Volume: 8 Issue: VII Month of publication: July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30358

Emotion based Facial Recognition

Dr. Uma Maheshwaran[1], Gautam KK [2], Lavanya D[3], Enza John[4], Akshay Shyam Bhagwat[5]
[1, 2, 3, 4, 5]Computer Science Department, Visvesvaraya Technological University

Abstract: This paper tends to design an artificially intelligent system capable of emotion recognition through facial expressions of unknown people. The network in this paper consists of three convolutional layers each followed by max pooling and ReLU. The network is trained on FER2013 dataset and tested on RaFD dataset thus giving a wide range of training images to the network, so that it can overcome the basic problem of recognition of unknown faces. The pertinence of the final model is depicted in a live video application that can instantaneously return users emotions based on their facial posture. The accuracy obtained by this method was 68%, which is better than the previous state-of-the-arts methods. The results provide an important insight on the significance of using different datasets for training and validation. This paper mainly focuses on neural network based artificially intelligent systems capable of deriving the emotion of a person through pictures of his or her face. The primary research question being: "How can an artificial neural network, be used for interpreting the facial expression of a human?"

Keywords: Face Expression Recognition (FER), RaFD, ReLU, artificial neural network, FER2013 .

I. INTRODUCTION

Emotions are an important property of humans and are essential for effective interactions among the society. Humans communication can be either verbal of nonverbal, which it has been shown most of them refer to nonverbal communication. In nonverbal communication, emotion plays effective role because it conveys humans feeling about the subject, and in the psychology research it is proven that facial expressions is more effective then spoken word in conversation.

Facial expressions are one of the natural means to communicate the emotions and these emotions can be used in entertainment and Human Machine Interface (HMI) fields. Information from facial expression distributed in different area of face and each of them has different information so that mouth and eyes include more information that cheek and forehead. There were shown on several psychological studies which culture and environment can influence the impact of emotion and the way of expressing feeling for human beings. In many of these studies shown that gender, cultural background, age have bias in expressing emotion while there is not clear evidence on importance of environment for tendency the emotion.

Emotion recognition methods can be divided into two main groups: First group work on static images and second one work on dynamic image sequences. In the static approaches, temporal information is not considered and they just use current image information, while in the dynamic approaches images temporal information used in order to recognize expressed emotion in frame sequences.

Automatic emotion expression recognition include three steps: face image acquisition, feature extraction, and facial emotion expression recognition. In the optimal extracted features, within-class variations of expression should be minimum while between-class variations should be maximum. If the extracted features are not suite for task in hand and do not have enough information, even the best classifier may be unsuccessful to have best performance.

Feature extraction for emotion recognition can be divided into two approaches: Geometric feature-based methods and appearance-based methods. In the first methods, location and shape of parts of the face such as eyes, mouth, eyebrows and nose are considered, while in the second methods, particular regions or whole of face are considered.

Because of differentiating expressions' feature space is a difficult problem, so expression recognition is still a challenging task for computers. Some problems may be due to that, extracted features from two faces with equal expression may be different, while extracted features from one face with two expression may be equal, or some expression such as "fear" and "sad" are very similar.

In today's world, with the advancements in the areas of technology various music players are deployed with features like reversing the media, fast forwarding it, streaming playback with multicast streams.

Although these features satisfy the basic requirements of the user, yet one has to manually surf for the song from a large set of songs, according to the current circumstance and mood.

This is a time-consuming task that needs some effort and patience. The main objective of this work is to develop an intelligent system that can easily recognize the emotion through facial expression and accordingly play a music track based on that particular expression/emotion recognized.

II. IMPLEMENTTION

The implementation of the facial recognition system includes the following.

