Real Time Emotion Recognition from Facial Expressions Using CNN Architecture

Mehmet Akif OZDEMIR[1],Berkay ELAGOZ[1], AysegulALAYBEYOGLU[2], Reza SADIGHZADEH[3]and Aydin AKAN[1]
[1]Department of Biomedical Engineering, [2]Department of Computer Engineering, [3]Business Administration
Izmir Katip Celebi University
Izmir, Turkey
makif.ozdemir@ikc.edu.tr, berkayelagoz@gmail.com, aysegul.alaybeyoglu@ikc.edu.tr, riza@taimaksan.com, aydin.akan@ikc.edu.tr

***Abstract*—Emotion is an important topic in different fields such as biomedical engineering, psychology, neuroscience and health. Emotion recognition could be useful for diagnosis of brain and psychological disorders. In recent years, deep learning has progressed much in the field of image classification. In this study, we proposed a Convolutional Neural Network (CNN) based LeNet architecture for facial expression recognition. First of all, we merged 3 datasets (JAFFE, KDEF and our custom dataset). Then we trained our LeNet architecture for emotion states classification. In this study, we achieved accuracy of 96.43% and validation accuracy of 91.81% for classification of 7 different emotions through facial expressions.**

Keywords—Convolutional Neural Network; Deep Learning; Emotion Recognition; Facial Expressions, Real Time Detection.

I. Introduction

Although there are many studies in the literature on emotion, there is no common or singular definition in the literature about emotion[1]. Emotion is the appearance or reflection of a feeling. Distinct from feeling, emotion can be either real or sham. For example, feeling of pain can directly represent the feeling. But emotions are not felt exactly. Emotions present inner situations psychologically [2, 3].

Emotion is an important, complex and extensive research topic in the fields of biomedical engineering [4], psychology [5], neuroscience [6] and health [7]. Emotion detection is an important research area in biomedical engineering. Studies in this area focus on predicting human emotion and computer-assisted diagnosis of psychological disorders. There are different methods in literature to detect emotional states such as electroencephalography (EEG), galvanic skin response (GSR), speech analysis, facial expression, multimodal, visual scanning behavior [8-10].

In recent years, with the popularization of deep learning, great progress has been made in image classification. Convolutional neural networks (CNNs) is an artificial neural network type that proposed by Yann LeChun in 1988 [11]. Convolutional neural networks are one of the most popular deep learning architectures for image classification, recognition, and segmentation.

Convolutional neural networks built like a human brain with artificial neurons and consist of hierarchical multiply hidden layers. These artificial neurons take input from image, multiply weight, add bias and then apply activation function. So that, artificial neurons can be used in image classification, recognition, and segmentation by perform simple convolutions. By feeding the convolutional neural network with more data (huge amount of data), a better and highly accurate deep learning model can be achieved.

Deep learning based facial expression recognition is one of these methods to detect emotion state (e.g., anger, fear, neutral, happiness, disgust, sadness and surprise) of human. This method aims to detect facial expressions automatically to identify emotional state with high accuracy. In this method, labeled facial images from facial expression dataset are sent to CNN and CNN is trained by these images. Then, proposed CNN model makes a determination which facial expression is performed.

Chang et al. used CNN model based on ResNet to extract feature from Fer2013 and CK+ dataset. Proposed complexity perception classification algorithm (CPC) was applied with different classifiers (Softmax, LinearSVM, and RandomForest). CNN+Softmax with CPC has achieved 71.35% and 98.78% recognition accuracies for Fer2013 and CK+ respectively [12].

Clawson et al. proposed two human centric CNN architecture for facial expression recognition on CK+ dataset. CNN A consists of 1 convolutional layer and 1 max pooling layer. CNN B consists of 2 convolutional layers and 2 max pooling layers. These architectures trained with 0.0001 initial learning rate, 300 epochs and 10 batch size. According to results, proposed model has achieved 93.3% accuracy on CK+ images [13].

Nguyen et al. proposed multi-level18-layer CNN model similar to VGG. These model does not take only high-level features also takes mid-level features. Plain CNN model has reached 69.21% accuracy and proposed multi-level CNN model has reached 73.03% accuracy on Fer2013 dataset [14].

Cao et al. proposed CNN model with K-means clustering idea and SVM classifier which has achieved 80.29% accuracy. K-means clustering model determines initial value of the convolution kernel of CNN. SVM layers takes features from trained CNN model to classify Fer2013 images [15].

