Journal of Engineering Sciences

Vol 10, Issue 12, DEC/2019
ISSN NO:0377-9254

Facial Emotion Recognition using CNN

[1]A Jahnavi, [2]Shaik Taj Mahaboob
[1]PG Student, [2]Assistant Professor
[1]jahnavijanu520@gmail.com
[1]Electronics and Communication Engineering,
[1]JNTUA College of Engineering Pulivendula, Pulivendula, India

***Abstract*—**

With respect to its essential and commercial potentiality, a significant area in the computer vision as well as artificial intelligence fields is Facial emotion recognition (FER). Since the visual expressions with in interpersonal communication serves as the primary data channels, the study in which the facial images are mainly implemented is focused by this paper even if various sensors are used in order to conduct FER. The investigations performed from the previous years are provided in this paper. Initially, the description of the conventional FER methods and an overview of the descriptive FER systems classes in addition to their main algorithms are provided. With the help of deep networks that enables the "end-to-end" learning, the Deep-learning-based FER approaches are proposed. On behalf of the spatial characteristics of aparticular frame, an updated hybrid deep-learning methodthat combines a convolutional neural network (CNN)is focused in this paper for the sequential characteristics of consecutive structures. At the end, a short overviewis provided regarding the openly existing valuation parameters. Moreover, a comparison which is a standard on behalf of a quantitative comparison of FER researches is presented using the benchmark outcomes. The basic CNN variation is used by the existing work that is performed in order to generate the effective outcomes.A facial recognition system with the help of ALEXNET is proposed in this paper that produces enhanced outcomes compared to the basic CNN.

Keywords: FER (Facial emotion recognition), Deep learning methods, CNN (Convolutional neural network), Temporal features.

I. Introduction

The intentions of other people can be understood by the facial expressions that plays a significant role in human communication.Generally, the emotional conditions like happiness, sorrow and anger of others are inferred by the human beings with the help of facial expressions as well as vocal tone. The 1/3rdof the human communication is conveyed using the verbal components and the 2/3rdis by the nonverbal components. Facial expressions are the major datachannels that carrythe emotions in-between various nonverbal components. Thus, a special focus is gained by the facial expressions in the previous year's using the applications in perceptual, cognitive sciences in addition to the affective computing as well as computer animations.

Even though several sensors like electromyography (EMG) and electrocardiogram are present, arise in Interest in automated facial emotion recognition (FER) is observed with the fast growth of artificial intelligent procedures which includes human computer interaction (HCI), virtual reality (VR), augment reality (AR), advanced driver assistant systems (ADASs) as well as entertainment. Here, the FER is referred as the recognition of the facial expression which considers main features of facial emotion expression recognition.

On the basis of automatic FER, the investigations are primarily divided into 2 groups in terms of the features which are produced or handmade with the help of neural network outputs.

Figure 1 shows that the composition of FER is done using 3 main stages i.e., (1) detection of face as well as facial component, (2) extractionof feature as well as (3) classification of expression in conventional FER approaches. Initially, the detection of facial components (for example, eyes as well as nose) or landmarks is performed from the face region as well as the detection of the face imageis done from an input image. Later, the extraction of the several spatial as well as temporal features is done with the help of the facial components. Finally with the help of extracted

features, the recognition outcomes are produced by the pre-trained FE classifiers.

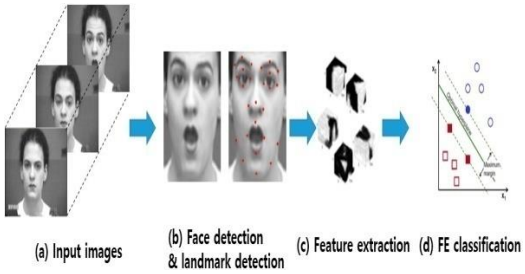


Fig.1. Process implemented in conventional FER approaches: using the input images (a), the detection of the face region and facial landmarks is performed (b), the extraction of the spatial and temporal features is done with the help of face components as well as landmarks (c), and on the basis of the facial categories, the determination of the facial expressionis done by pre-trained design categorizers(face images are taken from CK+ dataset) (d).