A. Dataset

Neural networks, DNN’s in particular, are known for their need for large amounts of training data. Moreover, the choice of images used for training is responsible for a big part of the performance of the eventual model. This implies the need for a both qualitative and quantitative dataset. For emotion recognition, several datasets are available for research, varying from a few hundred high resolution photos to tens of thousands smaller images. Out of which Face Expression Recognition Challange (FERC2013) dataset is selected as it is popular for FER system training and Radbound Faces (RaFD) datasets is selected for cross validation. The datasets differ mainly on quantity, quality, and 'cleanness' of the images. The FERC-2013 set for example has about 32000 low resolution images, where the RaFD provides 8000 high resolution photos. Furthermore it can be noticed that the facial expressions in the RaFD are posed (i.e. 'clean'), while the FERC-2013 set shows emotions 'in the wild'. This makes pictures from the FERC-2013 set harder to interpret, but given the large size of the dataset, the diversity can be beneficial for the robustness of a model. It is clear that, once trained upon the FERC-2013 set, images from 'clean' datasets can easily be classified, but not vice versa. Please note that non-frontal faces and pictures with the label contemptuous are taken out of the RaFD data, since these are not represented in the FERC-2013 training set. Furthermore, with use of the Haar Feature-Based Cascaded Classifier inside the OpenCV framework, all data is preprocessed. For every image, only the square part containing the face is taken, rescaled, and converted into an array with 48x48 grey-scale values.

B. Network

It consists four learned layers, three convolutional and one fully connected. Important features of the network are described in this section.

C. Training

The network is programmed with the use of the TFLearn library on top of TensorFlow, running on Python. This environment reduces the complexity of the code as neuronlayers are created instead of single neurons. The advantage of this set up is that we can get real-time feedback on the training progress and accuracy, which increases the reusability of the trained model.

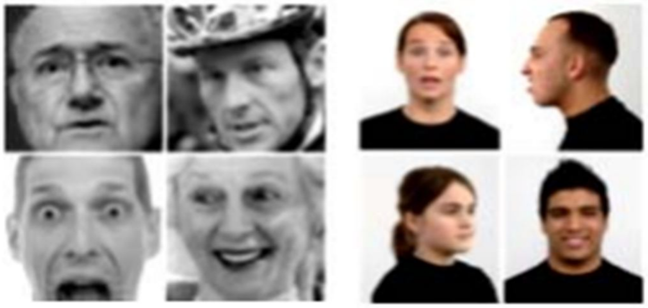


Fig 2.1 Samples from the FERC-2013 (left) and RaFD (right) datasets.

D. ReLU

The Rectified Linear Unit is widely being used as an activation function by Neural Network researchers in the past few years. The reason to this is that ReLU computes the activation function as: f (x) = max(0, x) that means the activation is threshold to 0. This helps in a faster convergence of stochastic gradient when compared to the sigmoid and tanh functions. This facilitates a faster learning.

E. Local Contrast Normalization

ReLUs do not require input normalization to avoid saturation. Even if a few training examples provide positive input to a ReLU, learning will occur in that neuron. Nonetheless local normalization scheme is very helpful in generalization. If i Ax,y is the activity of the neuron with kernel i and position (x,y), then the contrast normalized activity i Bx,y obtained by applying ReLU is given as;

$$B_{x,y}^{i} = A_{x,y}^{i} /(K + \alpha \sum_{j=\max(0,i-m/2)}^{\min(M-1,i+m/2)} (A_{x,y}^{i})^{2})^{\beta}$$

(1)

Where m= no. of adjacent kernel maps at a position (x,y) and M is the total no. of kernels in the layer. The constants K, m, Į and ǜ are hyper-parameters whose default values in TensorFlow are; K = 1, m = 5, Į = 1 and ǜ = 0.5.

F. Max Pooling

The pooling layer in a CNN is like a grid of pooling units with each pixel having a width p, summing up a q x q neighborhood centered at the pooling unit's location. When p is set as p<q we obtain max pooling.