978-1-7281-2420-9/19/$31.00 ©2019 IEEE

Ahmed et al. merged different facial expression datasets which are CK, CK+, Fer2013, the MUG facial expression database, KDEF, AKDEF, and KinFaceW-I/II. Data augmentation was applied merged dataset. Proposed CNN model consists of 3 convolutional layers with 32, 64, 128 filters and kernel size are 3x3. According to results, proposed model has reached 96.24% accuracy [16].

Christou et al. proposed 13 layer CNN model that used on Fer2013 dataset and achieved 91.12% accuracy on validation dataset [17].

Sajjanhar et al. worked on CK+, JAFFE and FACES datasets. They trained and used pre-trained CNN models such as Inception-V3, VGG-16, VGG-19 and VGG-Face. According to results, highest accuracy (97.16%) was obtained with VGG-19 model on FACES dataset [18].

Chen et al. proposed two-stage framework based on Difference Convolutional Neural Network (DCNN) that trained with CK+ and BU-4DFE datasets. Results showed proposed model achieved 95.4% accuracy on CK+ dataset and 77.4% on BU-4DFE [19].

In this study, we proposed CNN based LeNet architecture for facial expression recognition to estimate emotion states of human. We merged 3 different datasets (KDEF, JAFFE and our custom dataset). Then, proposed LeNet architecture was trained with final dataset for classification of 7 emotion states (happy, sad, surprised, angry, disgust, afraid and neutral). The aim of the study is to obtain deep learning model that achieve higher accuracy rate for emotion recognition through facial expression.

II. Methods

A. *Facial Expression Dataset*

There are many open accesses facial expression dataset in literature. We used 3 facial expression datasets. These are JAFFE, KDEF and our custom dataset.

JAFFE dataset contains 213 images with 7 facial expressions (happy, sad, surprised, angry, disgust, afraid and neutral). These images were taken from 10 Japanese female models [20]. An example of images from JAFFE dataset are shown in Fig. 1.

KDEF dataset contains 4900 images with 7 facial expressions (happy, sad, surprised, angry, disgust, afraid and neutral). Participants are 35 males and 35 females. Dataset contains 5 different angles. We used only straight position in this study [21]. An example of images from KDEF dataset are shown in Fig. 2.

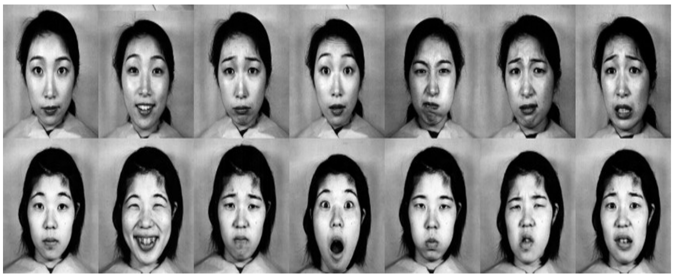

Fig.1. Example of images from JAFFE facial expression dataset.

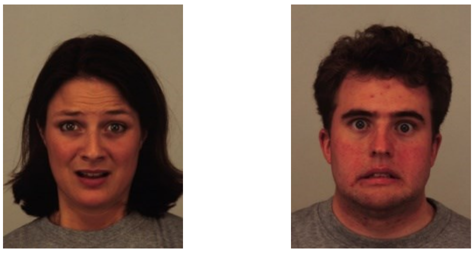

Fig.2. Example of images from KDEF facial expression dataset.

Our custom dataset contains 140 images with 7 facial expressions (happy, sad, surprised, angry, disgust, afraid and neutral). Participants are 1 male and 1 female. Each facial expression was expressed 10 times by 1 participant.

B. *Image Preprocessing*

Containing approximately equal numbers of face images which is seven different facial expressions were different resolutions, because of there were 3 different databases. Therefore, first of all, the face circumference was detected using the *Haar Cascade* library from the pictures. Then, these detected rectangular facial expressions were clipped and recorded to the same size. Also, the pixel values in the images were converted to gray images size of 64x64 to be placed in neural networks. This process was done to avoid unnecessary density in the neural networks.

