International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 4115–4123

FINET: Facial Expression Recognition Based on Fusion Inherited Network

Victor Mokaya[1], Mukesh Kr Gupta[2], Mukesh Bansal[3]

[1]*Department of computer science & information Technology, Suresh GyanVihar University, India*
[2]*Department of Electrical engineering, Suresh GyanVihar University, Jaipur, India*
[3]*Department of software engineering, Altran Private Limited, India*

Abstract

Machine learning models built from complex hand-crafted features and classification processes are challenging to design and arenotrobustized. Due to this fact, convolutional architecture is incorporated to automatically extract and classify class labels with high levels of accuracy. In this work we propose a lightweight network FiNet: fusion inherited network for universal facial expression recognition. FiNet consist of two fusion bocks for contextual feature extraction from salient regions. Block-1 uses a single convolution filter to capture and preserve the domain features. Block-2 dispenses unwanted features from the enriched inherited features by employing a single filter. The inherited results of block one and two are fused into block 3 which merge all the comprehensive region features for classification by discriminating for higher adaptability. The two-stage fusion network significantly allows the network to only conserve spatial features which increases the discriminative power of FiNet. Space computation complexity is achieved by limiting the number of parameters and incorporating smaller database size which proves the reliability of FiNet. FiNet size is 3.6MB as compared to VGG16:500.5MB, and ResNet 88.4MB. Comparative analysis on ResNet and VGG16 proved that the results of our proposed model outperformed existing futuristic models. The recognition accuracy results of our model on JAFFE and CK+ database was 84.37% and 96.62% respectively. The proposed model provides higher adaptability to lower computational power and storage as compared to traditional systems. FiNet size is 3.6MB compared to VGG16:500.5MB, and ResNet 88.4MB. FiNet provides an opportunity to be deployed in smart gadgets.

Keywords:*Facial Expression Recognition, fusion, space computation complexity, Feature Classification, FiNet*

1. INTRODUCTION

Expressions convey the most effective information about the mental process and intentions of the speaker during communication [1]. Nowadays with advancement of human computer interaction (HCI) it is necessary to monitor all the practical and essential activities about education, entertainment, psychiatric treatment, reviewing online shopping experience, and monitoring driver fatigue for long distance travels. At present HCI systems provide better and reliable performance results but their robustness still needs improvements to enhance efficiency. Currently the principal feature expression recognition (FER) approaches are: Action Units (AUs) [2] and appearance-based methods. Decoding facial landmarks reveals a composition of AUs; 12 on the upper and 18 on the lower face. They record the expansion or contraction of the facial muscles (eye, nose, mouth, burrows, lips, eyebrows, and eye corners). The appearance of the primary details however remains intact during this task. Appearance based methods largely rely on large image samples to detect multiple characteristics (face shape, eye color, mouth closed, mouth opened and skin color). Another class of approaches which incorporates hand crafted feature descriptors include Principal Component Analysis (PCA) [3], Discrete Cosine Transform (DCT) [4], Gabor features [5], Haar-like features [6] and Local Binary Pattern (LBP) [7]. After extraction, the above features are fed into a feature classifier in form of feature map or feature vector for classification. Nowadays with providence of computing power and cutting-edge technology, convolution neural network (CNN) have become the sort after techniques for computer vision tasks. Major applications include; object detection, face

ISSN: 2233-7857 IJFGCN

recognition, facial expression detection, haze removal, and anomaly detection among others. Implementation of CNN includes adjusting and updating the weights of the inputs in order to learn the sufficient class labels. Latest proposed models in literature include; a model similar to VGG16 by [8] was achieved through fine tuning the network parameters. A Weighted Mixture Deep Neural Network (WMDNN) [9] with two color channels for gray-scale and color images were proposed According to [10]an ensemble model for transfer learning using VGG16 and SVM classifier is proposed. In work done by [11], a novel Deep Attentive Multi-path CNN (DAM-CNN) is proposed which yields a dynamic image representation for FER. According to [12] a three VGG-net and Long Short-Term Memory (LSTM) [13] is proposed. The VGG-net extracts static and motive features whereby three types of attention mechanisms are jointly integrated for discriminative visual representation. The descriptive micro-expression features are fed into LSTM to extract spatial features for micro-expressions recognition. According to work by [14]a region-based pattern with an extensive index for response to emotions known as (RETRaIN) is developed.

