2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Deep Convolutional Neural Network for Facial Expression Recognition using Facial Parts

Lucy Nwosu[1], Hui Wang[1], Jiang Lu[1], Ishaq Unwala[1]Xiaokun Yang[1]and Ting Zhang[2]
[1]Dept. of Computer Engineering, University of Houston – Clear Lake, Houston, TX 77058
Email: Lucy.Nwosu@uhcl.edu,LuJ@uhcl.edu
[2]Department of CSET, University of Houston – Downtown, Houston, TX 77002
Email: zhangtin@uhd.edu

***Abstract*— This paper proposes the design of a Facial Expression Recognition (FER) system based on deep convolutional neural network by using facial parts. In this work, a simple solution for facial expression recognition that uses a combination of algorithms for face detection, feature extraction and classification is discussed. The proposed method uses a two-channel convolutional neural network in which Facial Parts (FPs) are used as input to the first convolutional layer, the extracted eyes are used as input to the first channel while the mouth is the input into the second channel. Information from both channels converges in a fully connected layer which is used to learn global information from these local features and is then used for classification. Experiments are carried out on the Japanese Female Facial Expression (JAFFE) and the Extended Cohn-Kanada (CK+) datasets to determine the recognition accuracy for the proposed FER system. The results achieved shows that the system provides improved classification accuracy when compared to other methods.**

Keywords*— *Facial expression recognition, Convolutional Neural Networks, Facial Parts.

I. INTRODUCTION

The face, also called "the organ of emotion" is the most powerful "channel" of nonverbal communication. The face can send many subtle signals as an array of facial expressions—a smile of happiness, a frown of sadness or disapproval, the wide-open eyes of surprise, or a lip curled in disgust. These signals if recognized by machines can make human-machine interaction more robust and harmonious. Facial expression recognition (FER) is one of the most important nonverbal channels through which Human Machine Interaction (HMI) systems can recognize humans' internal emotions and intent. It is a type of biometric authentication that focuses on uniquely recognizing human facial appearance based upon one or more physical or behavioral traits and inside emotions portrayed on one's face. An FER has many application areas such as human machine interaction as used in social robots, security-surveillance and computer interactive games. It is used in behavioral science to provide social information (origin, age, and gender) and used in medicine science- for pain monitoring, treatment of mental retardation, depression and anxiety. Although humans recognize facial expressions virtually without effort, reliable expression recognitionby machine is still a challenge. Key challenges in FERs include achieving optimal preprocessing, feature extraction or selection, and classification, particularly under conditions of input data variability, view or pose of the head, environmental clutter and illumination, multiple sources of facial variability.

Over the years various methods and algorithms has been employed for improving the performance of the FER. To overcome the problem of having an input with multiple well known standard face datasets, Ali [14] developed a deep neural network architecture that takes registered facial images as the input and classifies them. To recognize the facial expressions across different facial views, Tong [10], proposed a novel deep neural network (DNN)- driven method in which scale invariant feature transform (SIFT) features corresponding to a set of landmark points are first extracted from each facial image. Many studies have also been conducted using engineered features (e.g. Scale Invariant Feature Transform (SIFT) [10] and Gabor [9], [6]) where the classifiers hyper-parameters are tuned to give best recognition accuracies across a single database. Recently, several works on FER successfully uses Convolutional Neural Networks (CNNs) [2-4] for feature extraction and classification. CNNs can be found in literature and as a special type of multilayer perceptron (MLP) that focuses on the local relationship between pixels by using receptive fields. They have been shown to achieve high recognition rates in various image recognition tasks, these methods however use different CNN architectures to achieve desired result.

This paper gives an overview of a proposed methodology and technique for face detection, feature extraction, and classification that uses CNN architecture for improving the performance of the FER and overcome the above challenges. The rest of this paper is organized as follows: Section II gives a background on existing works done by researchers using CNN for facial expression recognition, Section III describes the proposed architecture and method, Section IV reports the experimental results, evaluation and discussion based on result. Finally, Section V concludes the paper.

II. RELATED WORK

In recent years, researchers have made considerable progress in developing expression classifiers [7-9]. In order to obtain better representation of facial expressions several deep learning techniques that work on data representations, such as Convolutional Neural Networks (CNN) has been