C. Convolutional *Neural* Network Architecture

With the proposed CNN architecture, it is aimed to educate the pixel values in the rectangular region containing facial expressions quickly and functionally and to make quick queries with the deep artificial neural network model formed. The proposed CNN structure is summarized in Fig. 3. The network mimics the LeNet structure used in classification of 2D facial expression data and includes the two convolutional layers, two max-pooling layers, and one fully connected layer. The convolutional layers with kernel size of 2x2 are stacked together which are followed by max-pooling layer with kernel size of 2x2 and stride of 2. After all operations of convolutional layers and max-pooling layers, each frame feeds to the fully connected layers and prediction of frames was processed with Softmax classifier as seven different facial emotional state.

D. *Network Training*

In training of network, test size determined as 25%. Batch size has been set as 32 and epoch number was found as 500 to converge parameters of network. Learning rate defined as 10^{-3}. All kernel size defined as 2x2 with stride of 2 for convolutional layers and max-pooling layers respectively. Number of convolutional layers represented as 16 and 32 respectively. Summary of proposed CNN architecture is shown in Table I.

E. *Real Time Testing*

After training of proposed CNN architecture, the trained model was tested in real time. First of all, human faces were detected with the *Haar Cascade* library within 30 images per second of the computer camera. After that, the detected images were sent to the model and the classes they belong to were queried. As a result of the predictions, the possibility of belonging to which class the facial expression was shown on a

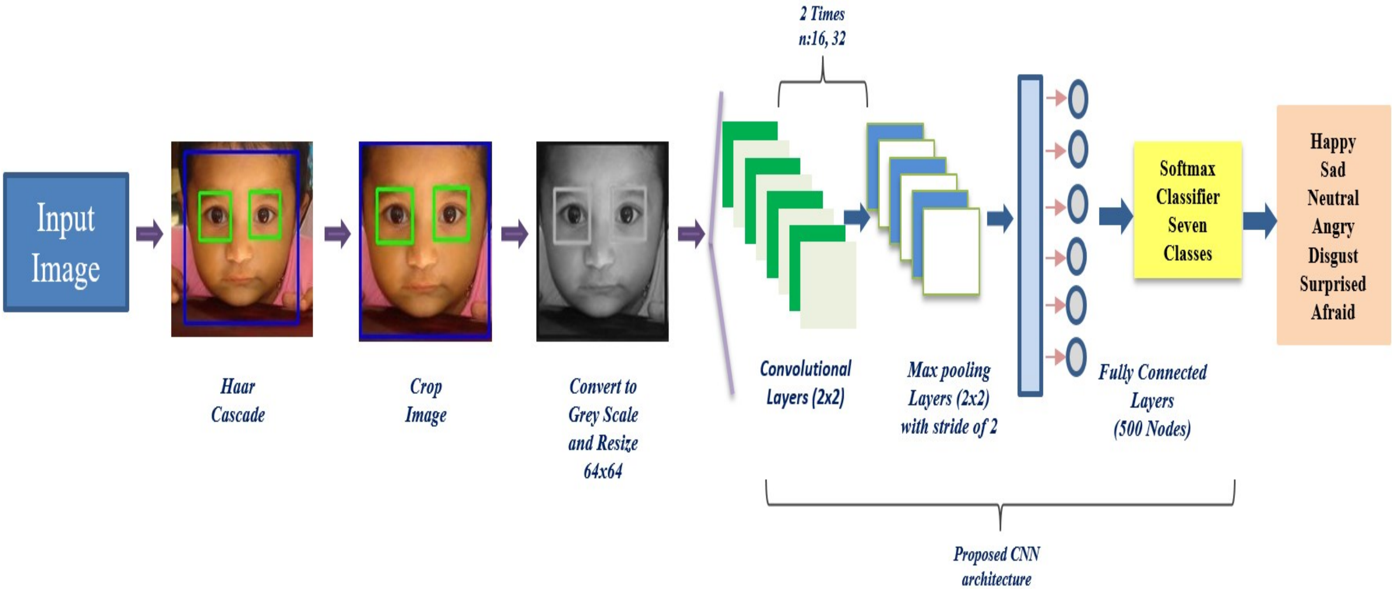
2 Times
n:16, 32
Input Image
Haar Cascade
Crop Image
Convert to Grey Scale and Resize 64x64
Convolutional Layers (2x2)
Max pooling Layers (2x2) with stride of 2
Fully Connected Layers (500 Nodes)
Softmax Classifier Seven Classes
Happy
Sad
Neutral
Angry
Disgust
Surprised
Afraid
Proposed CNN architecture

Fig. 3. Proposed CNN model diagram for facial emotion recognition.

separate screen and the emotion in which class was higher was overwritten on the *Haar Cascade* frame. This process was performed on every 30 frames that occurred every second of the camera image obtained in real time.