2. PROPOSED METHOD

FiNet has been implemented by reducing the block size of the convolution layers and increasing the filter size. The blocks consist of down-sampling max pooling with filter size 2x2 to extract macro features comprehensively. VGG16-Net has been developed on deep dense networks suitable for computer visionapplications. However, deep dense has been unlikely to capture macro level features of expression locations due to larger convolution and pooling tasks. Therefore, to reduce space computation complexity a shallow lightweight network is proposed to extract adequate spatial information from facial expressions. Additionally, an embedded discriminative layer to enhance class category capabilities is incorporated.

Our proposed FiNet architecture has three blocks. The proposed architecture is as follows: the initial block has one 5x5 convolution layer and 6 filters over the input image to retrieve fine features (brows, burrows, eye lines etc.); the intermediate block of one 5x5 convolution layer and 16 filters extracts course feature (eyes, mouth, lips, chin, nose etc.). The resultant inherited features from layer one and two are fused together and fed to the final block with one 5x5 convolutionlayer and 32 filters to employ discriminative extraction of active feature regions. In each block except the last there exists a max pooling layer with filter size 2x2 to capture feature maps of highest pooling value. The two output layers consist of 84 neurons and a classification layer with 7 neurons to represent the 7 expression classes.The proposed model is shown in Table 1 below. Figure 1. Shows the flowchart of experimental procedure.

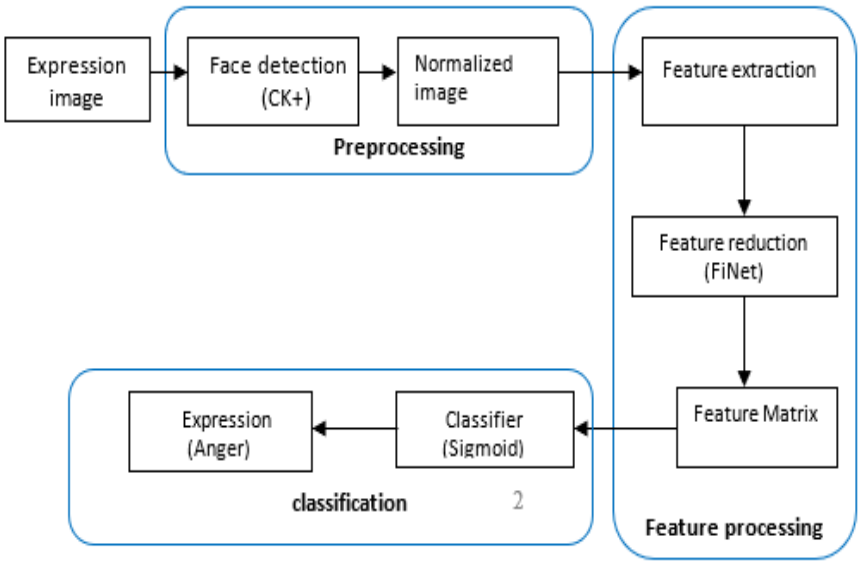


Fig.1.Flow chart of experimental procedure

ISSN: 2233-7857 IJFGCN

Table 1.The proposed FiNet network

Input image: 128 x128x1
Conv2d: 5x5x6 + ReLu
Maxpooling: 2x2
Conv2d:5x5x16 +ReLu
Maxpooling:2x2
Conv2d:5x5x32 +ReLu
FC:84 + ReLu
FC: 7

2.1. Pre-processing

In the pre-processing we normalized the images for saliency and to retain spatial features. Original image size 256 x 256 was cropped and scaled to 128 x 128 by elimination of ground influences. Further on we detected the eyes which were the salient regions from the cropped image. The face was further rotated by the angle between y-axis and x-axis, this was necessary inorder to preserve the salient regions as shown in figure.2 below.

