© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Automatic Facial Expression Recognition using Convolutional Neural Network (CNN)

Eftekhar Ahmed*, Tasnim Azad Abir, Jinat Ara Siraji
Department of Electronics and Communication Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh

ABSTRACT

Facial Expression Recognition is has been widely used in Artificial Intelligence, Human-Computer Interaction, and Security Monitoring. Convolution neural network (CNN) works as a depth learning architecture and it can extract the essential features of the image. In the case of large changes in shooting conditions, CNN's effect is better than the methods of Support Vector Machines (SVM) and Principal Component Analysis (PCA). Therefore, we are proposing a method based on CNN. The purpose is to classify each facial image as one of the seven facial expressions considered here. A new convolution neural network structure has been designed according to the characteristics of facial expression recognition. To extract implicit features convolution kernel is being used and max-pooling is being used to reduce the dimensions of the extracted implicit features. In comparison to AlexNet network, we can improve the recognition accuracy about on the FER and CK+ facial expression database with the help of Batch Normalization (BN) layer to our network. A facial expression recognition system is constructed for the convenience of application, and all the experimental results show that the system can reach the real-time needs.

Keywords: Convolutional Neural Network, Distributed Neural Network, Support Vector Machine, Principle Component Analysis

I. INTRODUCTION

Face provides a wide range of information about the identity, age, sex, race as well as the emotional and mental state. Facial expressions play a crucial role in social interactions and commonly used in the behavioral interpretation of emotions. Human expression recognition is influenced by certain context. When a subject is being investigated, the investigator might be diverted by the subject's voice tone or argument and may forget to keep track of the facial expressions. Automatic facial expression recognition systems are exempt from such contextual interference [1]. Such systems can be beneficial in many fields, like gaming applications, criminal

interrogations, psychiatry, animations etc. We want to improve a CNN architecture which will be able to detect the meaning of the human facial expression. Convolution Neural Network (CNN) is a new type of neural networks, it is a combination of traditional artificial neural network and deep learning technology [2]. With the development of deep learning, applying convolution neural network into a classification problem has attained impressive success. Critical and unforeseen features are extracted by deep learning methods through iterative weight update by backpropagation and error optimization [3-4].

When we talk face to face we can easily understand the facial expression of others. But still now it remains challenging for machines to understand the meaning of human's facial expression. The main objectives of our works are the improvement of face detection algorithm and implementation of the feature extraction technology. We also want to get better effect of classification than before and build a perfect architecture of convolution neural network. To train machine about how to understand the meaning of the human's facial expression accurately there are still many problems, such as low precision, slow identification, etc. [5]. This work will develop an improved technique to detect the face expression by feature extraction technology.

II. METHODS AND MATERIAL

A. Deep Learning and Artificial Neural Network

Deep learning is basically a branch of machine learning. It is a technique which can directly extract features from data like text, image, or sound and learn the task by itself [6]. Deep learning methods can perform object recognition and classification much accurately than human [7]. Using Graphical Processing Unit (GPU) large deep network can complete the task of learning easily within a very short period of time. For very much effective learning we need sufficient quality data and the library has already been very much enriched with a great amount of dataset by different services. As most of the deep learning method follows the neural network architecture, deep learning models are often mentioned as the deep neural network [8]. A very much popular deep learning model is convolution neural network (CNN) which deals with image data. When a neural network is designed with multiple layers it is known as the deep neural network. The number of layers can be two to hundreds [9].

B. Artificial Neural Network

Artificial Neural Network (ANN) simulate and processes data in a similar fashion the human brain analyzes and processes information [10]. It is based on the idea that working of human brain can be imitated by making the right connections of silicon and wires as living neurons and dendrites. Neurons are connected to other cells by Axons. Dendrites create electric impulses in response to stimuli from external environment or inputs from the sensory organs [11]. A neuron sends the stimuli to other neuron to handle the issue. In this way these impulses travel through the whole neural network. The Multiple nodes that behaves like the biological neurons of human brain

compose ANN. Connected through links, neurons can interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons called its activation or node value [12]. There is a weight associated with each link. By altering weight values, the ANNs are capable of learning. The ANN is based on a collection of artificial neurons. Each connection between artificial neurons can transmit a signal from one to another. The artificial neuron can process the signals and transmit the signal to the artificial neurons connected to it [13].