978-1-5386-1956-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.213

developed. The CNN used in [5], has a conceptual similarity to the one used to build the FACS model. The input is decomposed into features and each deeper layer has a more complex representation, which builds upon the previous layer. The final feature representations are then used for classification. CNNs differ from the FACS model [6] by the fact that, if shown enough samples, the network can actually learn a general representation of a smile instead of using a fixed and pre-defined structure representation. . In 2015, Yu and Zhang used CNN for FER in EmotiW. They used an ensemble of CNNs [13] and randomly perturbed the input image to achieve a 2-3% boost in accuracy. Kahou et al. [11], in their work evaluated two experiments using CNNs, in which a standard CNN was pre-trained using the Acted Facial Expression in the Wild (AFEW) dataset [12]for the first experiment and the Toronto Face Datasetfor the second experiment. Both experiments were used for the recognition of six facial expressions, based on the universal facial expressions and one neutral expression. The paper successfully showed that pertaining improved network performance. Dennis H. et. al in his work[4], applied a face image to two channels of CNN, the information from the two networks was combined to give generate a 94.4% recognition accuracy.A CNN network that uses an ensemble of the outputs of three CNNs for classification is analyzed in [1]. The CNNs uses the face, mouth and eyes as input. The method when compared to other methods achieved a high recognition rate.

In this paper, a simple CNN architecture to predict basic expressions from facial parts is proposed. This approach combines standard methods such as Viola Jones method for face and facial regions detection/extraction [17,18] and convolutional neural network for feature extraction and classification. The main contributions of this paper are:

1) Using facial parts as input to reduce the size of input and FER processing time when compared to other CNN architectures that used the whole face for facial expression recognition.
2) Proposing a two-channel CNN architecture for feature extraction and expression recognition

III. PROPOSED METHOD

The proposed facial expression recognition method used in this paper is based on a two-channel architecture that is able to recognize facial expressions. Fig.1 shows the process diagram of the proposed facial expression recognition system which is divided into three stages- Image Pre-Processing which involves Face and facial parts detection using viola-Jones algorithm, facial Feature extraction and feature classification using CNN.The detected facial parts are cropped and extracted and then used as the input into the inception layer of the CNN. The Training phase involves the feature extraction and classification using convolution neural network. It is expected that using the facial parts instead of the whole face image as input to the inception layer will reduce training time and increase the probability of high level feature extraction thereby increasing system accuracy.

A. Image Pre-processing

This is the first phase of the proposed system and it involves the Face Detection and FPs detection/extraction. The Viola-Jones face detection framework, which is a robust algorithm capable of processing images extremely rapidly for real-time situations, is used [7]. This algorithm detects face region irrespective of variance in size, background, brightness and spatial transformation of the raw input image. The face/FP detection is achieved by combining classifiers in a cascade structure that is capable of increasing the detection performance while reducing computational complexity. Final classifier is computed by the linear combination of all weak classifiers, which separates the positive and negative in terms of the weighted error (weight of each learner is directly proportional to its accuracy).

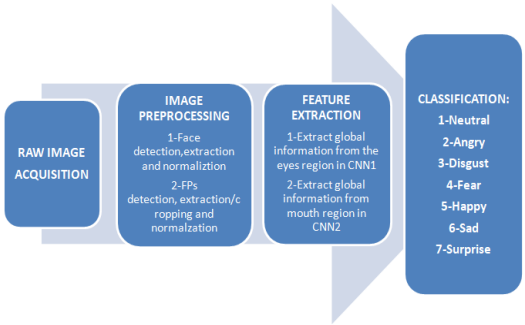

Fig.1. Process diagram of the proposed FER system

The face is first detected, cropped, extracted and normalized to a size of 64 x 64 pixels, and then facial parts (both eyes and mouth) are detected, cropped and extracted from the normalized face image. The extracted facial parts are resized to an equal size of 32 x 64 pixels. The reduced image scale helps to reduce the information that has to be learned by the network and also makes training faster and with less memory cost [1]. Fig. 2 represents a set of extracted FPs from an image in the JAFFEE database without resizing.

Fig.2. (a) raw imge; (b) cropped face and extracted FPs

B. The CNN based feature extraction and classification

CNNs employ the concepts of receptive field and weight sharing. Through these concepts, the number of trainable parameters is being reduced and the propagation of information through the layers can be calculated by convolution. A signal is convolved with a filter map, containing the shared weights to produce a feature map.

The proposed CNN architecture proposed in Fig. 3 consists of two channels of CNNs that takes in the eyes and mouth separately. Each of the CNN channels receives an input grayscale image of 32 x 64 pixels; Information from both channels converges in a fully connected layer (FC) which is used to learn global information from these local features and is then used for classification.

