Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 2, 2021, 259-268.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Facial Expression Recognition based on Image Gradient and Deep Convolutional Neural Network

Mohammad Reza Falahzadeh[1], Fardad Farokhi[1], Ali Harimi[2*] and Reza Sabbaghi-Nadooshan[1]

1. Department of Technical and engineering, Central Tehran Branch, Islamic Azad University, Iran.
2. Department of Technical and engineering, Shahrood Branch, Islamic Azad University, Iran.

Article Info

Article History:
Received 15 July 2020
Revised 13 December 2020
Accepted 03 January 2021

DOI:10.22044/jadm.2021.9898.2121

Keywords:
Facial Expression Recognition, Deep Convolutional Neural Network, Three-channel of the Image, AlexNet DCNN, Transfer Learning.

**Corresponding author: a.harimi@iau-shahrood.ac.ir (A. Harimi).*

Abstract

Facial expression recognition (FER), which is one of the basic ways of interacting with machines, has attracted much attention in the recent years. In this paper, a novel FER system based on a deep convolutional neural network (DCNN) is presented. Motivated by the powerful ability of DCNN in order to learn the features and image classification, the goal of this research work is to design a compatible and discriminative input for pre-trained AlexNet-DCNN. The proposed method consists of 4 steps. First, extracting three channels of the image including the original gray-level image in addition to the horizontal and vertical gradients of the image similar to the red, green, and blue color channels of an RGB image as the DCNN input. Secondly, data augmentation including scale, rotation, width shift, height shift, zoom, horizontal flip, and vertical flip of the images are prepared in addition to the original images for training DCNN. Then the AlexNet-DCNN model is applied in order to learn the high-level features corresponding to different emotion classes. Finally, transfer learning is implemented on the proposed model, and the presented model is fine-tuned on the target datasets. The average recognition accuracies of 92.41% and 93.66% are achieved for the JAFFE and CK+ datasets, respectively. The experimental results on two benchmark emotional datasets show a promising performance of the proposed model that can improve the performance of the current FER systems.

1. Introduction

Facial Expression Recognition (FER) is a topic in the field of pattern recognition, which includes a wide range of applications such as multimedia, treatment of mentally retarded patients, human-computer interaction, and lie detection [1]. The human face is one of the most important non-verbal communication tools, and its variations indicate the human's inner state, feeling, thinking or even illness [2]. FER may face challenges such as difficulties in finding faces in the scenes with complex and varied backgrounds, face rotation, exit the part of the face from camera view, and extraction of facial features with natural and artificial features such as beard or glasses [3]. Accordingly, reaching an accurate FER system is often a challenging task, and requires a successful extraction of the features and the appropriate classification schemes.

In the recent years, many efforts have been made in order to develop methods for FER. However, there are still many challenges in increasing the efficiency of these systems [4]. The major purpose of an FER system is to identify the basic emotions such as anger, fear, happiness, disgust, neutral, sadness, and surprise from the human face [5]. The previous research works in the field of FER have generally focused on extracting features such as SIFT and LBP and classification algorithms such as SVM, HMM, and GMM [6-11]. In the recent years, with the advent of deep learning networks and the acceptable results of these networks in the field of image processing, many research works have been performed on feature learning and image classification with these networks [1, 12-14]. Many deep learning algorithms such as Deep Belief Networks (DBNs), Deep Convolutional Neural Networks (DCNNs), and Long Short-Term Memory (LSTM) have been introduced in order to learn the features of FER [5, 15, 16]. Motivated by the promising performance of deep learning algorithms [17, 18], in this work, we aimed to employ a deep learning model to develop an effective facial expression recognition system. In this way, the success of DCNNs in image classification motivated us to use AlexNet-DCNN in our FER system. In order to address this issue, first, we have to provide an appropriate input for AlexNet-DCNN, which is pre-trained on the large ImageNet dataset [19]. This image should be quite compatible in shape, size, and range of numerical values to AlexNet in order to use the transfer learning technique. AlexNet can learn high-level features from the input image and classify them according to the corresponding emotions. However, instead of the red, green, and blue layers of an ordinary RGB color image, our input image is composed of the original gray-level image, its horizontal gradient Gx, and its vertical gradient Gy. Extensive experiments on the two public emotional datasets JAFFE [20] and CK+ [21] demonstrate the promising performance of our proposed method.