C. Convolutional Neural Network

A multilayer neural network with one or more convolutional layers and one or more fully connected layers is known as a convolutional neural network (CNN) [14]. The architecture of a CNN is designed to take advantage of the 2D structure of an input signal and it is easier to train and have many fewer parameters than fully connected networks having the same number of hidden layers. The convolutional layer's parameters consist of a set of learnable filters which are spatially small but extends through the full depth of the input volume [15]. At the time of forward pass each filters calculate the dot products between the entries of the filter and the input sliding across the width and height of the input volume. At every spatial position, the response of the filter is made by sliding over the width and height of the input volume producing a 2-dimensional activation map [16]. We will have an entire set of filters in each convolutional layer and each of them will produce a separate 2-dimensional activation map. These activation maps along the depth dimension will be stacked and produce the output volume. Dealing with high-dimensional inputs it is practical to connect each neuron to only a local region of the input volume [17]. The spatial extent of this connectivity called the receptive field of the neuron and the extent of the connectivity along the depth axis is always equal to the depth of the input volume. The connections are local in space but always full along the entire depth of the input volume.

It is common to periodically insert a pooling layer in-between successive convolutional layers in a convolutional network architecture [7]. By progressively reducing the spatial size of the representation, it reduces the amount of parameters and computation time in the network. Hence, also controls over fitting. In addition to max pooling, the pooling units can also perform other functions, such as *average pooling or even L2-norm pooling [18].*

It is common to keep track of the index of the max activation at the time of forward pass of a pooling layer so that gradient routing is efficient during back propagation [19]. In a fully connected layer neurons have full connections to all activations in the previous layer. Their activations then can be computed with a matrix multiplication followed by a bias offset. CNN use little pre-processing compared to other image classification algorithms.

III. SIMULATION TOOLS AND ALGORITHMS

A new CNN structure is proposed for facial expression recognition. Two publicly available databases FER and CK+ are used to carry out the experiments [15]. In order to remove non-expression feature of a facial image, we need to carry out the pre-process of face image, which includes face detection and face image cropping. The faces are detected and cropped using OpenCV library. Then, the facial expressions features are extracted using our convolution neural network which works under deep learning framework. Finally, the training model is used to classify each facial image as one of the seven facial expressions: angry, disgust, neutral, sad, fear, surprise and happy.

A. Network Architecture

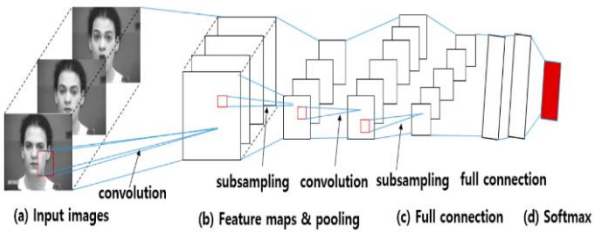

Figure.1 Facial expression recognition network structure.

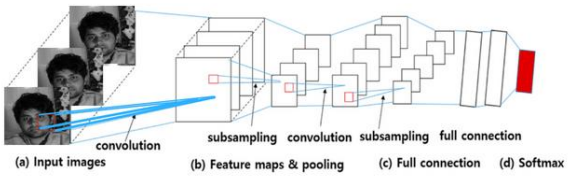

Figure.2 Facial expression recognition network structure using softmax pooling.