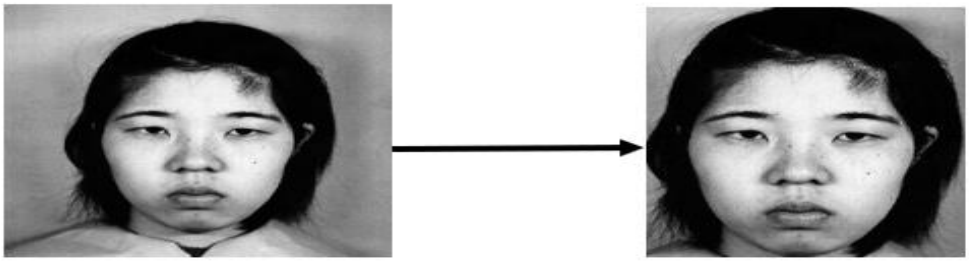


Fig .2.Pre-processed JAFFE dataset

2.2. Feature training

For training we minimized the cost function using batch gradient descent and modified the weights using the back-propagation error algorithm [15]**.** Batch size was set to 255 and the momentum of the weights 0.15. Drop out was utilized to prevent overfitting to the fully connected layers (except for the last output layer) with probability of 0.25. Experimental results prove the learning rate increased to the tens digit after every 5 epochs and remained constant at the 90^{th} epoch. The network was trained with 100 epochs. In the experiment the bias was initialized by zero and all weights randomly initialized with normal distribution of mean zero and standard deviation set to as in equation Eq. (1) below.

ISSN: 2233-7857 IJFGCN

$$\delta = \sqrt{\frac{2}{x}} \quad (1)$$

Wherex = weight size on each neuron, c.l (Convolution Layer) and f.c.l (Fully Connected Layer). c.l of x = (filter size) x (filter size) x (depth of previous layer), f.c.l of x = number of neurons from previous layer in each iteration, 255 images were injected into the network. After each iteration the training data was shuffled randomly. Cross-entropy was applied in the following pattern in equation Eq. (2)

- M represents the size of images in the training set
- E represents the size of emotion labels
- $y_n$is the one-hot encoding of the true class label of n^{th}image.
- $\hat{y}$ probability distribution of emotion classes of n^{th} image using SoftMax function.

Cross- entropy equation is given by:

$$z = \frac{1}{M}\sum_{n=1}^{M}\sum_{P=1}^{E} y_n\,(p)\log\left(\hat{y}_n(p)\right) \quad (2)$$

Where $y_n(p) \in (0,1)$ and p is the true label of n^{th}image.$\hat{y}_n\,(p) \in (0,1)$ represents probability is the true label of n^{th} image.

2.3. Feature testing

The dataset was grouped into two sets. During training highest accuracy levels were injected into the test set for validation. Further, the final validation set accuracy presented the recognition intelligence of our system for the emotion classes. The results of the predicted labels are shown in figure 2 and figure 3 below.

Further, we regularized the weights of each layer to limit size of the individual layer by adding a value to the hyperparameter. The given neuron result is represented by rectified linear unit y and the dropout possibility probability given by y,d in equationsEq. (3) and Eq. (4) below.