The main contribution of this work is to develop a pre-processing stage providing a proper input for AlexNet-DCNN. Since face wrinkles convey important emotional features, we use horizontal and vertical gradients of the original image containing image details as the second and third layers of the network input. In other words, the proposed Gx and Gy are not only useful in providing a fit-size input for AlexNet but also enrich the network input in terms of the important emotional cues of the face image. Finally, we tuned the AlexNet-DCNN model under the proposed model. The remainder of this paper is organized as what follows. The proposed method is described in Section 2. In Section 3, the datasets used are introduced. In Section 4, the experimental results and simulation are shown. Finally, the conclusions and suggestions for the future works are presented.

2. Proposed Method

The proposed facial expression recognition method used in this work is based on a DCNN. The suggested model is shown in Figure 1. This method consists of four steps: generation of the DCNN input, data augmentation, the AlexNet-DCNN model, and transfer learning on the proposed model for emotion classification.

2.1. Generation of DCNN Input

The RGB color images contain 3 matrices of red, green, and blue. In FER, colors do not contain important information, so most papers have done image processing in the Grayscale mode in the field of FER [2]. On the other hand, most benchmark emotional datasets have a limited number of samples. Using DCNNs and obtaining the desirable results, the datasets with a large number of samples are required [22]. In order to tackle this problem, the Transfer Learning (TL) method is proposed [23-26]. This method uses the weights of the networks that have trained on large datasets, and good results have been achieved. Using the weights of the pre-trained network on the proposed model partially solves the problem of the small-size dataset .The pre-trained networks have been trained on the large ImageNet datasets with over one million samples such as AlexNet [19], VGG [27], ResNet [28], Inception (GoogleNet) [29], and DensNet [30]. The inputs of these networks are the RGB images consisting of the 3 channels R, G, and B. Hence, the inputs compatible with the networks must be provided to optimally utilize these networks and obtain accurate results. Accordingly, in the proposed model, the original image gradient x, and gradient y constitute three channels similar to the red, green, blue (RGB) image representation as to the DCNN input. This input is compatible with the input of the AlexNet pre-trained network since different emotions arise with changes in different parts of the face in the horizontal axis direction (x-axis) and the vertical axis direction (y-axis). It is expected that using the gradient x and the gradient y along with the original image can provide a useful emotional information. In this method, we extract three channels of image (Origin Image, Image Gradient X, and Image Gradient Y) similar to the RGB image representation in a way that is compatible with the AlexNet-DCNN input. This input is a matrix consisting of 3 channels, where the channel arrays are numeric in the range of 0 to 255. Figure 2 shows the Origin Image, Image Gradient X, and Image Gradient Y in the CK+ and JAFFE datasets. Figure 3 shows the generation of the AlexNet DCNN compatible Input.