Figure.1 shows the facial expression recognition network. The entire network includes four convolution layers, two pooling layers and two fully connection layers. The convolution layer is the feature extraction layer. The convolution operation is performed on the training convolution kernel and the previous layer of all the feature maps. The output of the activation function forms the neurons of the current layer, thus forming the characteristic map of the current convolution layer. Figure. 2 shows the facial expression recognition network structure using softmax pooling. The calculation is as follows [17]:

$$net_j^l = \sum_{i \in M_j} a_i^{l-1} \otimes \omega_{i,j}^l + \omega_b$$

$$a_{i,j}^l = F\left(net_{i,j}^l\right)$$

Where net_j^l represents the weighted input of layer l. a_i^{l-1} is the characteristic map of the output of the l-1 layer. $\omega_{i,j}^l$ denotes a convolution kernel matrix, it represent the connection weight between neurons.ω_b denotes the offset term of the jth feature map. In the experiment, set ω_b = 0, can improve the speed of network training, while reducing the learning parameters. $a_{i,j}^l$ is the j feature graph of the convolution l layer. F() represents the activation function. This model uses 96 filters with the size of 11×11×4.This layer extracts the low-level edge features. The original image is randomly cropped to the size of 227×227, after the first convolution of the image size becomes 55×55. To increase the nonlinear properties of network, we use ReLU(Rectified Linear Units) as the activation function. For any given input value x, ReLU is defined by:

$$F(x) = \max(0, x)$$

Where x is the input to the neuron. Using the ReLU activation function allows us to avoid the vanishing gradient problem caused by some other activation

functions. Increasing the number of convolution layers the feature dimension increases rapidly [12]. In order to avoid such a large dimension we used the pooling layer to reduce the dimension. The down-sampling is performed by Max-pooling. The down-sampling process does not change the number of feature graphs. It reduces the number of parameters by removing unnecessary information from each feature map. The models are trained and tested on the database from FER Kaggle [15]. The dataset has grayscale images of faces of the size 48x48 pixel. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. We have classified each face based on the emotion in the facial expression in to one of seven categories: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. The training set consists of 28,709 examples. The public test set used for the leaderboard consists of 3,589 examples. The final test set consists of another 3,589 examples.

For cross validation we used extend Cohn-Kanaded database (CK+) which includes 327 video sequences acted out by 118 participants [15]. Each sequence which consists of approximately 10 to 30 frames is labeled with one of seven expressions categories: angry, disgust, neutral, sad, fear, surprise, happy. Every sequence starts with the neutral emotion and then the frame depicts the emotion which is for the corresponding label. We selected images of human faces with obvious facial expressions to recognize facial expression. All 327 sequences of the CK+ database are used for evaluating the proposed model.

C. Python Development Environment

As python development environment we have used Spyder which is the scientific python development environment. It may also be used as a library providing powerful console-related widgets for PyQt-based applications.

D. AlexNet

AlexNet network used a relatively simple layout, compared to modern architectures [18]. The network was made up of 5 convolutional layers, max-pooling layers, dropout layers, and 3 fully connected layers as shown in figure. 3. The network was used for classification with 1000 possible categories.

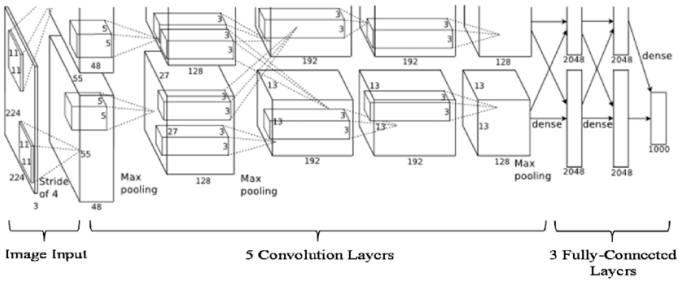

Figure. 3 AlexNet Architecture

The neural network developed for AlexNet was the coming out party for CNNs in the computer vision community. Tensor Flow mainly designed for deep neural network models is a tool for machine learning that contains a wide range of functionality.