$$ReLu(y) = \max(0, y) \quad (2)$$

$$Dropout\,(y, d) = \sum_{y,prob,1-d}^{y,prob,d} y \quad (3)$$

3. RESULTS AND DISCUSSION

Our model results were analyzed with futuristic architecture built on similar database. The different techniques used for analysis of results were; LBP [7], LDP [16], and LDN [17], ResNet [18], and VGG16[8]. Current CNN models like VGG16 -Net and ResNet have recorded impressive classification results. However, in FER tasks these models fail to conserve discriminative features. Recent studies [19] confirm that deep neural networks fail to learn sufficient features over smaller datasets. Moreover, present FER benchmark datasets have limited sample sizes. Therefore, to overcome this challenge we developed a shallow and lightweight model to extract sufficient features with smaller parameters. FiNet network has eight million (8M) parameters while VGG-16:138M and ResNet 31M. Thus, FiNet has a higher adaptability to handle real world challenging tasks together with lower computational power as compared to the latter. FiNet weighs 3.6MB compared to VGG16:500.5MB, and ResNet 88.4MB. This provides FiNet a niche higher to be deployed in smart gadgets. The performance and accuracy of the proposed work on CK+ dataset was recorded in Table

ISSN: 2233-7857 IJFGCN

3. The comparison of our model was evaluated with high accuracy benchmark model over the same database LBP [7], Local Directional Pattern (LDP) [16] and Local Directional Number (LDN) [17]. Our model performed better with high accuracy of 96.62% against the other methods that were compared for 7 classes. The evaluation of our results was further performed on JAFFE and CK+. Figure 3. and Figure 5 below illustrates the train loss versus validation loss for JAFFE and CK+ dataset. Figure. 4. and Figure. 6. below Illustrates the training accuracy and validation accuracy for both JAFFE and CK+ database for 7 motions. The performance accuracy of our results was evaluated with other benchmarks. Our model attained 84.38% accuracy against other models as illustrated above in Table 3.

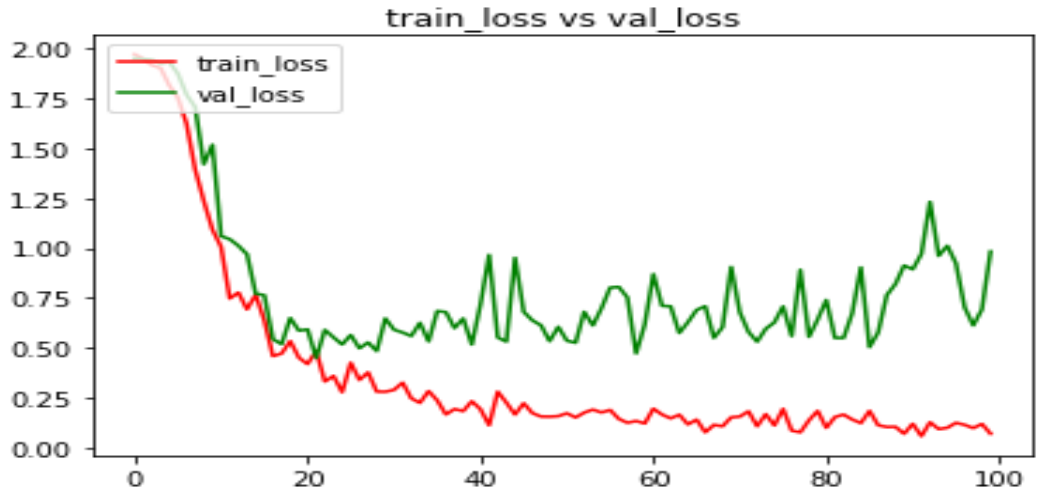
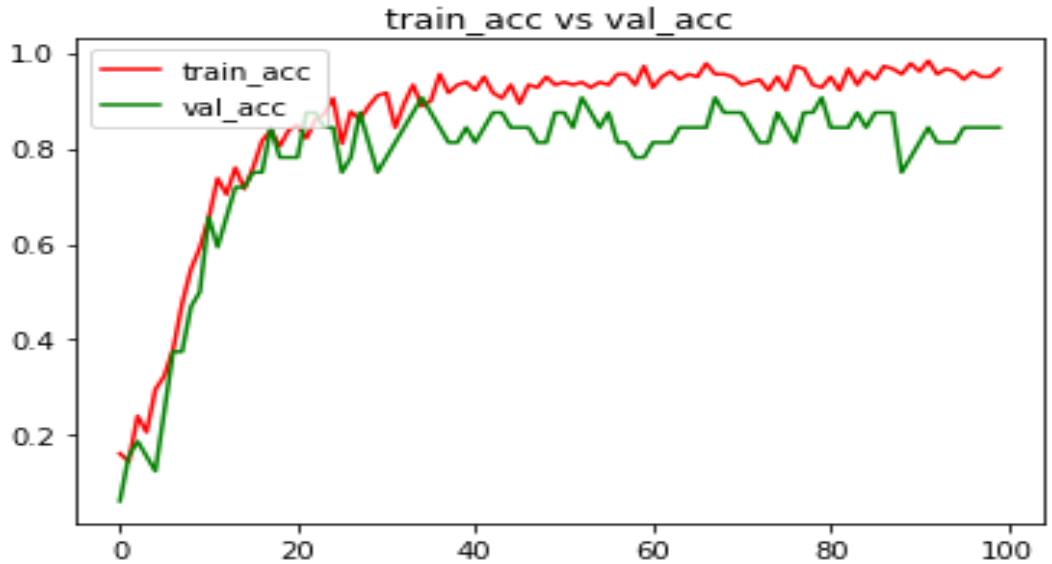