TABLE I. SUMMARY OF PROPOSED CNN ARCHITECTURE

Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 64, 64, 20)	520
activation_1 (Activation)	(None, 64, 64, 20)	0
max_pooling2d_1 (MaxPooling2)	(None, 32, 64, 20)	0
conv2d_2 (Conv2D)	(None, 32, 32, 50)	25050
activation_2 (Activation)	(None, 32, 32, 50)	0
max_pooling2d_2 (MaxPooling2)	(None, 16, 16, 50)	0
flatten_1 (Flatten)	(None, 12800)	0
dense_1 (Dense)	(None, 500)	6400500
activation_3 (Activation)	(None, 500)	0
dense_2 (Dense)	(None, 7)	3507
activation_4 (Activation)	(None, 7)	0
Total params: 6,429,577 Trainable params: 6,429,577 Non-trainable params: 0		
None		

**Conv2D: 2D Convolutional Layer, Maxpooling2: 2D Max pooling Layer*

III. RESULT AND DISCUSSION

In this study, *Keras* and *TensorFlow* libraries were used for training LeNet CNN architecture and prediction of emotion states with proposed deep learning model. Intel I7 8300 CPU was used for all experiments and training custom dataset. Proposed LeNet CNN model was set with mentioned parameters. Fig. 4. shows performance metrics (training accuracy and training loss, validation accuracy and validation loss) of proposed architecture during training and testing. According to experiment results, training loss was found 0.0887; training accuracy was found 96.43%; validation loss was found 0.2725 and validation accuracy was found 91.81%. When we look at our results, we get better results than mentioned studies in introduction section [9,11,12,14,16].

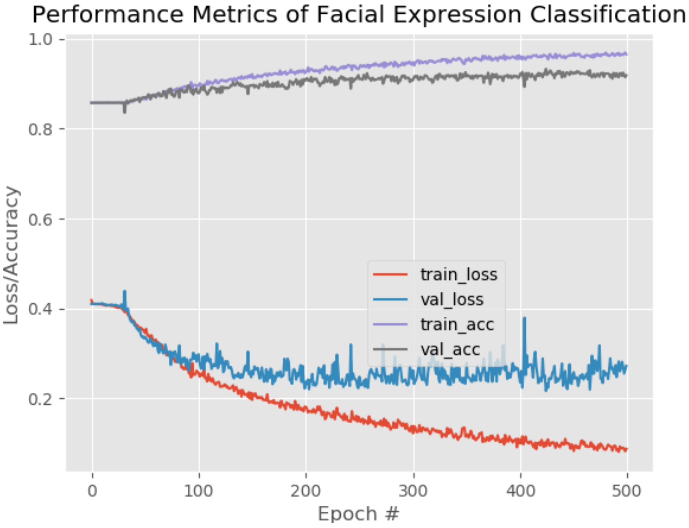

Fig. 4. Performance metrics of proposed architecture.

According to Fig. 5. confusion matrix, proposed LeNet model more accurate at prediction of surprised, fear, neutral emotion states and less accurate at prediction of sad emotion state.

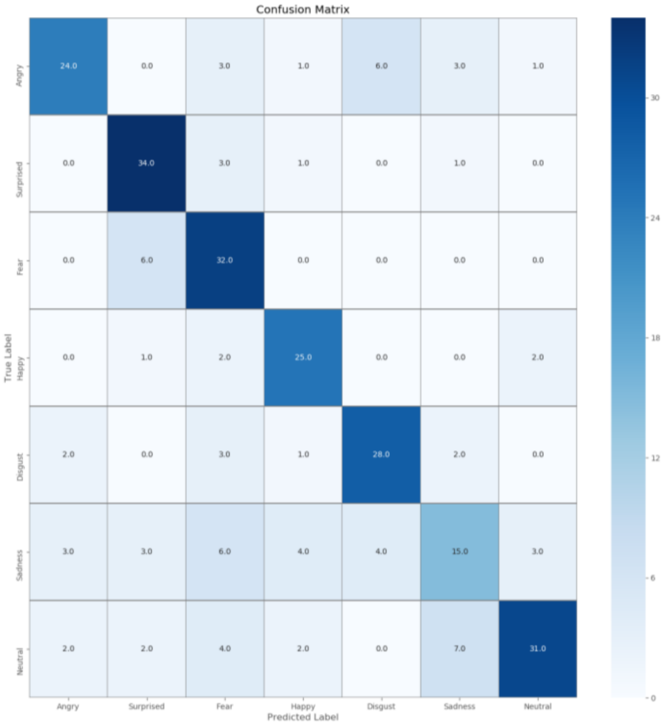

Fig. 5. Confusion matrix of propoesed architecture.

