we have taken series combination of Traditional auto encoder (TAE) and convolution neural network (CNN) for the detection of facial emotion. Here, features which are extracted by the auto encoder is further given to CNN with an expectation that the network may learn the features more precisely or deeply to give better classification accuracy and we trained this architecture with JAFFE database after applying Data Augmentation technique. Since every deep learning architecture classification performance mainly depends on amount of training data, we applied Data Augmentation to raise the amount of data. Fig-8 contains the above proposed architecture.

B. Parallel combination of Auto Encoder and Convolution-Pooling layers

We have taken parallel combination of Traditional auto encoder and Convolution-Pooling layers for feature extraction and extracted features are fused together which are further given as input to the fully connected layers for classification of the facial emotions. And training is done as previous architectures. Fig-9 show the Parallel combination architecture.

C. Convolution Auto Encoder in cascade with fully connected layers

In this subsection, we introduced an architecture to recognize facial emotions and architecture is in Fig-10. Here, Convolution Auto Encoder is used as feature extractor and fully connected layers are used as classifier. We trained all these architectures and got considerable classification performance.

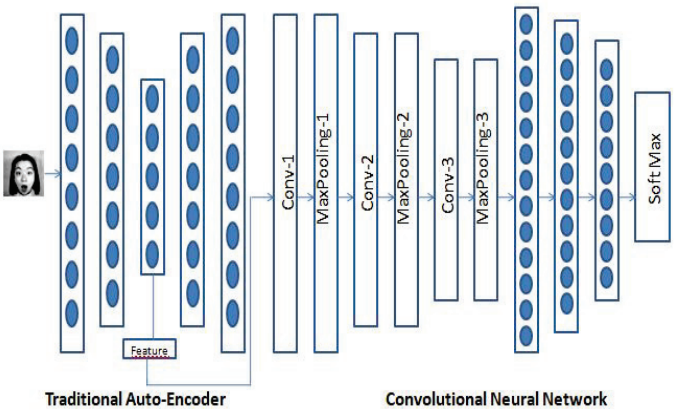

Fig. 8. Series combination of Traditional Auto Encoder and CNN

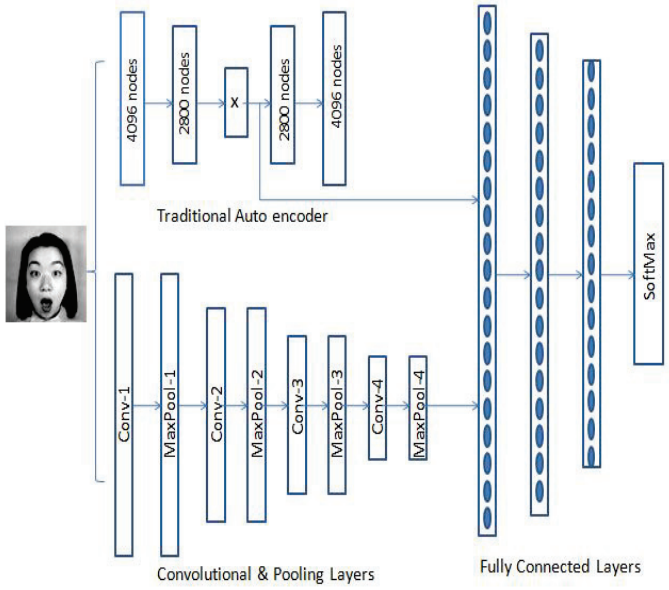

Fig. 9. Parallel combination of TAE and convolution-pooling layers

We implemented all the above architectures using Keras neural network library and trained with posed-emotion data set i.e. JAFFE. In these models we didn't used any benchmark architectures (such as VGGNet or Res Net) for feature extraction. But, to acquire much better classification accuracy we thought to use some base CNN like VGGNet-16 along with Auto Encoders for feature extraction and execution of our thought gave birth to below architectures (i.e. in Fig-11 and Fig-12).

D. Implementation of above proposed architectures using VGGNet-16 and Auto Encoders

In this section, we build new models of using the effectiveness of some neural networks to detect facial emotions.

1) Series combination of Traditional Auto Encoder and VGGNet-16: This architecture is present in Fig-11 in which the features extracted by the auto encoder is given as input to the VGGNet-16 for emotion detection.