Fig.3. Training loss vs validation loss for JAFFE dataset



ISSN: 2233-7857 IJFGCN

Fig.4.Training accuracy vs Validation accuracy for JAFFE dataset

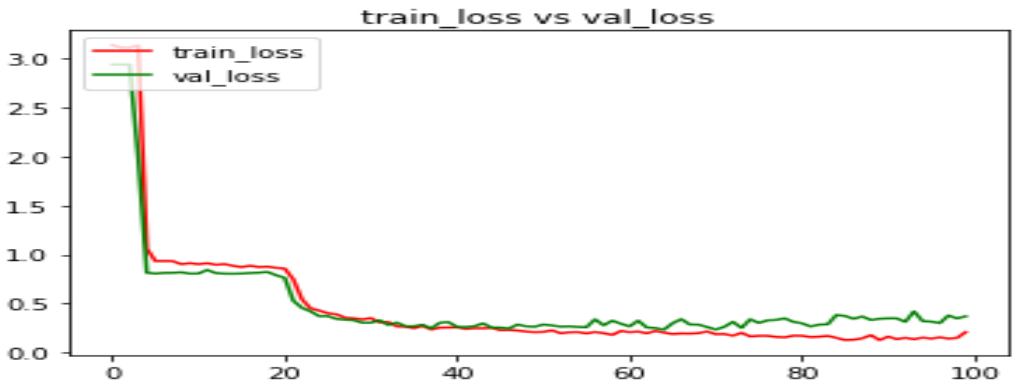


Fig.5. Training loss vs validation loss for CK+ dataset

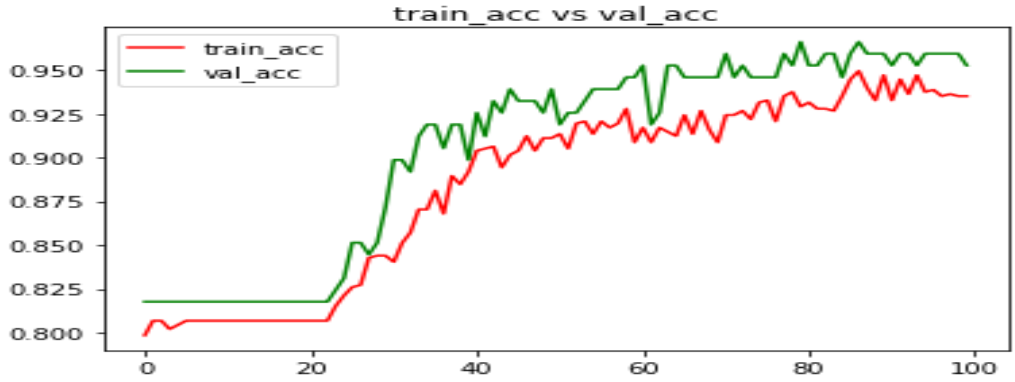


Fig.6. Training accuracy vs Validation accuracy for CK+ dataset

Table 2.Facial Expression Rate (%) of our model and other methods on JAFFE and CK+ for 7 classes.