IV. CONCLUSION

This paper proposed a low cost and functionality method for real time classification seven different emotions by facial expression based on LeNet CNN architecture. In this study, facial expression pictures, which can be said has a small number, were successfully trained in CNN and achieved high classification accuracy. Using the *Haar Cascade* library, the effect of unimportant pixels which is outside facial expressions was reduced. In addition, single-depth placement of the pixels in the pictures to networks did not only result in loss of success rate, but also reduced training time and number of networks. Using a custom database has provided higher validation and test accuracy than training in existing databases. The real-time test model has the functionality to query each image that occurs in every second.

Emotion estimation from facial expressions is the area of interest of many researchers in the literature. It is hoped that this study will be a source of studies that will help in the early detection of diseases from facial expressions and also studies of consumer behavior analysis.

ACKNOWLEDGMENT

This work was supported by Scientific Research Projects Coordinatorship of Izmir Katip Celebi University. Project number: 2019-ÖNAP-MÜMF-0003.

REFERENCES

[1] M. Cabanac, "What is emotion?," Behavioural processes, vol. 60, pp. 69-83, 2002.

[2] R. Roberts, "What an Emotion Is: a Sketch," The Philosophical Review, vol. 97, 1988.

[3] E. Shouse, "Feeling, emotion, affect," M/c journal, vol. 8, no. 6, p. 26, 2005.

[4] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312-323, 2019.

[5] J. M. B. Fugate, A. J. O'Hare, and W. S. Emmanuel, "Emotion words: Facing change," Journal of Experimental Social Psychology, vol. 79, pp. 264-274, 2018.

[6] J. P. Powers and K. S. LaBar, "Regulating emotion through distancing: A taxonomy, neurocognitive model, and supporting meta-analysis," Neuroscience & Biobehavioral Reviews, vol. 96, pp. 155-173, 2019.

[7] R. B. Lopez and B. T. Denny, "Negative affect mediates the relationship between use of emotion regulation strategies and general health in college-aged students," Personality and Individual Differences, vol. 151, p. 109529, 2019.

[8] S. Albanie, A. Nagrani, A. Vedaldi, and A. J. a. p. a. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," pp. 292-301, 2018.

[9] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5866-5870: IEEE.

[10] M. Degirmenci, M. A. Ozdemir, R. Sadighzadeh, and A. Akan, "Emotion Recognition from EEG Signals by Using Empirical Mode Decomposition," in 2018 Medical Technologies National Congress (TIPTEKNO), 2018, pp. 1-4.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, pp. 2278-2324, 1998.

[12] T. Chang, G. Wen, Y. Hu, and J. Ma, "Facial Expression Recognition Based on Complexity Perception Classification Algorithm," arXiv e-prints, Accessed on: February 01, 2018 Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180300185C

[13] K. Clawson, L. Delicato, and C. Bowerman, "Human Centric Facial Expression Recognition," 2018.

[14] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I. Na, and S. H. Kim, "Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks," International Journal of Pattern Recognition and Artificial Intelligence, 2018.

[15] T. Cao and M. Li, "Facial Expression Recognition Algorithm Based on the Combination of CNN and K-Means," presented at the Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 2019.

[16] T. Ahmed, S. Hossain, M. Hossain, R. Islam, and K. Andersson, "Facial Expression Recognition using Convolutional Neural Network with Data Augmentation," pp. 1-17, 2019.

[17] N. Christou and N. Kanojiya, "Human Facial Expression Recognition with Convolution Neural Networks," Singapore, 2019, pp. 539-545: Springer Singapore.

[18] A. Sajjanhar, Z. Wu, and Q. Wen, "Deep learning models for facial expression recognition," in 2018 Digital Image Computing: Techniques and Applications (DICTA), 2018, pp. 1-6: IEEE.

[19] J. Chen, Y. Lv, R. Xu, and C. Xu, "Automatic social signal analysis: Facial expression recognition using difference convolution neural network," Journal of Parallel and Distributed Computing, vol. 131, pp. 97-102, 2019.

[20] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," pp. 200-205, 1998.

[21] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces (KDEF)," CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, vol. 91, p. 630, 1998.