G. Network Architecture

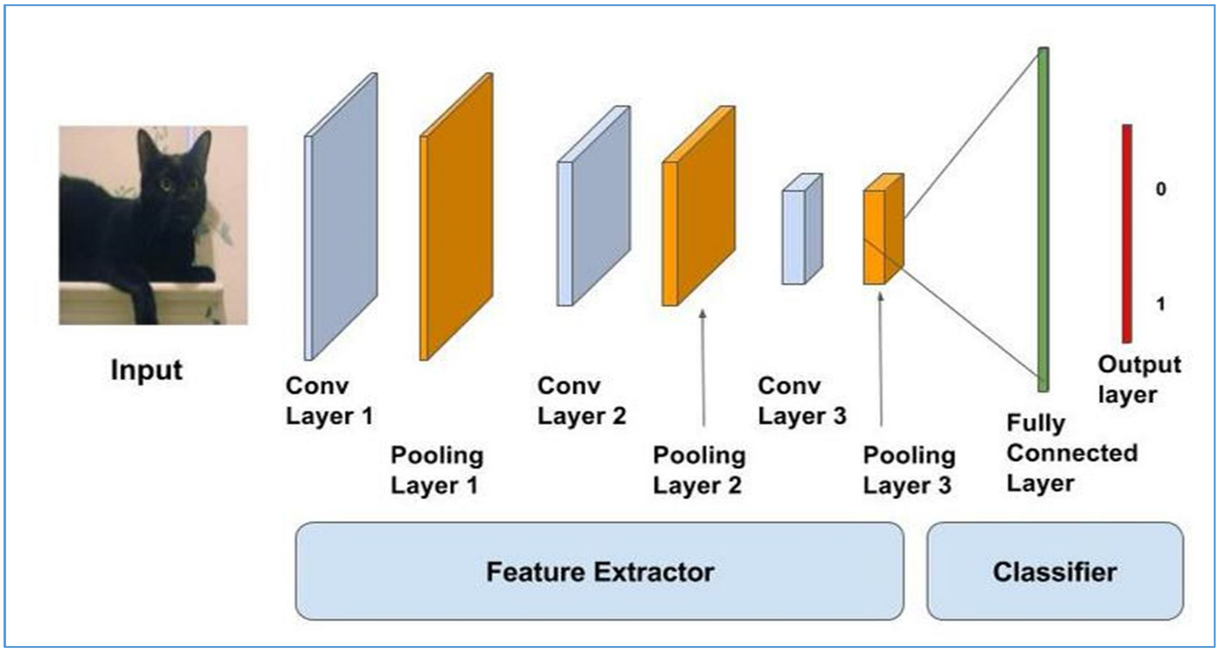
The network starts with an input layer of 48 by 48, matching the size of the input data. The next layer is a 5x5x64 convolutional layer with a kernel of size 5x5 and stride 1. The output of this layer is given to local contrast normalization layer, which is followed by a max-pooling layer. This is followed by another 5x5x64 convolutional layer which is connected to a max pooling layer with a stride of 2 and a 4x4x128 convolutional layer. Then finally a linear Fully Connected layer gives input to a softmax output layer. As we are classifying 7 emotions, 7 softmax units are applied mapping one class each. Dropout of probability 0.3 is applied to the fully connected layer to avoid over fitting. All the convolutional layers contain ReLU units.

III. PROPOSED SYSYTEM

In the proposed system, seven states of facial emotion are recognized by deep convolutional network which it includes three steps of feature learning, selection, and classification simultaneously. Training network with more than two layers was a difficult problem, but with the progress of GPUs, it is possible to train neural network with more than one layers. Deep neural network has three alternating types of layers which includes convolutional, sub-sampling and fully connected layers.

A. Convolutional Neural Networks

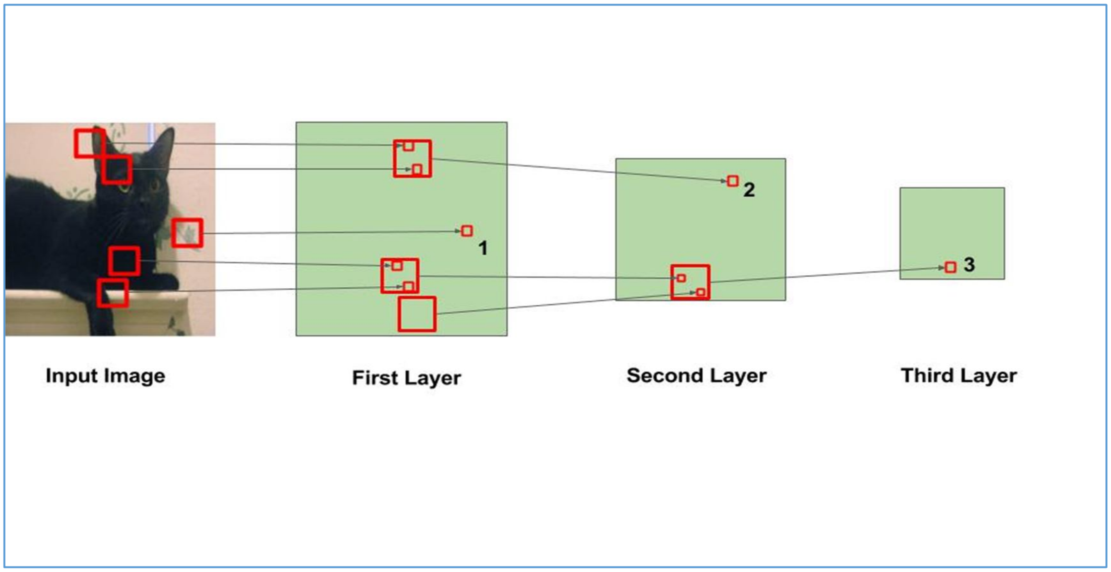
Convolutional Neural Networks are a form of Feed forward Neural Networks. Given below is a schema of a typical CNN. The first part consists of Convolutional and max-pooling layers which act as the feature extractor. The second part consists of the fully connected layer which performs non-linear transformations of the extracted features and acts as the classifier. In the diagram 3.1., the input is fed to the network of stacked Conv, Pool and Dense layers. The output can be a softmax layer indicating whether there is a cat or something else. You can also have a sigmoid layer to give you a probability of the image being a cat. Let us see the two layers in detail.