Method	**7 class labels JAFFE**	**7 class labels CK+**
	Accuracy %	Accuracy %
LBP [7]	85.23	89
LDP [16]	86.19	92.3
VGG16[8]	85.1	95.2
ResNet[18]	85.6	91.8
LDN [17]	81.42	91.68
FiNet	**85.38**	**96.62**

ISSN: 2233-7857 IJFGCN

3.1. Extended Cohn Kanadedatabase

The CK+ database consists of 1043 facial expression images from 123 subjects of different age groups [20]. Figure7. shows samples from CK+ dataset. From these, 981 images were used with seven expression states: Anger, Happy, Sad, Neutral, Fear, Contempt and Surprise. The images were divided as follows: 80% training sample and 20% testing sample

Fig. 7.Sample of CK+ Dataset

3.2. JAFFE database

Our performance metrics were conducted using the JAFFE [21]. Inside the database were 213 peak emotions of the ten subjects in the dataset. In this dataset each subject comprises of six universal emotions (Anger, Happy, Sad, Contempt, Neutral, Fear and Surprise). Figure 8. Shows sample images from JAFFE dataset. In our model 198 images were used, 80:20 ratio was used for training and testing phases.

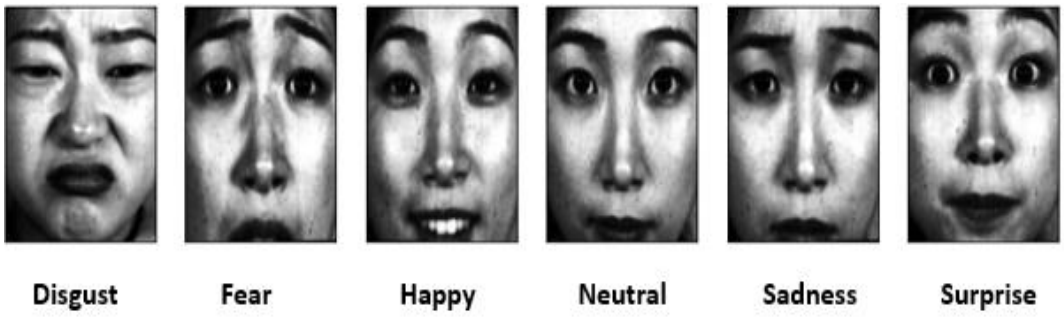


Fig.8**.** Sample of JAFFE Dataset

4. CONCLUSION

In this work a novel CNN architecture FiNet: fusion inherited network is proposed for macro facial expression recognition. We have proposed two blocks and single block of convolution layers fused together with filter size 5x5, to retrieve comprehensive and enriched features from salient regions. The weights of convolution filters are however 6,16 and 32 respectively for each block. FiNetutilizes the third block for enhanced discrimination of preservation of enriched features. The effectiveness of FiNet was examined on similar CNN variants for performance. The experimental setup was tested on two benchmark datasets Ck+ and JAFFE with profound classification accuracy results. We plan to enhance our work by using different transfer architecture to detect features from this network and SVM to categorize emotion classes from micro-features of facial expressions

ISSN: 2233-7857 IJFGCN

Acknowledgement
I wish to thank my supervisor Dr (prof) Mukesh Kr. Gupta for the immense support and guidance during the preparation of this work. The continued support and assistance from Dr Mukesh Bansal were also helpful to complete this work successfully.