IV. RESULTS AND DISCUSSIONS

We trained and tested the model on the database from FER Kaggle. The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. Training contains two columns, emotion and pixels. The emotion column contains a numeric code ranging from 0 to 6 for the emotion that is present in the facial expression image. The pixels column contains a string surrounded in quotes for each image. The contents of this string a space-separated pixel values in row major order. Testing contains only the pixels column and your task is to predict the emotion column. For cross validation we used extend Cohn-Kanaded database (CK+) which includes 327 video sequences acted out by 118 participants. Each sequence which consists of approximately 10 to 30 frames is labeled with one of seven expressions categories: angry, disgust, neutral, sad, fear, surprise, happy. Every sequence starts with the neutral emotion and then the frame depicts the emotion which is for the corresponding label. We selected images of human faces with obvious facial expressions to recognize facial expression. All 327 sequences of the CK+ database are used for evaluating the proposed model. Figure. 4. Shows the confusion matrix of the test datasets. Figure 5 and 6 presents the real time view of facial expression classification using DCNN and SVM respectively.

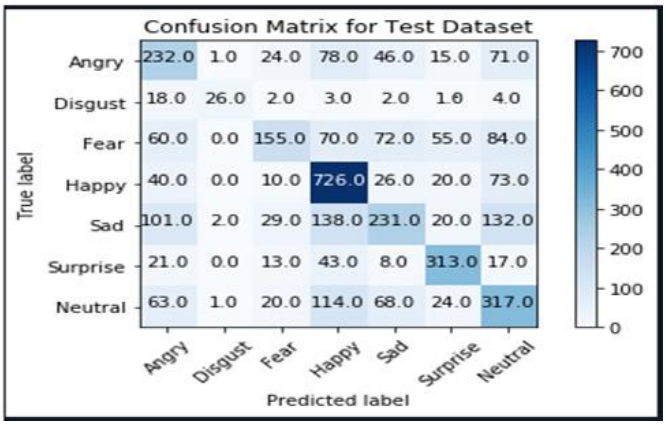

Figure. 4 Confusion Matrix

Table I Comparison of the accuracy of the network structures under two different methods

Method	Accuracy
SVM	76%
DCNN	88%

Table I presents the comparison between accuracy of facial expression classification using SVM and DCNN. It is evident that DCNN provides better classification accuracy. From Figure. 4 we can see the confusion matrix. It shows that the accuracy is good enough. We can see the observations. Wit5h two methods we made real time facial expression detection and found that the DCNN method is better than SVM.