The development of the deep learning has yielded advanced outcomes in numerous computer visions as a common methodology towards the machine learning compared to the conventional methods with the help of handcrafted components.

The dependency upon face-physics-based models in addition to additional pre-processing methods is reduced to a greater extent using the Deep-learning-based FER methods in addition to the "end-to-end" learning is enabled for presenting within the pipeline immediately with the input images. The convolutional neural network (CNN) is a best known network models and a kind of deep learning technique within the available deep-learning models. A feature map is produced by convoluting the input image using a filter collection with the layers of the convolution in the CNN-based approaches. Later, the combination of every single feature map towards the networks which are connected completely and the recognition of the face expression that belongs to a specific category are performed with the help of the SoftMax algorithm output. The techniques implemented in the CNN-based FER approaches are shown in Figure 2. With respect to applications of the frame or video images, the distribution of the FER is performed into 2 classes. Initially, the execution of the static (frame-based) FER that depends upon the static facial features is done with the help of the extraction of the handcrafted features using particular peak expression frames concerning the image sequences. Finally in facial expression structures, the dynamic expressions are captured by dynamic (video-based) FER using the spatio-temporal features. Since an extra temporal data is provided, an increased recognition rate is present in the dynamic FER compared to the static FER.

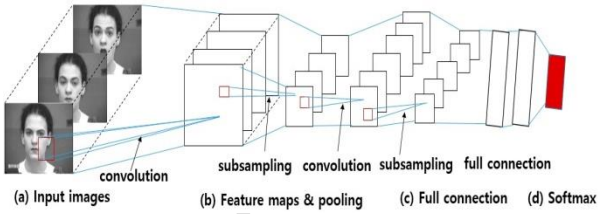


Fig.2. Procedure of CNN-based FER approaches: (a) the convolution of the input imagesis performed with the help of filters within the convolution layers. (b) The construction of the feature maps is done using the convolution results as well as max-pooling (sub-sampling) layers below the Spatial resolution of the feature maps assumed. (c) The entirely coupled neural-network layers are applied after the convolutional layers by the CNNs and (d) based on the SoftMax (face images are pickedup byCK+ dataset) output, the recognition of the single face expressionis performed.

II. Literature

Numerous conventional approaches were discussed for the automatic FER systems. The detection of the face region, extraction of the geometric components, appearance components, or a geometric combination and presence components upon the objective face is the similarity among these methods. A feature vector is constructed for the training [1] with the correlation among the facial componentson behalf of the geometrical features. On the basis of position as well as 52 facial landmark point's angle, implementation of two geometric features is done by Ghimire and Lee [2]. Initially, the calculation of the angle in addition to Euclidean distance amid every single pair of landmarks is performed in a frame. Later, the subtraction of the distance as well as angles from

the corresponding distance is done in addition to angles in the video sequence's initial frame. The presentation of the 2 approaches is done on behalf of the classifier i.e., with the help of multi-class AdaBoost using dynamic time warping or with the help of a SVM upon the enhanced feature vectors.

A histogram of a local binary pattern (LBP) that includes numerous block dimensions as of a universal face section by means of characteristic vectors is employed by Happy et al. [3] and a principal component analysis (PCA) is used within the classification of various facial expressions. Due to the inefficiency of reflecting the local variations of the facial components, the accuracyof recognition is often reduced despite of the implementation of this approach in the real-time.

With the division of the complete face region within the domain-specific local regions, the extraction of the extracted region-specific appearance is done in Ghimire et al. [4]. With the help of incremental search method, the determination of the significant local areas is performed that reduces the dimensions of the features and improves the accurateness of the recognition. The drawbacks of the 2 methods are complemented by combining certain methods [5] and greater outcomes are provided for the hybrid features.