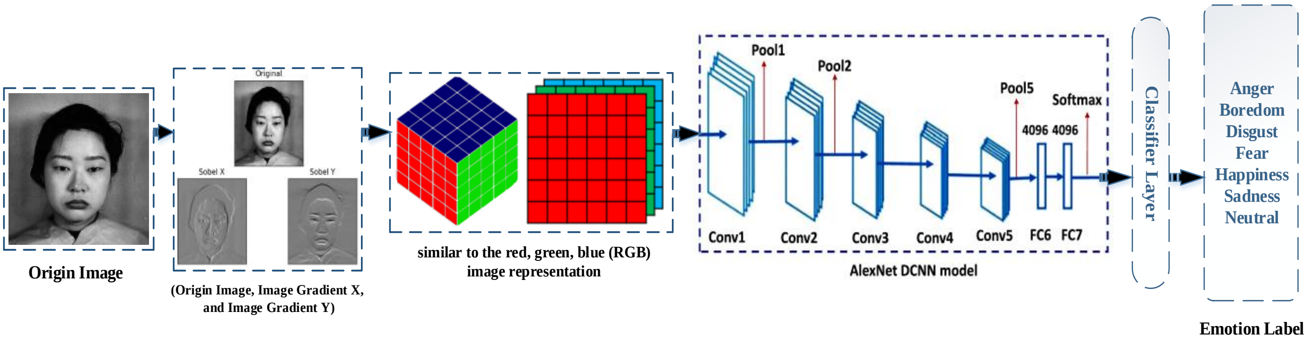


Figure 1. An overview of the proposed method for facial expression recognition: (1) Three channels of image are extracted similar to the RGB image representation. (2) The pre-processing stage and resized are performed, and the compatible network input of Alexnet-DCNN is provided. (3) The AlexNet-DCNN model pre-trained on large ImageNet Dataset is applied for learning the high-level features. (4) TL and fine-tuning is done in order to classify different emotions.

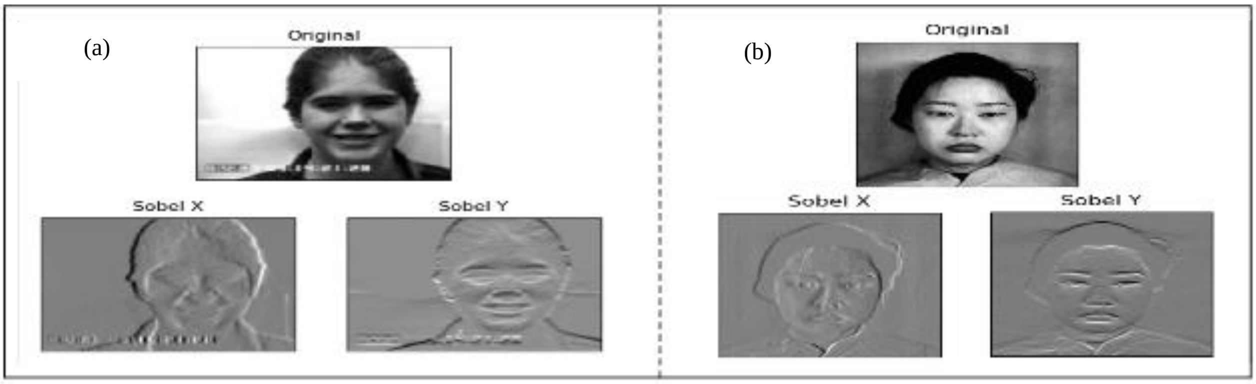


Figure 2. Original Image, Image Gradient X, and Image Gradient Y in the (a) CK+, and (b) JAFFE datasets.

Estimation of The intensity gradient at a pixel in the x and y direction, for an image f, is given by:

$$\frac{\partial f}{\partial x} = f(x+1, y) - f(x-1, y) \qquad (1)$$

$$\frac{\partial f}{\partial y} = f(x, y+1) - f(x, y-1) \qquad (2)$$

The image edges are the areas with strong intensity contrasts. The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image. The edges can be detected using the Sobel filter. This filter detects the edges by calculating the gradient of the image on each pixel. The Sobel filter has two 3 * 3 kernels for horizontal and vertical changes (Gx and Gy). The Gx and Gy computations are:

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} * f \quad G_y = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} * f \qquad (3)$$

Where Gx and Gy are the horizontal and vertical gradients of the original image, f. The Sign, *, denotes the convolution operator. In order to compute Gx and Gy, the appropriate kernel (window) moves over the input image, computing the value for one pixel and then shifting one pixel to the right. Once the end of the row is reached, we move down to the beginning of the next row.