The convolutional layer can be thought of as the eyes of the CNN. The neurons in this layer look for specific features. If they find the features they are looking for, they produce a high activation. Convolution can be thought of as a weighted sum between two signals (in terms of signal processing jargon) or functions (in terms of mathematics). In image processing, to calculate convolution at a particular location (x,y), we extract k x k sized chunk from the image centered at location (x , y). We then multiply the values in this chunk element-by-element with the convolution filter (also sized k x k) and then add them all to obtain a single output.

CNNs can learn hierarchical features. In the below figure, the big squares indicate the region over which the convolution operation is performed and the small squares indicate the output of the operation which is just a number. The following observations are to be noted:

1) In the first layer, the square marked 1 is obtained from the area in the image where the leaves are painted.
2) In the second layer, the square marked 2 is obtained from the bigger square in Layer 1. The numbers in this square are obtained from multiple regions from the input image. Specifically, the whole area around the left ear of the cat is responsible for the value at the square marked 2.
3) Similarly, in the third layer, this cascading effect results in the square marked 3 being obtained from a large region around the leg area.



We can say from the above that the initial layers are looking at smaller regions of the image and thus can only learn simple features like edges / corners etc. As we go deeper into the network, the neurons get information from larger parts of the image and from various other neurons. Thus, the neurons at the later layers can learn more complicated features like eyes / legs.

B. Components of a Convolutional Neural Network

Convolutional networks are composed of an input layer, an output layer, and one or more hidden layers. A convolutional network is different than a regular neural network in that the neurons in its layers are arranged in three dimensions (width, height, and depth dimensions). This allows the CNN to transform an input volume in three dimensions to an output volume. The hidden layers are a combination of convolution layers, pooling layers, normalization layers, and fully connected layers. CNNs use multiple conv layers to filter input volumes to greater levels of abstraction.

CNNs improve their detection capability for unusually placed objects by using pooling layers for limited translation and rotation invariance. Pooling also allows for the usage of more convolutional layers by reducing memory consumption. Normalization layers are used to normalize over local input regions by moving all inputs in a layer towards a mean of zero and variance of one. Other regularization techniques such as batch normalization, where we normalize across the activations for the entire batch, or dropout, where we ignore randomly chosen neurons during the training process, can also be used. Fully-connected layers have neurons that are functionally similar to convolutional layers (compute dot products) but are different in that they are connected to all activations in the previous layer. More recent CNNs use inception modules which use 1×1 convolutional kernels to reduce the memory consumption further while allowing for more efficient computation (and thus training). This makes CNNs suitable for a number of machine learning applications.

C. Max Pooling Layer

Pooling layer is mostly used immediately after the convolutional layer to reduce the spatial size (only width and height, not depth). This reduces the number of parameters, hence computation is reduced. Using fewer parameters avoids overfitting. Overfitting is the condition when a trained model works very well on training data, but does not work very well on test data. The most common form of pooling is Max pooling where we take a filter of size and apply the maximum operation over the sized part of the image.

In the figure below the Max pool layer with filter size 2×2 and stride 2 is shown. The output is the max value in a 2×2 region shown using encircled digits. The most common pooling operation is done with the filter of size 2×2 with a stride of 2. It essentially reduces the size of input by half.