Facial micro-expressions recognitionis proposed by Polikovsky et al. [6] that captures the video sequences by using a high speed device of 200 frames per second (fps). The histogram generation of the 3D-Gradients orientation is done using the movement within every single region on behalf of the FER once the face regions are divided within the particular regions. Due to the issues present 2D images as a result of the inherent variations in pose and illumination, the expression analysis research employs 3D, 4D (dynamic 3D) recordings in addition to FER of 2D images. In general, the extraction and classificationof features are mainly included in the recognition of 3D facial expression. Due to the type of information, there is a difference in the dynamic and static systems in 3D. Deformation model, active shape model, 2D representation's review as well as distance-based characteristics is extracted using the statistical models by the static systems.

Near-infrared (NIR) video sequences as well as the feature terms i.e., LBP-TOP (Local binary patterns from three orthogonal planes) are implemented by Zhao et al. [7]. The geometric as well as appearance face data are combined by implementing the component-based facial features. The classifiers that represent SVM and sparse are utilized for FER. With the extraction of difference of the horizontal as well as vertical temperaturesusing various face sub-regions, the infrared thermal videos are implemented by Shen et al. [8]. The implementation of the Adaboost algorithm is done that has feeble classifiers of k-Nearest Neighbor.

Depending upon the depth channel from the Microsoft Kinect sensor, the recognition of the facial expression and emotion is performed by Szwoch and Pieni ˛azek [9] where the camera is excluded. With the help of the relations among the specific emotions, the local movements within the face area is used by means of the feature as well as recognized facial expressions.

The detected face is tracked by the active appearance model (AAM) and the face region is detected using Kinect motion sensor on the basis of depth informationin Sujono and Gunawan [10]. The shape and texture model must be adjusted by AAM once a variation occurs in shape and texture when compared with the training outcome. Depending upon the prior information derived from FACS, The facial emotion is recognized by changing the significant features within the AAM and fuzzy logic.

With the help of color and depth data using Kinect sensor together, FER is implemented by Wei et al. [11]. The random forest algorithm is used for recognizing 6 facial expressions and the captured sensor data is used to extract the facial feature point's vector using the face tracking algorithm. In general, the features as well as classifiers are determined by the experts using the conventional methods. HoG, LBP, distance as well as angle relation amid landmarks are the handcrafted features in addition to SVM, AdaBoost, as well as random forest are the pre-trained classifiers which are employed in order to extract the feature and to recognize FE depending upon the features that are extracted. Compared to the deep learning-based methods,comparatively lower computing power as

well as memory is required by the conventional methods.

III. System analysis

Proposed method

The block diagram of the proposed method can be observed in the figure. 4. The training phase consisted of reading the input images, defining the layers followed by the options to the layers. Once the network is defined, the training of the images starts. This produces the model of the Deep learning. This model is used for recognition of the images in the testing face.

Convolutional Neural Network (CNN)

One of the deep learning neural networks is convolutional neural network (CNN). A great advance in the image recognition is represented by CNNs. The visual imagery is analyzed by them and they are responsible for the image classification in background. Based on the Facebook's photo tagging to self-driving cars, they can be identified. They are indirectly responsible entirely from healthcare to security.

The working of CNN is based on the feature extraction from the face emotions.The requirement of the physical feature extraction is eliminated here. The features are untrained and are informed when the network is trained by several faces. Due to this, the deep learning models are made highly precise on behalf of the responsibilities of computer vision. Over the tens or hundreds of the hidden layers, the detection of feature is learned by CNNs. The convolution of the features that are learned increases every single layer.

A. Classification with neural networks

Fig. 3 shows that Convolutional neural network (CNN) consists of 9 convolutional layers on behalf of the classification of the satellite image (remote sensing data). Using the similar image size and the satellite images (remote sensing data) of arbitrary size as the inputs, the outcome is produced by excluding some of the subsampling layersin FCN. Using various dimensions, the satellite image (remote sensing data) images can be processed by the network. Moreover, some additional post-processing for resizing the output towards the similar dimension as the input is not required by it. In fact, the proposed network is evaluated using various sizesbased on 2 standard datasets. The proposed network architecture is illustrated in Fig. 3. The entire data blobs within the architecture have similar height and width and a change is occurred in depth. During the entire CNN processing, no dimension is reduced.