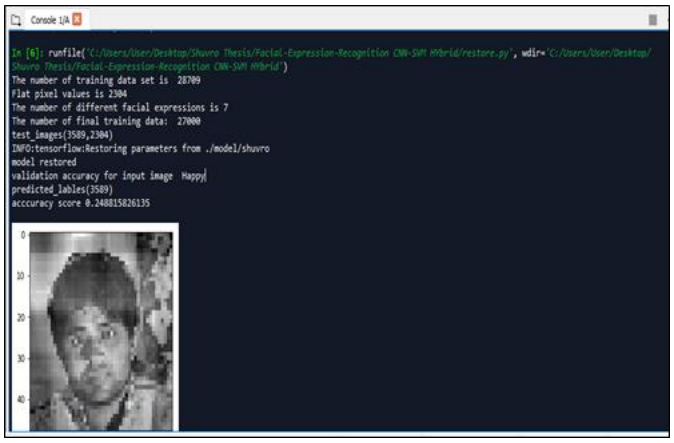

Figure.5 Real time facial expression for DCNN method

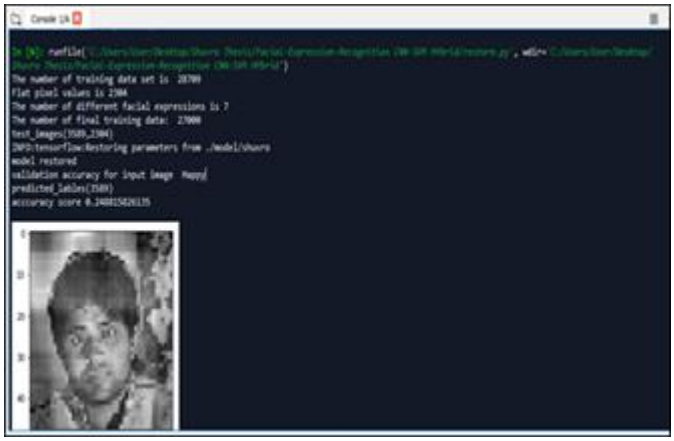

Figure.6 Real time facial expression for SVM method

V. CONCLUSION

The purpose of this work is to classify each facial image as one of the seven facial expressions: angry, disgust, neutral, sad, fear, surprise and happy. A new convolution neural network structure has been designed according to the characteristics of facial expression recognition. To extract implicit features convolution kernel is being used and max-pooling is being used to reduce the dimensions of the extracted implicit features. We built the structure of the neural network model. The purpose is to detect the meaning facial expression automatically. We trained the model, imported datasets, and test set. We made the comparison. We determined accuracy for train set FER and Test set FER. We used two methods which are SVM and DCNN. We found that the DCNN was the better one.

VI. REFERENCES

1. J. Yin "Face Festure Extraction Based on Principle Discriminant Information Analysis" ,IEEE International Conference on Automation and Logistics, 2007, pp. 1580-1584.
2. M Pantic and L.J.Rothkrantz "Automatic analysis of facial expressions: The state of the art. Pattern Analysis and Machine Intelligence", IEEE Transactions on, 2000, pp. 1424-1445.
3. Ian J. Goodfellow "Challenges in representation learning: A report on three machine learning contests" , Neural information processing .2013, pp. 117-124.
4. ZYu and C.Zhang "Image based static facial expression recognition with multiple deep network learning", the 2015 ACM on International Conference on Multimodal Interaction.2015, pp. 435-442.
5. SE.Kahou "Combining modality specific deep neural networks for emotion recognition in video" , the 15th ACM on International conference on multimodal interaction. 2013, pp. 543-550.

6. SMinchul "Baseline CNN structure analysis for facial expression recognition", 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) 2016, pp. 26-31.
7. PEkman and W.V.Friesen "Facial Action Coding System: A Technique for the Measurement of Facial Movement", Palo Alto: Consulting Psychologists Press, 1978.
8. TKanade "Recognizing action units for facial expression analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence.2001, pp. 97-115.
9. MS. Bartlett "Fully automatic facial action recognition in spontaneous behavior", IEEE Conference on Automatic Facial and Gesture Recognition, 2006, pp. 223-230.
10. C.Ira "Evaluation of expression recognition techniques.", Image and Video Retrieval. Springer Berlin Heidelberg, 2003, pp. 184-195.
11. M.Liu "Deeply learning deformable facial action parts model for dynamic expression analysis" ,Computer Vision–ACCV. Springer International Publishing.2004, pp. 143-157.
12. P.Burkert "DeXpression:Deep Convolutional Neural Network for Expression Recognition",IEEE 2015.
13. G.Ali "Boosted NNE collections for multicultural facial expression recognition. Pattern Recognition", 2016, pp.14–27.
14. P.Lucey "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression", Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on IEEE.2010, pp. 94–101.
15. M.Lyons "Coding facial expressions with gabor wavelets. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition",1998,pp. 200–205.
16. A.Krizhevsky "Imagenet classification with deep convolutional neural networks" Advances in neural information processing systems, 2012, pp . 1097–1105.
17. Y.Jia "Caffe: Convolutional architecture for fast feature embedding", Eprint Arxiv.2014,pp. 675-678.
18. K.Alex "Imagenet classification with deep convolutional neural networks", International Conference on Neural Information Processing Systems. 2012 , pp . 1097-1105