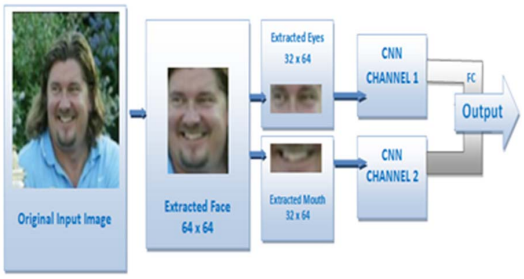

Fig.3. Illustration of the proposed CNN architecture

A given input image is passed through the CNN architecture as illustrated in Fig. 3. Each CNN channel is made up of 2 convolutional layers, 2 max-pooling layers, 1 full connection layers and 1 output layer. The first CNN layer applies a convolution kernel of 5 x 5, followed by a max pooling layer with kernel of 2 x 2 to reduce the image to half of its size using the maximum function. This inception layer learns 5 different kernels and outputs five feature maps. The second CNN layer applies a new convolution to each of the five feature maps from the first layer using 5 different filters of 5 x 5 followed by another max-pooling using a filter size of 2 x 2. The output from both channels converges to a fully connected hidden layer that has 256 neurons which connects all neurons of the previous layer to every neuron of its own layer. Classification is based on the class with the maximum probability.

IV. Experimental Results

The proposed system runs on a 3.0GHz Intel i5-2320 CPU, 8GB RAM and developed on MATLAB 2017 software. The performance of the proposed method is evaluated using the classification accuracy obtained from experiments performed on the JAFFE and the CK+ datasets.

A. JAFFE Dataset

Fig.4. A sample of images from the JAFFE dataset with their corresponding emotions.

The JAFFE dataset contains of 213 grayscale facial expression images of 7 facial expressions -disgust, Fear, Joy, surprise, sadness, anger, and neutral, posed by 10 Japanese females. A sample of images from the dataset is shown on figure 4. Each image is of size 256 x 256[9]. The images in this database were divided into 149 images (70%) for Training and 64 images (30%) for testing.

The network is trained for classification using the training subset of 70% while the testing subset of 30% is used to test the probability that a given facial image belongs to a particular facial expression class. Training of the CNN is achieved in a supervised manner using the standard back-propagation algorithm. The average recognition accuracy is used to evaluate the performance of the network. Image preprocessing took an average of 0.1s while image classification took an average of 0.2s. Table I shows the confusion matrix of the recognition accuracy for seven facial expressions using the training weights obtained with the best accuracy. The recognition accuracy achieved using this method is 97.71%.

Table I. Confusion Matrix on JAFFE

	Neutral	Surprise	Angry	Disgust	Fear	Happy	Sad
Neutral	97.6%	0.0%	0.3%	1.2%	0.3%	0.2%	0.0%
Surprise	0.0%	98.0%	0.1%	0.0%	0.5%	1.2%	0.2%
Angry	0.8%	0.0%	97.6%	0.9%	0.2%	0.4%	0.9%
Disgust	0.0%	0.0%	1.3%	98.3%	0.0%	0.2%	0.2%
Fear	0.3%	1.0%	0.7%	0.7%	96.6%	0.3%	0.7%
Happy	1.6%	1.5%	0.0%	0.0%	1.6%	97.9%	0.0%
Sad	1.0%	0.5%	0.3%	0.1%	1.1%	0.0%	98.0%

The method gives a high recognition accuracy of at least 98% for surprise, disgust, happy and sad while Angry, Neutral and Fear had lower accuracy between 96.6% – 97.6%.

B. CK+ Dataset

The CK+ dataset consists of 593 image sequences from 123 subjects from ages 18-50 of both male and females from different ethnic groups comprising of 69% female and 19% from African America and Asians [1].

Fig.5. A sample of images from the CK+ dataset with their corresponding emotions.

Here we use 350 images of size 640 x 480 or 640 x 490 which were randomly selected from 50 subjects in the database. The contempt expression in the dataset was not used in the experiment. An illustration of the datasets is seen in Fig. 5. A 10-fold cross validation was carried out using the selected 350 images, 245 images (70%) were used for training and 105 images (30%) were used for testing. The network is trained for classification using the training subset and tested using the testing subset. The CK+ dataset is used to show the robustness of the network. Table II shows the confusion matrix of the recognition accuracy for seven facial expressions on the CK+ dataset.

Table II. Confusion Matrix on CK+

	Neutral	Surprise	Angry	Disgust	Fear	Happy	Sad
Neutral	98.7%	0.4%	1.2%	1.1%	0.3%	0.5%	0.0%
Surprise	0.0%	98.4%	0.5%	0.0%	0.4%	0.9%	0.1%
Angry	1.3%	0.0%	93.7%	0.5%	0.0%	0.5%	3.6%
Disgust	0.0%	0.0%	2.3%	97.2%	0.0%	0.2%	0.2%
Fear	0.3%	1.0%	2.3%	3.5%	92.2%	0.5%	0.7%
Happy	0.0%	2.3%	0.0%	0.0%	1.6%	98.1%	0.0%
Sad	0.2%	0.1%	4.1%	0.0%	4.0%	0.0%	91.8%

The CK+ dataset achieved a recognition accuracy of 95.71%. This result is lower than that achieved by using the JAFFE because of the varied nature of the images in the

CK+ dataset which were captured under more challenging pose and illumination conditions. This results in a number of the Sad expressions being confused with Anger or Fear expressions. The recognition accuracy for the different expressions using the CK+ dataset shows that the CK+ dataset achieved a higher recognition for the neutral and surprise expressions than that achieved with the JAFFE dataset but slightly lower accuracy for angry, disgust, fear, happy and sad ranging from 91.8% - 98.1%.