2.2. Data Augmentation

Data augmentation means increasing the number of samples by making changes in the current data and adding new data to the original input signals. Data augmentation is a powerful way to make data robust against some of the challenges such as the difficulty of finding faces in different scenes as well as reducing the risk of over-fitting [31]. In this work, we used scale, rotation, width shift, height shift, zoom, horizontal flip, and vertical flip of the images in addition to the original images for training DCNN. Figure 4 shows the examples of data augmentation in the JAFFE dataset related to the disgust emotion.

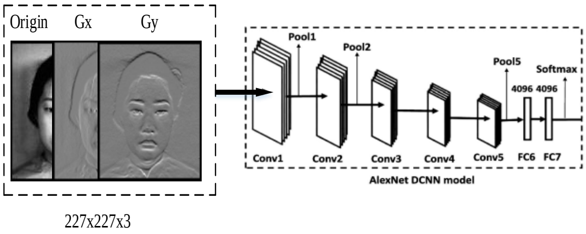


Figure 3. Generation of AlexNet DCNN compatible input.

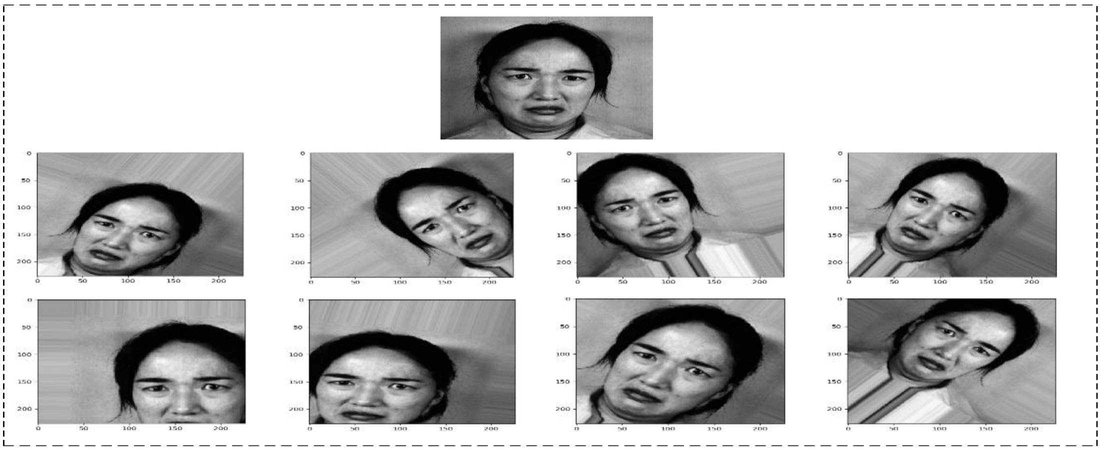


Figure 4. Examples of data augmentation in the JAFFE dataset related to disgust

2.3. AlexNet-DCNN Model

At this stage, first, pre-processing is performed since the input size of AlexNet [19] is 227 × 227 ×,3; all the dataset samples must be resized to 227 × 227 × 3.

2.3.1 AlexNet Architecture

AlexNet is a DCNN proposed by Alex Krizhevsky and his colleagues. AlexNet won the ImageNet ILSVRC-2012 competition. The The ImageNet ILSVRC competition has been held every year since 2010. Its purpose is to identify and classify the large-scale images. The participating networks must separate images with 1000 different classes. The criterion for measuring the accuracy of networks is the top 5 class error [19]. The AlexNet architecture consists of eight layers, five convolutional layers and three fully connected layers. The network uses two GPUs for processing, and the size of the input image is 227 × 227 × 3. Table 1 presents the specifications of the AlexNet architectural layers.