REFERENCES

[1] Ekman, Paul.,"*Facial expression and emotion*". American psychologist, 1993.
[2] Emily Prince, Martin Katherine, Messinger Daniel, Allen Mike,"Facial action coding system". *The SAGE Encyclopedia of Communication Research Methods*, 2017
[3] Lever, Jake, Martin Krzywinski, and Naomi Altman, Points of significance: "*Principal component analysis*", Nature Publishing Group. 2017,
[4] Hayat, Khizar, and TanzeelaQazi. "*Forgery detection in digital images via discrete wavelet and discrete cosine transforms*." Computers & Electrical Engineering Vol.62,pp 448-458, 2017.
[5] Wu, T., Bartlett, M. S., &Movellan, J. R. "*Facial expression recognition using gabor motion energy filters*".IEEE computer society conference on computer vision and pattern recognition-workshops", pp. 42-47, 2010.
[6] Mohamed, B., Issam, A., Mohamed, A., &Abdellatif, B. "*ECG image classification in real time based on the haar-like features and artificial neural networks*", Procedia Computer Science, Vol.*73*, pp.32-39, 2015.
[7] Lin, J. H., Lazarow, J., Yang, A., Hong, D., Gupta, R., & Tu, Z., "*Local binary pattern networks*", The IEEE Winter Conference on Applications of Computer Vision pp. 825-834, 2020.
[8] Simonyan, Karen, and Andrew Zisserman. "*Very deep convolutional networks for large-scale image recognition."* arXiv preprint arXiv, 2014.
[9] Yang, B., Cao, J., Ni, R., & Zhang, Y, *"Facial expression recognition using weighted mixture deep neural network based on double-channel facial images"*. IEEE Access, Vol.*6*, pp.4630-4640. 2017.
[10] Nyein, T., &Oo, A. N. "*University Classroom Attendance System Using FaceNet and Support Vector Machine*". International Conference on Advanced Information Technologies (ICAIT), pp. 171-176, IEEE. 2019.
[11] Cha, Young. Jin, Wooram Choi, and Oral Büyüköztürk, "*Deep learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering*". 2019
[12] Yang, T., Wu, Y., Zhao, J., & Guan, L. "*Semantic segmentation via highly fused convolutional network with multiple soft cost functions"*, Cognitive Systems *Research*, Vol.*53*, 2019.
[13] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., &Schmidhuber, J. LSTM: "A search space odyssey". IEEE transactions on neural networks and learning systems, Vol.*28*, 2016.
[14] Verma, M., Vipparthi, S. K., & Singh, G "*Region Based Extensive Response Index Pattern for Facial Expression Recognition*". International Conference on Communication, Computing and Internet of Things (IC3IoT), pp. 20-24,2018.
[15] Cilimkovic, Mirza. "*Neural networks and back propagation algorithm*." Institute of Technology Blanchardstown, Blanchardstown Road North Dublin 2015.
[16] Rivera, A. R., Castillo, J. R., &Chae, O. O. "*Local directional number pattern for face analysis: Face and expression recognition"*. IEEE transactions on image processing, *22*(5), pp.1740-1752. 2012.
[17] Pillai, A., Soundrapandiyan, R., Satapathy, S., Satapathy, S. C., Jung, K. H., & Krishnan, R. "*Local diagonal extrema number pattern: A new feature descriptor for face recognition*", Future Generation Computer Systems, Vol. 81, pp.297-306,2018.
[18] He, K., Zhang, X., Ren, S., & Sun, J. "*Deep residual learning for image recognition*". Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770-778, 2016.
[19] Schindler Alexander, Thomas Lidy, and Andreas Rauber. "*Comparing Shallow versus Deep Neural Network Architectures for Automatic Music Genre Classification. in FMT"*. 2016

ISSN: 2233-7857 IJFGCN

[20] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. "*The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression*". IEEE computer society conference on computer vision and pattern recognition-workshops pp. 94-101. 2010.

[21] Lyons, M., Akamatsu, S., Kamachi, M., &Gyoba, J. "*Coding facial expressions with gabor wavelets*". In Proceedings Third IEEE international conference on automatic face and gesture recognition, pp. 200-205.1998.

ISSN: 2233-7857 IJFGCN