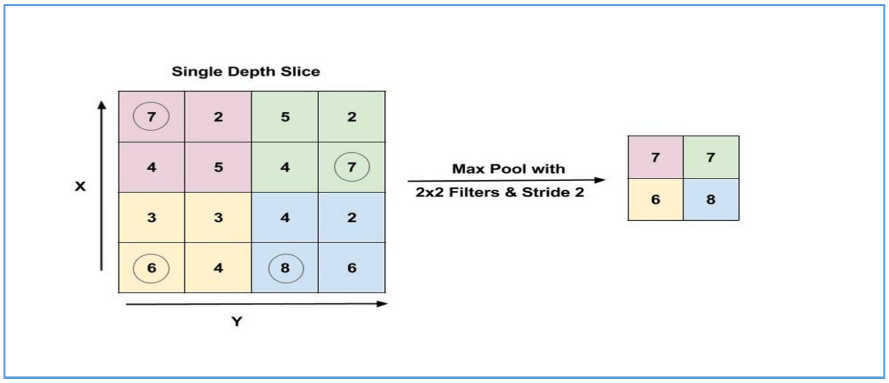


Fig. 3.3. Max Pooling layer

D. Emotion Recognition and Playing Music

To detect the actual emotion on the face we could use the above model. Models trained on a single individual, work much better when used on the same individual, often because in that case there is less variance between the data (here: facial features). If we minimize the variance by keeping the face the same, most of the detected differences will be due to the fact that a different emotion is expressed. The last and the most important part of this system is the playing of music based on the current emotion detected of an individual. Once the facial expression of the user is classified, the user's corresponding emotional state is recognized. A number of songs from various domains pertaining to a number of emotions is collected and put up in the list. Each emotion category has a number of songs listed in it. When the user's expression is classified, songs belonging to that category are then played.

IV. RESULTS

An overview of the network architecture is shown in Figure 4.1. The network was trained with learning rate of 0.001for 100 epochs in the final run, to make sure accuracy converges to the optimum. The network training was done with 20000 pictures from the FERC-2013 dataset. The ratios of the emotions --present in this set are given in Figure 4.1. Newly compiled validation set (20000 images) and test set (10000 images) from the FERC-2013 dataset are used together with the well balanced RaFD test set.

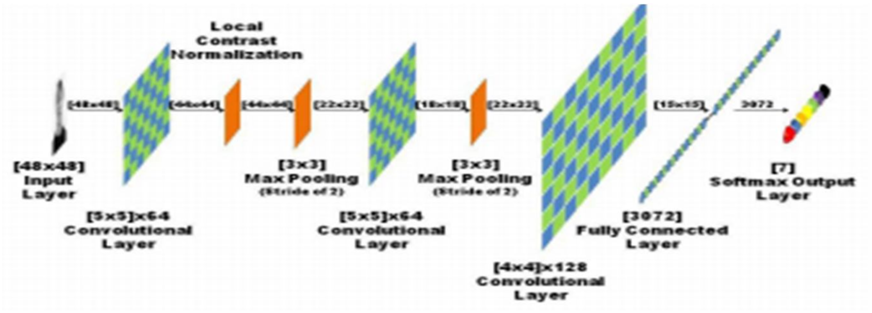


Fig 4.1 Network Architecture for the final model

On all validation and test sets the accuracy was 68%, underlining that more data and longer training can improve the performance of the network. It is to be noted that the accuracy on the RaFD test set, which contains completely different pictures than the training data is better 71%. This shows the effective generalizing capabilities of this final model. The performance matrix of the final model shows maximum accuracy for happy (90%) and minimum for sad (28%). It is to be noted that though the images for disgust emotion are less its accuracy is much better 62%. Live emotion recognition through video was performed with the help of developed application, which directly processes webcam footage through the final model. As mentioned above using the OpenCV face recognition program, the face from real-time video is tracked, extracted, and scaled to a usable 48x48 pixel input image. This data when fed to the input of the neural network model, returned values of the output layer. These values represented the likelihood of each emotion depicted by the user. The output with the highest value was assumed to be the current emotion of the user and was depicted by an emoticon on the left of the screen.

V. CONCLUSION

On all validation and test sets the accuracy was 68%, underlining that more data and longer training can improve the performance of the network. It is to be noted that the accuracy on the RaFD A three layered CNN for Emotion recognition problem was developed and its performance was evaluated using a live application. The training took 5 days on CPU, which is less considering the present state-of-the-art systems. The accuracy of the final model was 68%, which is better than the current FER systems. The results demonstrated that in spite of FER2013 training dataset having lesser images for disgusted label the emotion could be classified. Also the additional RaFD test dataset improves the performance of the network. The final model worked exceptionally well on unknown faces.