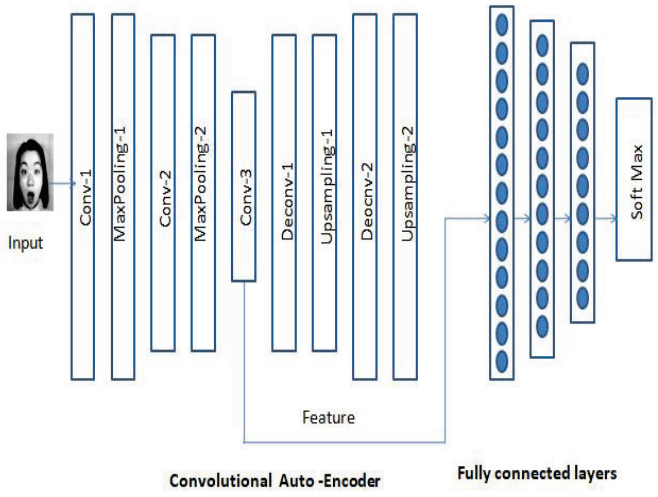

Fig. 10. Cascade of convolution auto encoder and Fully connected layers

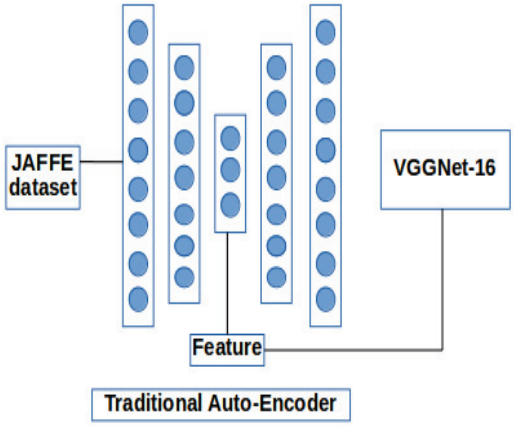

Fig. 11. Series combination of Traditional Auto Encoder and VGGNet-16

2) Parallel combination of Traditional Auto Encoder and VGGNet-16: In VGGNet-16, features extracted by Convolution-pooling layers is given to fully connected layers for classification. In our proposed model , features extracted by auto encoder and Convolution-Pooling layers of VGGNet-16 is fused together and then given to a fully connected layers for classification. Block diagram of this architecture is explained in Fig-12.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

We trained all the three architectures up to 50 epochs by applying data augmentation technique. We accomplished different classification accuracy for each architecture.

A. Series combination architectures

1) Accuracy - Series Combination of Traditional Auto Encoder and CNN : By training this series combination architecture which is in Fig-8, we experienced classification accuracy of 15-28% which was plotted in Fig-13.

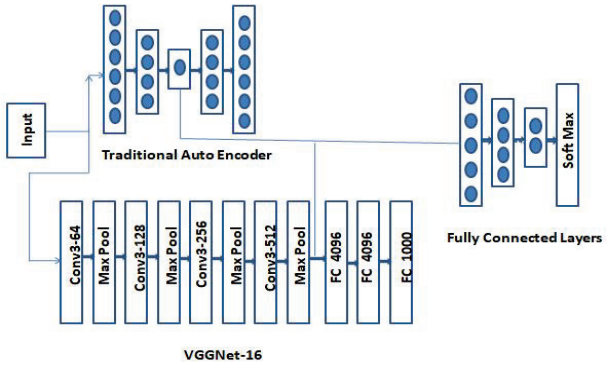

Fig. 12. Parallel combination of Traditional Auto Encoder and VGGNet-16

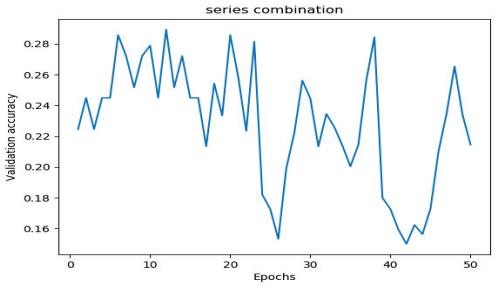

Fig. 13. Accuracy - series combination of Traditional Auto Encoder and CNN

2) Accuracy - Series Combination of Traditional Auto Encoder and VGGNet-16: Even though, utilization of benchmark CNN architecture (i.e. VGGNet-16) in series with TAE doesn't yield a considerable classification accuracy for facial emotions. We got classification accuracy of 11-15% for this architecture . Accuracy vs No.of Epochs graph was drawn in Fig-14.

From these results we can conclude that, series combination of Deep Learning architectures for feature extraction is not at all suitable for detecting facial emotions.

B. Parallel Combination Architectures

1) Accuracy - Parallel Combination of Traditional Auto Encoder and Convolution-Pooling layers: After training the parallel combination architecture which is in Fig-9, we got considerable classification performance i.e. between 30-51% which was drawn in Fig-15 and for each epoch it is taking 12-14 sec to train the network. For JAFFE database this is one of the best architecture for detection of facial emotions.