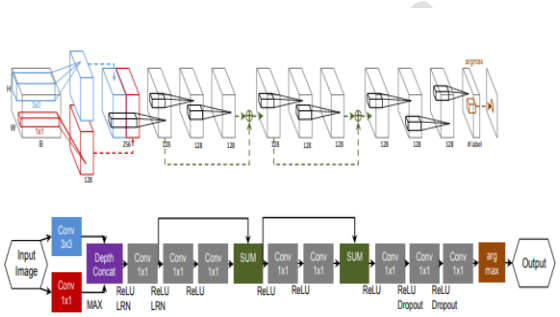


Fig.3: The input as well as output blobs of convolutional layers along with the connections are illustrated in the 1st row. The representation of the filters of every single convolutional layer is done below the output blob. Network flow chart is presented in the 2nd row.

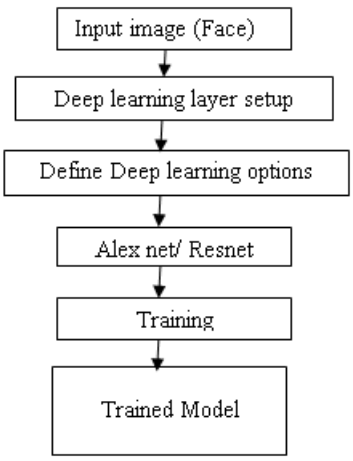
An inception module where an input image is convoluted using 2 convolutional filters of various dimensions (1×1×B and 3×3×B where, the quantity of spectral bands is represented by B) is used by application of the input satellite image (remote sensing data)by the first convolutional layer. Though the spectral correlations are addressed by the 1×1×B filters, thelocal spatial correlations of the input image are exploited by using the 3×3×B filters. Fig. 3 shows that a mutual feature map which is applied as an input for the layers of subsequent convolutions is formed by combining the first convolutional layer output as well as two convolutional feature maps. The convolutional filters with sizes above 3×3×B are not used in the 1st convolutional layer for preventing the limited spatial statistics of the input satellite image (remote sensing data) by overflowing. The nonlinear features are extracted from the joint spatio-spectral feature mapusing 1×1×B filtersby the subsequent convolutional layers. The deep networkis reduced by demonstrating the usage of 2 moduleswith the residual learning method. In terms of inputs of the layers, the responsibility of the residual learningis

observing the layers by help of the formula given below equation(1):

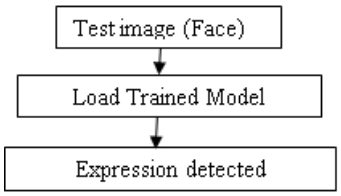
$$y = F(x, \{W_i\}) + x, \quad (1)$$

Where, the layer's input and output vectors considered is represented by x and y respectively. Function F represents the enduring mapping of convolution filters Wi that must be studied. The component-wise adding of F as well as x to the dimension of F along with x must be similar and is represented by the operation F + x.

Two convolutional layersare employed by the residual mappingin this architecture that is proposed. The initial layer in the module nonlinear is made by ReLU (Rectified Linear Unit). By considering the computational budget constraint, the network's depth and width are increased by proving the efficiency of the inception module and the residual learning. A few training samples are used in order to learn the deep network efficiently. The implementation of training data is reduced synchronously when ever a failure is occurred in the training the 7th and 8th convolutional layers. The fully connected layers of AlexNet in addition to the layer combination within the bottom most 3convolutional layers are similar. After every single inception module i.e., the 2nd, 3rd, 5th, 7th, 8th convolutional layers in addition to 2 residual learning components, the implementation of ReLU is done.LRN (Local Response Normalization) normalizes the first two convolutional layer's output. This method employs the training set system as shown in figure 4(A).



A. Training stage



B. Testing Stage

Fig.4. Proposed system black diagram

C. The Alex Net Architecture

A fundamental, simple in addition to efficient CNN architecture that includes the cascaded stages, such aspooling layers, convolution layers, rectified linear unit (ReLU) layers as well as completely connected layers is called as ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012) that is discussed in AlexNet by Alex Krizhevsky et al. in the 2012. The 1st layer, the 2nd layer, the 3rd layer and the 4th layer along with the pooling layer as well as the 5th layer along with 3entirely-connected layers are mainly included in AlexNet. The entire cost function is optimized using the stochastic gradient descent (SGD) procedure by extracting the convolution kernels for AlexNet architecture throughout the procedure of back-propagation optimization. In general, the convolved feature maps are generated by implementing the convolutional layers on the input feature maps using convolutional kernels that are descending. The data in the particular neighborhood window is aggregated using a max pooling process otherwise a mean pooling operation in order to operate the pooling layers upon the convolved feature maps. Compared to ReLU non-linearity layer as well as the procedure of dropout adjustment, AlexNet is widely applied in various practical strategies.

The training phase is accelerated and the over fitting is prevented by a half-wave rectifier function called as ReLUwhich is given in Equation (2). The co-adaptations of the neuronsis reduced by fixing several input or hidden neurons randomly in the dropout technique and it can be implemented in the AlexNet architecture with fully connected layers.

$$f(x) = \max(x, 0) \quad (2)$$

Transmission of CNN network parameters from nat ural image datasets to HSR remote sensing image d atasets is permitted through transfer methods and pr e-training method. Due to the resemblances among natural imagery datasets as well as datasets for

remote sensing of images along with the compatibility of the category, it can be achieved for some extent. A well-trained AlexNet architecture is obtained by analyzing the wide-ranging and complex ImageNet datasets. Moreover, the subsequent classification frame work is modified by some of the essential well-trained network parameters. Thus, the responsibility of HSR remote sensing imagery scene categorization can be performed by the Alex Net architecture with the help of a pre-training system. Architecture of the Alex Net is made by an throughout categorization channel by a pre-training mechanism, depending upon the suitable as well as wide-ranging demonstration capability of the Alex Net's pre-trained architecture compared to the HSR remote sensing imaging scene classification. Figure 5 illustrates the architecture of pre-trained AlexNet network.

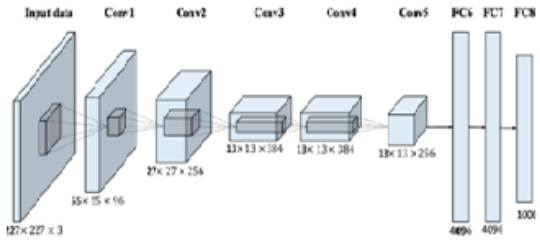


Fig.5. TheAlexNet architecture

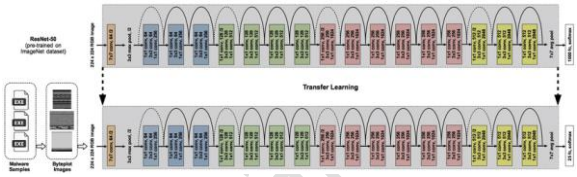


Fig.6. ResNet-50 layers that are pre-trained upon the ImageNet dataset are transmitted towards the DCNN model that replaces final 1000 completely-connected (fc) soft-max layer using a 25 completely-connected soft-max layer as well as disabling the convolutional layer parameters within the process of training.

C. ResNets50

Skipping the convolutional layer's blocks with the help of shortcut connections is the main concept of the deep convolutional networks called as Residual Networks (ResNets). (i)On behalf of the similar dimension of the output feature map, the equal amount of filters are present in every layer; and (ii) whenever, the distribution of the feature map size is done into two equal halves and the sum of filters is folded, are the designing rules followed by the blocks which are called "bottleneck". The convolutional layers with 2strides are used to perform the down-samplingand the normalization is implemented followed by every single convolution and earlier to the beginning of ReLU. The implementation of identity shortcut is done whenever the input as well as the output is considered with similar dimensions. The dimensions through 1 × 1 convolutions are matched by using the projection shortcut by increasing the dimensions. The execution of 2 strides is done once the shortcuts pass by the feature maps of two sizes in two situations. Using soft-max activation, a 1,000 fully-connected (fc) layer is used to conclude a network. 50 total numbers of weighted layers are used by trainable parameters of 23,534,592.

Fig.6 represents the original ResNet50 architecture. ResNet-50 is used by means of the basic model and pre-trained is used on behalf of the task of detecting the object upon the ImageNet dataset in this method. As the collection and interpretation of several malware samples continues to pose major tasks, the transmission of the ResNet-50 comprehensively trained on the ImageNet is assumed in spite of the difference among the natural images and malware within the plot images for making the malware image recognition tasks additionally efficient. The initial 49 layers of ResNet-50 are transmitted with the help of transfer learning techniques [11]that are abandoned upon the assignment of malware categorization. These layers are assumed as feature extraction layers. The bottleneck features are the learned features of the extraction layers which generates the activation maps. Fig. 6 shows that, as there are 25 classes, 1,000 numbers of completely connected soft-max are replaced by the trained 25 numbers of completely connected soft-max and a 25 numbers of completely connected soft-max is trained with the help of bottleneck features of malware byte plot images which are considered as input.

IV. Results

The experiments are performed on the FER 13. The FER 13 had images of size 48x48. Different sets of images have been considered for training and testing.

Table 1: Confusion matrix-CNN

Journal of Engineering Sciences
Vol 10, Issue 12, DEC/2019
ISSN NO:0377-9254

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
Angry	**14**	0	3	1	2	2	3	56%
	56%	0.0%	12%	4%	8%	8%	12%	44%
Disgust	2	**15**	5	1	1	1	0	60%
	8%	**60%**	20%	4%	4%	4%	0.0%	40%
Fear	3	2	**10**	3	1	4	2	40%
	12%	8%	**40%**	12%	4%	16%	8%	60%
Happy	0	1	1	**23**	0	0	0	92%
	0.0%	4%	4%	**92%**	0.0%	0.0%	0.0%	8%
Neutral	2	3	3	3	**9**	4	1	36%
	8%	12%	12%	12%	**36%**	16%	4%	64%
Sad	2	0	4	3	0	**16**	0	64%
	8%	0.0%	16%	12%	0.0%	**64%**	0.0%	36%
Surprise	0	0	5	2	0	0	**18**	72%
	0.0%	0.0%	20%	8%	0.0%	0.0%	**72%**	28%
								60%
								40%

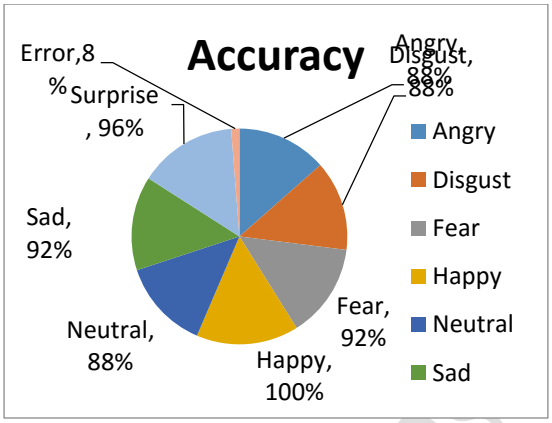


Fig.8. Alexnet pie chart

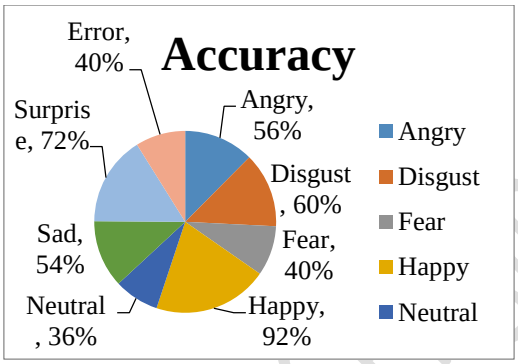


Fig.7. CNN pie chart

Table 2: Confusion matrix Alexnet

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
Angry	**22**	0	0	0	2	1	0	88%
	88%	0.0%	0.0%	0.0%	8%	4%	0.0%	12%
Disgust	1	**22**	2	0	0	0	0	88%
	4%	**88%**	8%	0.0%	0.0%	0.0%	0.0%	12%
Fear	0	0	**23**	0	1	1	0	92%
	0.0%	0.0%	**92%**	0.0%	4%	4%	0.0%	8%
Happy	0	0	0	**25**	0	0	0	100%
	0.0%	0.0%	0.0%	**100%**	0.0%	0.0%	0.0%	0.0%
Neutral	1	0	1	0	**22**	4	1	88%
	4%	0.0%	4%	0.0%	**88%**	16%	4%	12%
Sad	0	1	0	0	1	**23**	0	92%
	0.0%	4%	0.0%	0.0%	4%	**92%**	0.0%	8%
Surprise	0	0	5	2	0	0	**24**	96%
	0.0%	0.0%	20%	8%	0.0%	0.0%	**96%**	4%
								92%
								8%

Table 3: Confusion Matrix Resnet50

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
Angry	**10**	1	3	1	3	4	3	40%
	40%	4%	12%	4%	12%	16%	12%	60%
Disgust	2	**11**	5	2	2	2	1	44%
	8%	**44%**	20%	8%	8%	8%	4%	56%
Fear	3	1	**12**	0	1	1	4	48%
	12%	4%	**48%**	0.0%	4%	4%	16%	52%
Happy	1	1	0	**22**	0	0	1	88%
	4%	4%	0.0%	**88%**	0.0%	0.0%	4%	12%
Neutral	2	2	1	1	**14**	4	1	56%
	8%	8%	4%	4%	**56%**	16%	4%	44%
Sad	0	2	1	1	3	**17**	1	68%
	0.0%	8%	4%	4%	12%	**68%**	4%	32%
Surprise	0	0	3	2	0	0	**20**	80%
	0.0%	0.0%	12%	8%	0.0%	0.0%	**80%**	20%
								61%
								39%

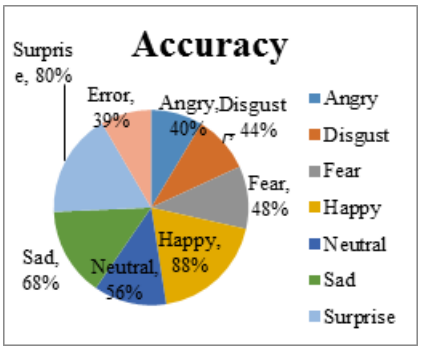


Fig.9. Resnet 50 pie chart

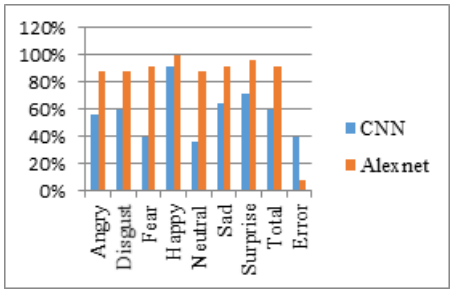


Fig.10. Comparison between CNN and AlexNet

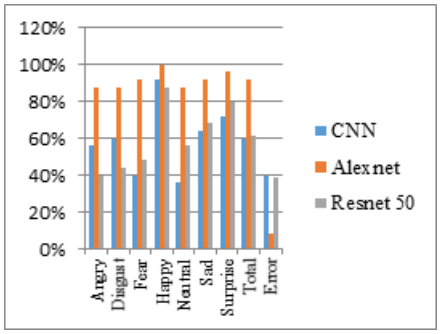


Fig.11. Comparison between CNN, AlexNet and ResNet 50

The testing stage is done by giving the image as input. The trained models are already loaded and the input is compared with those models and finally the expression is detected as presented in fig.4 (B).