VI. ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance. So with gratitude we acknowledge all those whose guidance and encouragement served as beacon of light and crowned our effort with success. We are thankful to our Management for being a constant inspiration and providing all the facilities that needed throughout the Internship. We would like to thank our Principal Dr.P.Mahabaleshwarappa, for his constant guidance, advice and encouragement to complete the Internship. We consider it a privilege and honor to express my sincere gratitude to our beloved HOD, Mrs. Prameela Devi for her constant encouragement and all the support provided during this Internship. We convey our sincere thanks to our guide Dr. Uma Maheshwaran, Associate Professor, Dept. of CSE for his valuable guidance throughout the tenure of this Internship, and those support and encouragement made this work possible. It’s also a great pleasure to express my deepest gratitude to all my faculty members of my department for their cooperation and constructive criticism offered, which helped me a lot during my project work. Finally, we would like to thank all our family members and friends whose encouragement and support was invaluable.

REFERENCES

[1] Zhao, X., Zhang, S., 2016. A review on facial expression recognition : feature extraction and classification. IETE Tech. Rev. 33, 505–517.

[2] Poursaberi, A., Noubari, H.A., Gavrilova, M., Yanushkevich, S.N., 2012. Gauss Laguerre wavelet textural feature fusion with geometrical information for facial expression identification. EURASIP J. Image Video Process., 1–13

[3] Taylor, P., Siddiqi, M.H., Ali, R., Sattar, A., Khan, A.M., Siddiqi, M.H., Ali, R., Sattar, A., Khan, A.M., Lee, S., 2014. Depth camera-based facial expression recognition system using multilayer scheme. IETE Tech. Rev. 31, 277–286.

[4] Owusu, E., Zhan, Y., Mao, Q.R., 2014. A neural-ada boost based facial expression recognition system. Expert Syst. Appl. 41, 3383–3390.

[5] Biswas, S., 2015. An Efficient Expression Recognition Method using Contourletm Transform. Int. Conf. Percept. Mach. Intell. pp. 167–174.

[6] Cossetin, M.J., Nievola, J.C., Koerich, A.L., 2016. Facial expression recognition using a pairwise feature selection and classification approach. IEEE Int. Jt. Conf. Neural Networks, pp. 5149–5155.

[7] Happy, S.L., Member, S., Routray, A., 2015. Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. 6, 1–12.

[8] Hernandez-matamoros, A., Bonarini, A., Escamilla-hernandez, E., Nakano-miyatake, M., 2015., A Facial Expression Recognition with Automatic Segmentation of Face Regions. Int. Conf. Intell. Softw. Methodol. Tools, Tech. 529–540.

[9] Dahmane, M., Meunier, J., 2014. Prototype-based modeling for facial expression analysis. IEEE Trans. Multimed. 16, 1574–1584.

[10] Zhang, L., Member, S., Tjondronegoro, D., 2011. Facial expression recognition using facial movement features. IEEE Trans. Affect. Comput. 2, 219–229.

[11] Nigam, S., Singh, R., Misra, A.K., 2018. Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. Multimed. Tools Appl., 1–23

[12] Hegde, G.P., Seetha, M., Hegde, N., 2016. Kernel locality preserving symmetrical weighted fisher discriminant analysis based subspace approach for expression recognition. Eng. Sci. Technol. Int. J. 19, 1321–1333.

[13] Salmam, F.Z., Madani, A., Kissi, M., 2016. Facial Expression Recognition using Decision Trees. IEEE 13th Int. Conf. Comput. Graph. Imaging Vis., 125–130.

[14] Rashid, T.A., 2016. Convolutional neural networks based method for improving facial expression recognition. Intell. Syst. Technol. Appl. 73–84.

[15] Cui, R., Liu, M., Liu, M., 2016. Facial expression recognition based on ensemble of mulitple cNNs. Chinese Conf. Biometric Recognit. 511–518.