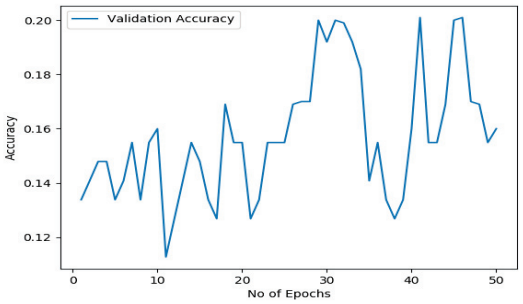

Fig. 14. Accuracy - series combination of TAE and VGGNet-16

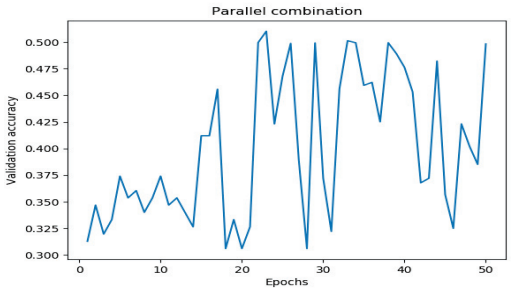

Fig. 15. Accuracy of parallel combination

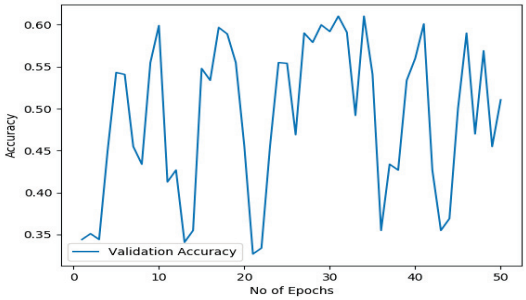

Fig. 16. Accuracy - parallel combination of Traditional Auto Encoder and VGGNet-16

2) Accuracy - Parallel Combination of Traditional Auto Encoder and VGGNet-16: We thought, for feature extraction, usage of benchmark Convolution Neural Networks instead of Convolution-Pooling layers may result better classification and we succeeded in this. After training this architecture we got classification accuracy of 35-60% which was drawn in Fig-16.

After observing these results , we came to a conclusion that parallel combination of Deep Learning architectures for feature extraction can perform well for classifying the given input.

C. Cascade of Convolution Auto-Encoder with Fully Connected Layers

Third architecture is a cascade combination of convolution auto encoder and fully connected layers which yields classification accuracy of 40-45% and accuracy results are drawn in Fig-17.

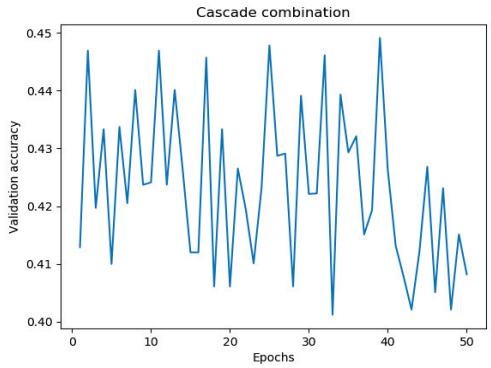

Fig. 17. Accuracy of cascade combination

VI. FUTURE WORK

In future, we are interested to implement a system which can detect emotions by considering Speech features and EEG features along with facial expression features.

VII. CONCLUSION

In this research paper, we introduced some interesting architectures for emotion detection using Deep Learning architectures. After training our models we conclude that, among all the proposed architectures, parallel combination of Deep Learning architectures evolve highest classification accuracy towards facial emotion detection whereas the series combination of Deep Learning architectures yield lowest classification accuracy.

VIII. REFERENCES

[1] S. Alizadeh, A. Fazel, "Convolutional Neural Networks for Facial Expression Recognition", CoRR, abs/1704.06756,2017.

[2] Ariel Ruiz-Garcia, Mark Elshaw, Abdulrahman Altahhan, Vasile Palade, "Stacked Deep Convolutional Auto-Encoders for emotion recognition from Facial expressions", 2017 International Joint Conference on Neural Networks(IJCNN).

[3] " Facial emotion analysis using deep convolution neural network ", G A Rajesh Kumar ; Ravi Kant Kumar ; Goutam Sanyal ," 2017 International Conference on Signal Processing and Communication (ICSPC) ".

[4] " Facial expression recognition combined with robust face detection in a convolutional neural network", M.Matsugu ; K.Mori ; Y.Mitari ; Y.Keneda, " Proceedings of the International Joint Conference on Neural Networks, 2003 ".

[5] " Facial smile detection using convolutional neural networks ", Dinh Viet Sang ; Le Tran Bao Cuong ; Do Phan Thuan , " 2017 9th International Conference on Knowledge and Systems Engineering (KSE) ".

[6] " Real Time Facial Expression Recognition Based on Deep Convolutional Spatial Neural Networks ", B Subarna ; Daleesha M Viswanathan, " 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR) ".

[7] Andre Teixeira Lopes et al, "A Facial Expression Recognition System Using Convolutional Networks", Vol. 00, pg. 273 – 280,2015.

[8] Arushi Raghuvanshi and Vivek Choksi, "Facial Expression Recognition with Convolutional Neural Networks", CS231n Course Projects, Winter 2016.

[9] " Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units ", Prudhvi Raj Dachapally ,Published 2017 in ArXiv.

[10] https://medium.com/@chrisprinzz/facial-emotion-detection-using-deep-learning-44dbce28349c