The average recognition accuracy of the proposed method is compared to other methods for facial expression recognition. Table III and Table IV shows the comparison of the performance accuracy obtained with this method and that of other methods on the JAFFE and CK+ database respectively. The proposed method is seen to achieve superior recognition accuracy when compared to other methods already existing in literature.

Table III. Comparison on JAFEE Dataset

Approach	Recognition Accuracy
SVM[5]	95.60%
Gabor[9]	93.30%
2-Channel CNN[4]	94.40%
Proposed Method	97.71%

Table IV. Comparison on CK+ Dataset

Approach	Recognition Accuracy
SVM[15]	95.10%
Gabor[9]	90.62%
3DCNN[16]	95.00%
Proposed Method	95.72%

V. Conclusion and Future Work

This work presents a new convolution neural network architecture that uses feature information from facial parts (Eyes and Mouth) as input into two separate CNN channels. The output from the two channels converges into a fully connected layer and the result used for classification. This method is aimed to have an advantage over using the whole face as an input by having an increased recognition accuracy and reduced cost. Experimental results based on the JAFFEE and a dataset created from randomly selected image samples from the CK+ database confirms the effectiveness and robustness of this method. It is shown that our proposed method can achieve the average expression recognition accuracy of 97.71% and 95.72% respectively for the two datasets. Another interesting aspect of this work that can be explored in the future would be to test this approach on more databases.

References

[1] C. Ruoxuan, L. Minyi, and L. Manhua "Facial Expression Recognition Based on Ensemble of Multiple CNNs," CCBR 2016, LNCS 9967, pp. 511-578, Springer International Publishing AG 2016

[2] P. Christopher and M. Kampel "Facial Expression Recognition using Convolutional Neural Networks: State of the art, "Computer Vision Lab, Vienna, Austria, arXiv:1612.02903v1, 9 Dec. 2016

[3] Arushi Raghuvanshi and Vivek Choksi, "Facial Expression Recognition with Convolutional Neural Networks", *CS231n Course Projects*, Winter 2016.

[4] Dennis Hamester et al., "Face Expression Recognition with a 2-Channel Convolutional Neural Network", *International Joint Conference on Neural Networks (IJCNN)*, 2015.

[5] H.C.Santiago, T.Ren,and G. D.C. Cavalcanti "Facial expression Recognition based on Motion Estimation," Neural Networks (IJCNN), 2016 International Joint Conference , Electronic ISSN:2161-4407. 03 November 2016

[6] J. Li, and E.Y . Lam "Facial expression recognition using deep neural networks,"Imaging Systems and Techniques (IST), 2015 IEEE International Conference on, 1-6

[7] Y. Muttu, H. G. Virani, "Effective face detection feature extraction neural network based approaches for facial expression recognition", *IEEE International Conference on Information Processing (ICIP)*, pp. 102-107, Dec 2015.

[8] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes and S. Wermter, "Understanding how deep neural networks learn face expressions," in *IEEE International Joint Conference on Neural Networks (IJCNN),* 2016.

[9] S.A.M Al-Sumaidaee, Facial Expression Recognition Using Local Gabor Gradient Code-Horizontal Diagonal Dedscriptor" School of Electrical and Electronic Engineering, Newcastle University, England, UK, 2015.

[10] Zhang, T., Zheng, W., Cui, Z., et al.: A deep neural network driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. **99**, 1 (2016)

[11] Kahou, Samira Ebrahimi, et al. "Combining modality specific deep neural networks for emotion recognition in video." Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013.

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting Large,Richly Annotated Facial-Expression Databases from Movies", IEEE Multimedia, vol. 19, no. 3, pp. 34–41, 2012.

[13] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. ICMI Proceedings.

[14] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10, IEEE, 2016

[15] Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)

[16] Byeon, Y.-H., Kwak, K.-C.: Facial expression recognition using 3d convolutional neural network. Int. J. Adv. Comput. Sci. Appl. **5**(12) (2014)

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proc. of CVPR, 2001.

[18] J. Lu, X. Fu, and T. Zhang, "A Smart System for Face Detection with Spatial Correlation Improvement in IoT Environment" in 2017 IEEE conference on ubiquitous intelligence and computing, 2017, in press.