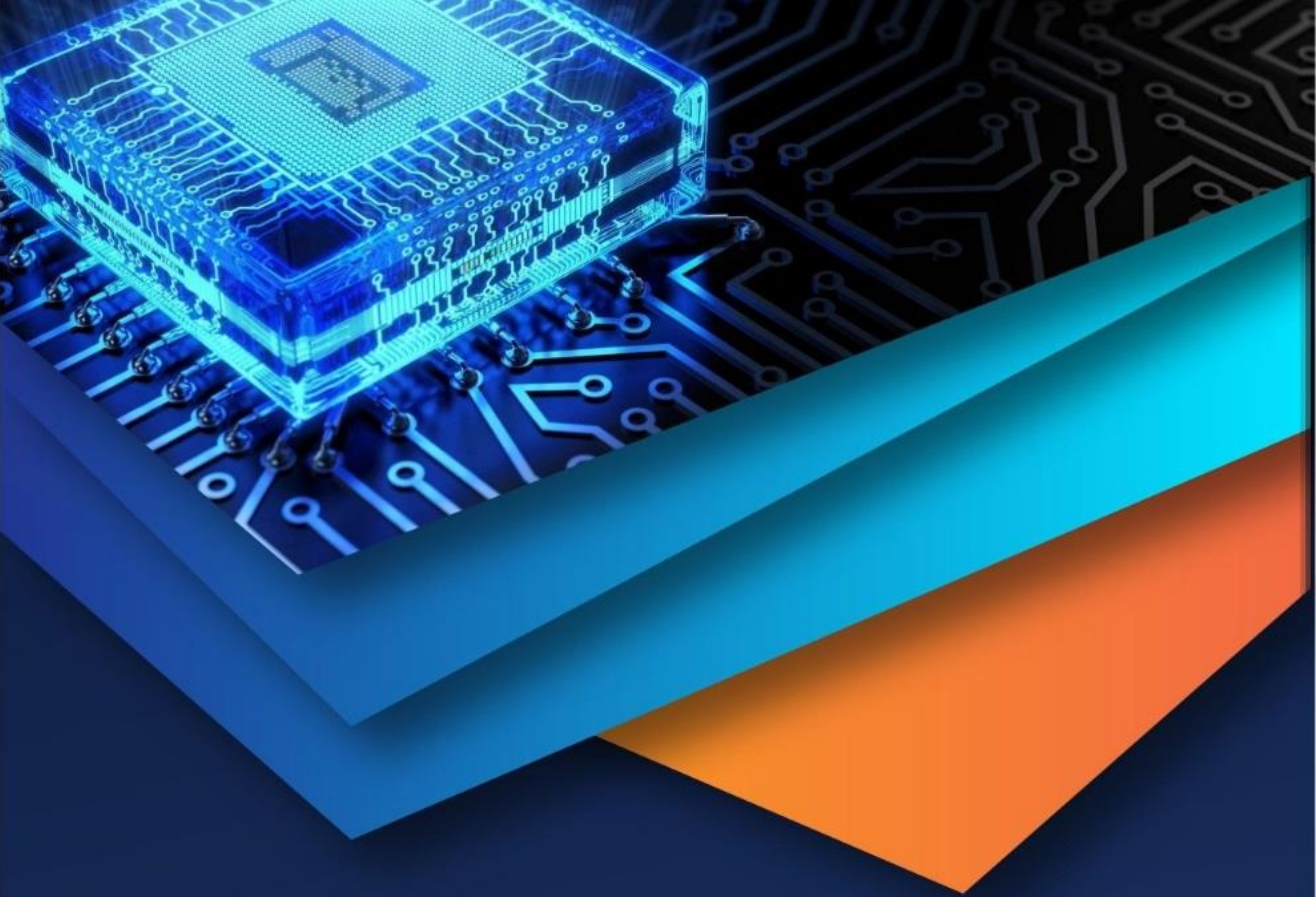
[16] Wu, Y., Qiu, W., 2017. Facial Expression Recognition based on Improved Deep Belief Networks. AIP Conf. Proc. 1864.

[17] Siddiqi, M.H., Ali, R., Khan, A.M., Park, Y., Lee, S., 2015. Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans. Image Process. 24, 1386–1398.

[18] Kumar, S., Bhuyan, M.K., Chakraborty, B.K., 2016. Extraction of informative regions of a face for facial expression recognition. IET Comput. Vis. 10, 567–576.

[19] Demir, Y., 2014. A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. Neural Comput. Appl. 27, 131–142.

[20] Zhao, G., Pietikäinen, M., 2009. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recognit. Lett. 30, 1117–1127.



10.22214/IJRASET

45.98

IMPACT FACTOR:
7.129

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY