FACIAL EXPRESSION RECOGNITION USING DEEP LEARNING

Ms. T Chandraleka[1], Karishma Agarwal[2], Aparna Singh[3], Utkarsh Singh[4]

[1]*Assistant Professor, Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamil Nadu, India*

[2,3,4]*UG Scholar, Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamil Nadu, India*

ABSTRACT

Human beings are emotional and expressive. Their state of being i.e. their emotions, decide a lot of things. Whether they are understanding something or if they are confused, one can get an idea, by seeing their facial expression. This is the idea behind this paper, which proposes to build a system to recognize various emotions of a person by capturing his/her facial expressions using the deep learning technology. The application of such a system would be to use it in a school or any other educational system to keep a track of how many students actually understand what is being taught and how many don't. As per the statistics, teaching methodologies can be improved to get better results. The six human emotions include: anger, fear, happiness, sadness, disgust and surprise. We will be using convolutional neural networks to analyze our datasets. Bearing in mind the challenge faced by teachers to understand the understanding of a student, we thereby propose a solution which is automatic and accurate using the revolutionary and most practical technology-' Deep Learning'. We employed deep convolutional neural networks to train our model. Also, we analyzed various classifiers and their performance which includes Distance Based Algorithms (DBA), Neural Networks and Normal Bayesian Classifier (NBC).

Keywords: *Facial Expression, Emotion, Convolutional Neural Networks.*

1. INTRODUCTION

Human beings interact with each other verbally as well as with gestures to emphasize more on certain parts of their speech and to display their emotions. While talking about gestures, it is noted that facial expressions play a major role in conveying emotions and are an important part of conversation. Though a lot of conversations happen verbally, it is pertinent to understand the messages send and received through gestures. Automatic recognition of facial expressions is a challenging endeavor but has a lot of real world applications such as in stock behavioral science and in clinical practice. Till now, a lot of methods are used by computers to predict facial expressions, however, it is recognizing facial expression of subjects with reasonably higher accuracies is still not achievable. In recent times, deep learning has become a viable tool for many computer-vision applications. One of the main motivation behind deep learning is that learning several hierarchical levels of feature abstractions fosters the disentanglement of the different aspects in the training dataset, as the different levels in the model can represent different aspects in the training dataset. A prominent deep learning algorithm, known as Convolutional Neural Network(CNN) can be used to examine various layers of the given input image and comprehend the label i.e. the expression. In this paper, we use deep CNN layers to classify facial expressions, providing the input image as a single three dimensional image and applying algorithms to predict the gesture. The accuracy of facial expression recognition (FER) also depends on the type of dataset we employ. Thus, the contributions of this paper are as follows:

1. We posit that learning together from different modalities of the data sets available can provide a more robust classifier for FER.
2. We propose a system wherein faces of students can be taken as input and analyzed by the classifier to recognize the expression.
3. We also propose better applications of this system that is, usage by small-scale self -employed people to increase their sales by predicting customer satisfaction.
4. We also provide a comparison between different state-of-the-art approaches and our method of doing the same.

The rest of this paper is organized as follows. Section-2 discusses the perusal of various methods used for FER. Section 3 elaborates more on the datasets. Section-4 describes the necessary background for our problem formulation. In section-5 we present in detail the observations and key findings. We conclude in Section-6 by highlighting the results and applications.

2. ANALYSIS OF CLASSIFIER ALGORITHMS EMPLOYED FOR FER

Facial feature extraction methods can be classified into two main categories, according to the way of obtaining the face patterns the "Holistic approach" (Template matching) which considers the hole face region in an image as the system's input data and represents each face as a vector whose components codify the grey level of each face pixel, and the "Feature approach" (Geometric, feature-based matching) which establishes certain facial landmarks related to face elements, such as eyes, nose, mouth and ears and computes features as distances between landmarks, relative positions or elements' sizes. A big majority of these systems are based on the holistic approach because, as remarked in [1], the template approach is more reliable than the feature one, and its implementation is simpler [2]. Feature extraction is the first step in facial expression recognition followed by classifier to classify input face expressions. A lot of algorithms have been employed by various scholars to get the highest accuracies in the field of FER. Basic procedure includes- data collection, face detection, pre-processing, feature extraction and classification. Among supervised algorithms, the following stand out:

1) Distance based algorithms, which make use of the simplest way of classifying a sample template in certain space by computing a similarity measure or distance to all the existing pattern templates (considered as vectors in a multidimensional space), in order to determine the closest one. The k-Nearest Neighbors (kNN) method is the most popular one, where given a positive integer number, k and a sample template, the k training templates with the smallest distance to the sample ones are selected.
2) Neural networks are learning structures which adapt to changes, proposed by a fixed number of interconnected computing units called neurons, capable of approximating non-linear output signals from input signals (feature vectors).
3) The Normal Bayesian Classifier (NBC) is based on the idea that feature vectors from each user are normally distributed. The data distribution function is therefore a Gaussian mixture with one component per class [3].

   YavuzKahraman et al. [4] implemented geometry-based features for face expression recognition. This Technique searched for 153 possible distances among 18 critical candidate points/landmarks. Correlation-based feature subset selection (CFS) method was applied to select 16 of these distances having significant contribution to accuracy. This CFS+ANN method has 91.2% correct classification rate. Zhou Ji-liu et al. [5] used automatic fiducial point location algorithm locating 58 fiducial points and calculated the Euclidean distances between the center of gravity coordinate and the fiducial points coordinates of the face. Person's neutral expression and the other seven basic expressions are used to extract geometric deformation difference features. This feature vector acts as input to multiclass SVM classifier which classifies data input seven basic expressions.Jun Wang et al proposed a deep convolutional neural network (CNN) for facial expression recognition system which is used for deeper feature representation of facial expression to achieve automatic recognition. The proposed system results 76.7442% and 80.303% accuracy in the JAFFE and CK+, respectively.

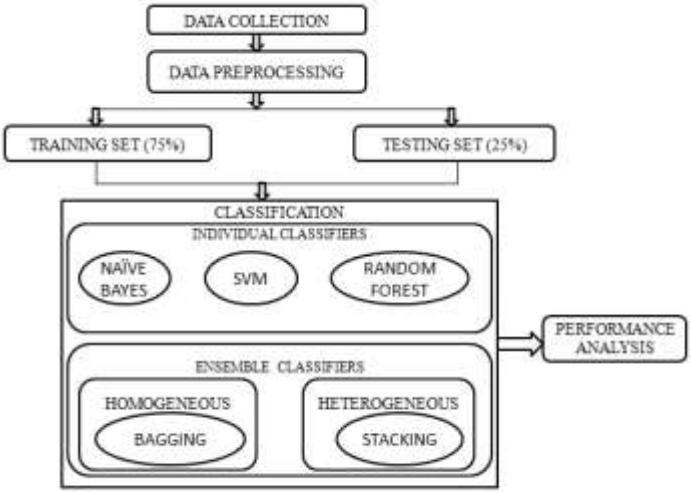

Fig.1 Block Diagram about proposed system

3. DATASETS

During the assessment stage of new techniques, in order to compare the performance of several methods, it is recommendable to use a standard testing data set. There are many databases currently in use and each one has been developed under a different set of requirements (a complete list can be found in [9] and http://www.face-rec. org/databases/).

The Extended Cohn-Kanade (CK+) (LUCEY et al., 2010) database contains 497 sequencesof 100 subjects. Each sequence contains about fifteen images and starts with the neutral expression of a subject and proceeds to a peek expression. All images in thedataset are 640 by 480 pixel arrays with 8-bit precision for grayscale values. Each image has a descriptor with its facial points, these points were used to normalize the facialexpression image.

4. FACIAL EXPRESSION RECOGNITION SYSTEM USING CNN

The first stage of the methods is a pre-processing step that aims to extract the best set of features that describes the facial changes caused by an expression. Once the images are pre-processed they can be either used to train the network or to test it (i.e. recognitionstep). In the training step, a set of pre-processed images are given to the network withtheir respective labels so that the best set of network weights for classification can befound. In the testing step, the network is configured with the weight set found duringthe training and the recognitions are performed. The recognition outputs the confidencelevel of each expression. The maximum confidence level is used to infer the expressionin the image. In order to increase the number of training samples a synthetic imagegeneration method is used during the pre-processing stage.

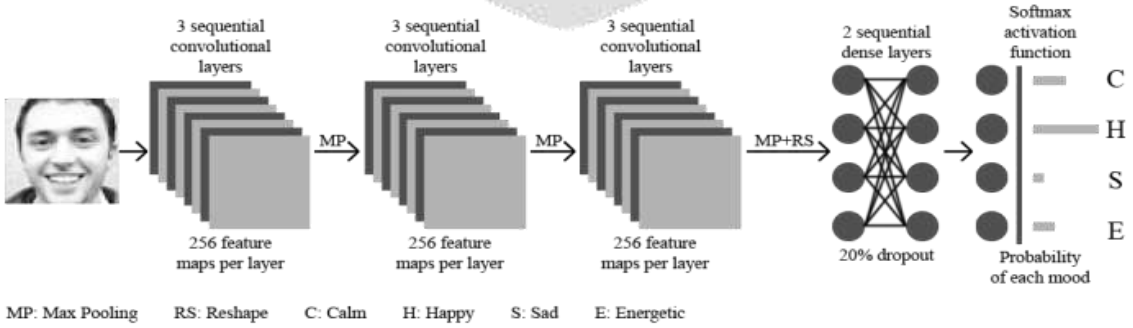

Fig.2 Various Layers used in the system including feature maps and Softmax activation function

4.1 SPATIAL NORMALIZATION

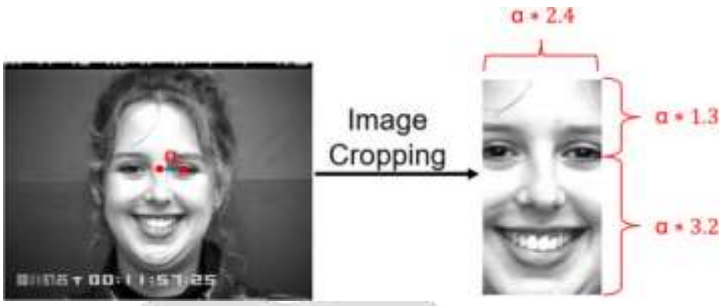

Fig.3 Cropping phase in spatial normalization

The spatial normalization procedure comprises three steps: rotation correction, image cropping and down-sampling. This procedure helps thefacial parts (eyes, mouth, nose and eyebrows) to be in the same pixel space helping theclassifier to associate which image parts are related to each expression.

4.1.1 ROTATION CORRECTION

This is the first step in spatial normalization wherein two informations are needed- facial expression and center of both the eyes. Based on these a geometric rotation and translation is carried out, to align two eyes center with horizontal axis and to keep face centralized in the image.

4.1.2 IMAGE CROPPING

This is the second step which involves the cropping of rotation corrected image thus removing the unnecessary background details and image patches that are not related to the expression.

4.1.3 DOWN SAMPLING

This is the last procedure for spatial normalization. After cropping, images will be of different sizes, therefore we need to down sample the image using a linear interpolation to 32X32 pixels in order to remove the variation in face size and keep the facial parts in same pixel space.

4.2 INTENSITY NORMALIZATION

The image brightness and contrast can vary even in images of the same person in the same expression increasing, therefore, the variation in the feature vector. Such variations increase the complexity of the problem that the classifier has to solve for each expression.

In order to reduce these issues an intensity normalization was applied. A method adapted from a bio-inspired technique described in (WANDELL, 1995), called contrastive equalization, was used. Basically, the normalization is a two-step procedure: firstly a subtractive local contrast normalization is performed; and secondly, a divisive local contrastnormalization is applied.

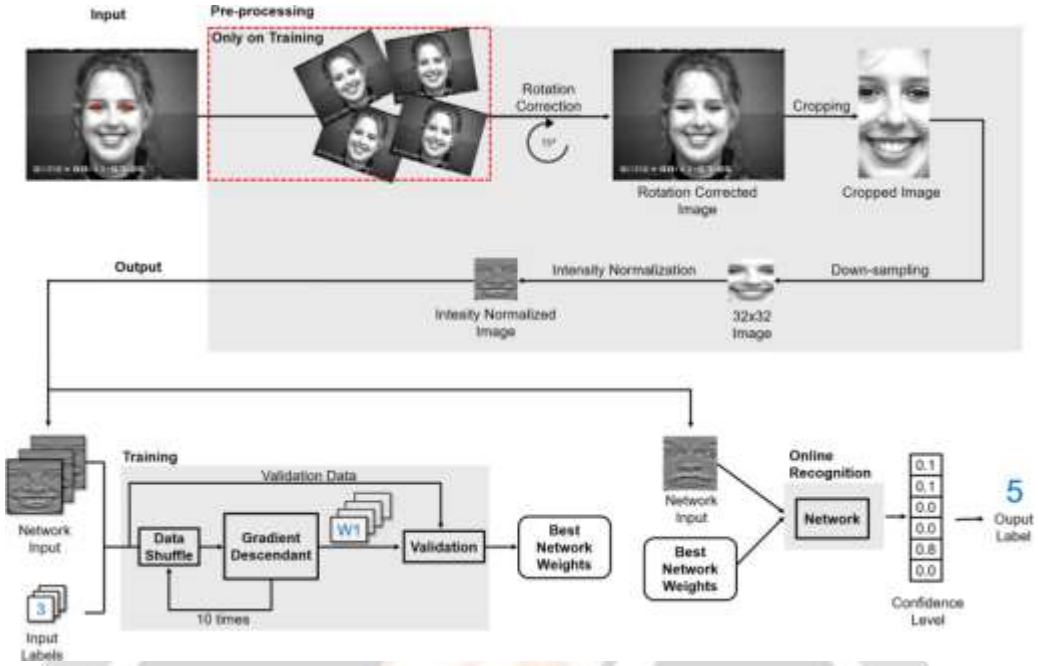

Fig. 4 Intensity normalization and its various sub steps

4.3 CONVOLUTION NEURAL NETWORK FOR FACIAL EXPRESSION CLASSIFICATION OF THE INTENSITY NORMALIZED IMAGE

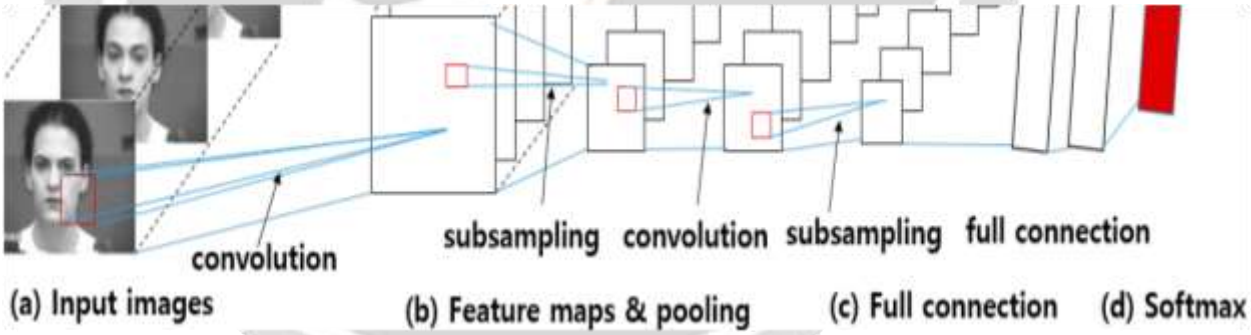

Figure 5: Architecture of the proposed Convolutional Neural Network for the current method. It comprises five layers: the first layer (convolution type) outputs 32 maps; the second layer (subsampling type) reduces the map size by half; the third layer (convolution type) outputs 64 maps for each input; the fourth layer (subsampling type) reduces the map once more by half; the fifth layer (fully connected type) and the final output withN nodes representing each one of the expression are responsible for classifying the facial image.

The architecture of our Convolutional Neural Network is represented in Figure 8. The network receives as input a 32_32 grayscale image and outputs the confidence of each expression. The class with the maximum value is used as the expression in the image. Our CNN architecture comprises 2 convolutional layers, 2 sub-sampling layers and one fully connected layer. The first layer of the CNN is a convolution layer,that applies a convolution kernel of 5x5 and outputs an image of 28x28 pixels. This layer is followed bya sub-sampling layer that uses max-pooling (with kernel size 2x2) to reduce the image to half of its size. Subsequently, a new convolution with a 7x7 kernel is applied to the feature vector and is followed by another sub-sampling, again with a 2x2 kernel. The output is given to a fully

connected hidden layer that has 256 neurons. Finally, the network has six or seven output nodes (one for each expression that outputs their confidence level) that are fully connected to the previous layer.

Two main classes of experiments were performed: experiments training and testing in the same database (CK+ or JAFFE or BU-3DFE) and experiments training in one database (CK+) and testing in another (JAFFE or BU-3DFE), i.e. cross-database experiment. To perform the experiments, the databases used in each experiment were divided into three sets: Training set, which is used to train the system; Validation set, which is used to dynamically tune the meta-parameters of the system (e.g. choose the best network weights out of 10 runs with random training samples presentation order); and the test set, which is used to actually measure the accuracy of the system.

5. RESULT

5.1 Training with CK+ and test with CK+

Table below shows the best result achieved(using both normalizations and the synthetic samples) using both classifiers. As can be seen the binary classifiers increase the accuracy. It happens because in this approach the hit can be achieved n times(one for each expression), instead of using just one classifier, where each sample has just one chance to be properly classified.

	Angry	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**
$C_{6classE}$	93.33%	100.00%	96.00%	98.55%	84.52%	99.20%
C_{binE}	98.27%	99.37%	99.24%	99.68%	98.17%	98.81%

Table 1:The training parameters that achieves the results

5.2 Training with JAFFE or CK+ and Test with the JAFFE

This database was also used for experimentation to measure the performance of the system. Methodology used was same for this database too.

Classifier	**Train**	**Test**	**6-expressions(%)**	**7-expressions(%)**
C_{nclass}	**CK+**	**JAFFE**	**45.23**	**36.07**
C_{bin}	**CK+**	**JAFFE**	**81.74**	**81.73**

Table 2: JAFFE accuracy cross database sets

The whole experimentation time including all the k-fold configurations of the proposed method was 32 hours to the method proposed in Section 3.1 and only 16 hours to the method proposed in Section 3.2, and the recognition is real time (only 0.01 second for each image), almost 100 images per second.

6. CONCLUSION

In this paper a deep learning based approach has been proposed to recognize facial expression. The classification of facial expressions has been practiced using the textured and geometric modalities as the source of information. Different architectures and depth of convolution filters, loss function that can give feedback on how much data is retained within the image are investigated. The proposed CNN based approach has been compared with two feature-based approaches. Demonstrated their performance using different processing and visualization techniques illustrating their advantage and disadvantage. The result demonstrated the deep CNN are capable of learning facial characteristics and improving facial expression results.

Due to the current problem of facial pose changes in the scene, it would be of great interest to design a 3D model that will be robust to these variations. Also, taking advantage of set analysis demonstrated in this work, it would be interesting to develop a system that in addition to use the sets analysis, add an adaptive methodology to estimate the status of each facial action unit to improve performance of the emotion recognition system. In future we plan to conduct more related research on video database by identifying sentiments from each video frames. The main advantage of video sentiment analysis is that it express both audio as well as video responses. The tone of the speaker is determined by the vocal modulation from recorded responses whereas sentiment nature of speaker is provided by visual data.

7. REFERENCES

[1]Miguel F. Arriaga-Gomez, Ignacio de Mendizabal-Vazquez ´, Rodrigo Ros-Gomez ´ and Carmen Sanchez- ´ Avila, “A Comparative Survey on Supervised Classifiers for Face Recognition” ,IEEE

[2] R. Brunelli and T. Poggio, “Face Recognition: Features versus Templates,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042–1052, 1993.

[3] Alptekin D.,YavuzKahraman "Facial Expression Recognition Using Geometric Features" The 23rd International Conference on Systems,Signals and Image Processing. 23-25 May 2016, Bratislava,Slovakia.

[4] Lei Gang, Li Xiao-hua* ,Zhou Ji-liu, Gong Xiao-gang "Geometric feature based facial expression recognition using multiclass support vector machines" IEEE International Conference on Granular Computing,2009

[5] Ke Shan, JunqiGuo, Wenwan You, Di Lu, RongfangBie "Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural- Network Structure " IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA),2017