This method considers seven expressions namely anger, sad, happy, hatred, fear, disinterested as well assurprising in CNN, AlexNet and ResNet 50 networks. By comparing these three networks AlexNet has more ability to detect the expression than CNN and ResNet 50from fig10 and fig11. The above tables 1, 2, 3 represents the detection of different face expressions and its accuracy values and also same as represented in pie chart for each network.

Conclusion

An overview of the FER methods is discussed in this paper. The distribution of those methods is done by conventional FER methods which involve the detection offace as well as facial element, extraction of feature in addition to the classification steps of expression. Adaboost and random forest are included in the classification algorithms within the conventional FER. Compared to deep-learning-based FER procedures, the dependency upon the face-physics-grounded patterns in addition to additional pre-processing methods can be reduced to a greater extent and the “end-to-end” learning is enabled within the channel immediately from the input images. The model learned using various FER datasets are analysed by a visualization of CNN and the ability of the networks which are trained by the emotion detection are demonstrated over the datasets as well astasks related to different FER. CNN architecture analysis is presented by some new researches for the facial emotions the previously applied CNN methods are outperformed by the sequential aver aging on behalf of aggregation. Nevertheless, several constraints are present in the deep-learning-based FER approaches which include the large-scale datasets, huge computing power, and large memory requirements. Moreover, these approaches consume additional time for training as well as testing the phases.In this paper it is concluded that Alexnet has more accuracy and performance than CNN and ResNet 50 networks.

Standard metricsare provided in direct to compare and the evaluation parameters of the FER-based methods. In terms of recognition, the calculation of the evaluation metricsis performed broadly and moreover the precision and recall are implemented here. Nevertheless to be willing to recognize consecutive facial expressions, a novel evaluation technique is proposed which produces various advantages on its future applications that are to be performed.

References

[1] Suk, M.; Prabhakaran, B. Real-time mobile facial expression recognition system—A case study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 24–27 June 2014; pp. 132–137.

[2] Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 2013, 13, 7714–7734. [CrossRef] [PubMed]

[3] Happy, S.L.; George, A.; Routray, A. A real time facial expression classification system using local binary patterns. In Proceedings of the 4th International Conference on Intelligent Human

Computer Interaction, Kharagpur, India, 27–29 December 2012; pp. 1–5.

[4] Ghimire, D.; Jeong, S.; Lee, J.; Park, S.H. Facial expression recognition based on local region specific featuresand support vector machines. Multimed. Tools Appl. **2017**, 76, 7803–7821. [CrossRef].

[5] Benitez-Quiroz, C.F.; Srinivasan, R.; Martinez, A.M. EmotioNet: An accurate, real-time algorithm for theautomatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5562–5570.

[6] Polikovsky, S.; Kameda, Y.; Ohta, Y. Facial micro-expressions recognition using high speed camera and3D-gradient descriptor. In Proceedings of the 3rd International Conference on Crime Detection andPrevention, London, UK, 3 December 2009; pp. 1–6.

[7] Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial expression recognition from near-infraredvideos. Image Vis. Comput. 2011, 29, 607–619. [CrossRef]

[8] Shen, P.; Wang, S.; Liu, Z. Facial expression recognition from infrared thermal videos. Intell. Auton. Syst.2013, 12, 323–333.

[9] Szwoch,M.; Pieniaz˙ek, P. Facial emotion recognition using depth data. In Proceedings of the 8th InternationalConference on Human System Interactions, Warsaw, Poland, 25–27 June 2015; pp. 271–277.

[10] Gunawan, A.A.S. Face expression detection on Kinect using active appearance model and fuzzy logic.ProcediaComput. Sci. 2015, 59, 268–274.

[11] Wei,W.; Jia, Q.; Chen, G. Real-time facial expression recognition for affective computing based on Kinect.In Proceedings of the IEEE 11th Conference on Industrial Electronics and Applications, Hefei, China, 5–7 June2016; pp. 161–165.