As shown in this table, In all the neurons' layers and convolutional layers, the Rectified Linear Unit (Relu) [32] activation function is used .The Relu activation function can be expressed as follows:

$$\mathrm{Re}\,lu(x_i) = \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases} \tag{4}$$

Where x_i is the ith input to the current convolution layer. The soft-max classifier is also utilized in order to recognize the emotion according to the learned features. The softmax activation function can be calculated as follows:

$$soft\ \max(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}} \tag{5}$$

Where z_i is the ith weighted input sample after passing the last layer in the DCNN model, and K is the total number of samples in the last layer. In the following, we introduce the computations and principles of a convolutional layer, Transfer Learning, and Fine-tuning, respectively.

Convolutional layer*:* In this layer, the CNN network uses different kernels to convolve the input image and the middle feature maps. This can be denoted as:

$$Z(i,j) = X(i,j) \otimes W(i,j) \tag{6}$$

$$Z(i,j) = X(i,j) \times W(i,j) = \sum_{s=0}^{a-1} \sum_{t=0}^{b-1} x(s,t).w(i-s,j-t) \tag{7}$$

here *Z(i,j)* denotes the *(i,j)* element of the result, *X(i,j)* and *W(i,j)* denotes the input and convolution kernel of size $a \times b$, respectively.

2.4. Transfer Learning (TL)

In order to train the deep learning networks, a large dataset is required. In cases where a small dataset is available, the TL method is recommended. In the TL method, the weights of a trained network on the large dataset are transferred to the target network. Thus, instead of starting from an unknown point, training starts from a point that is closer to the global minimum.

TL is a very effective way to train the deep-learning networks on small datasets using a pre-trained network [23, 24].

Table 1. Specifications of the AlexNet architectural layers[17].

Layer	Feature Map	Size	Kernel Size	Stride	Activation Function
Input	1	227×227×3	-	-	-
Convolution_1	96	55×55×3×96	11×11	4	RELU
Max pool	96	27×27×3×96	3×3	2	-
Convolution_2	256	27×27×3×256	5×5	1	RELU
Max pool	256	13×13×3×256	3×3	2	-
Convolution_3	384	13×13×3×384	3×3	1	RELU
Convolution_4	384	13×13×3×384	3×3	1	RELU
Convolution_5	256	13×13×3×256	3×3	1	RELU
Max pool	256	6×6×3×256	3×3	2	-
Fully connected_1	-	4096	-	-	RELU
Fully connected_2	-	4096	-	-	RELU
Out put	-	7	-	-	SOFTMAX

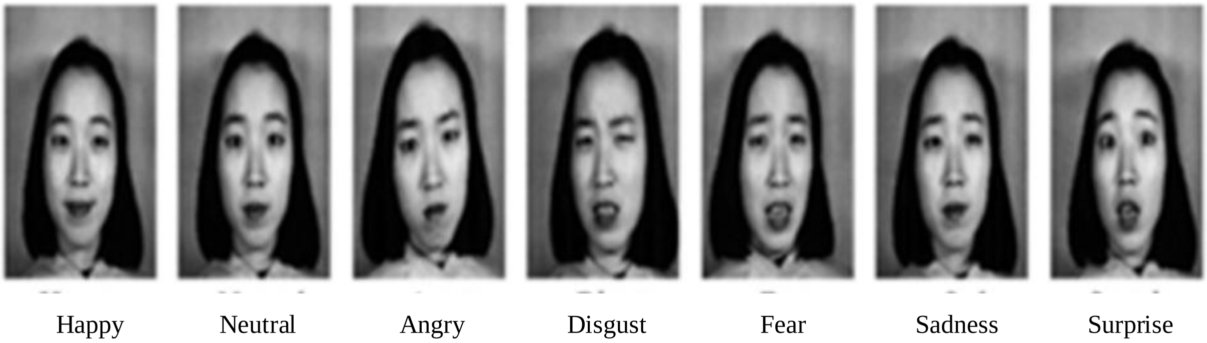


Figure 5. Example of images from the JAFFE dataset with related emotions.

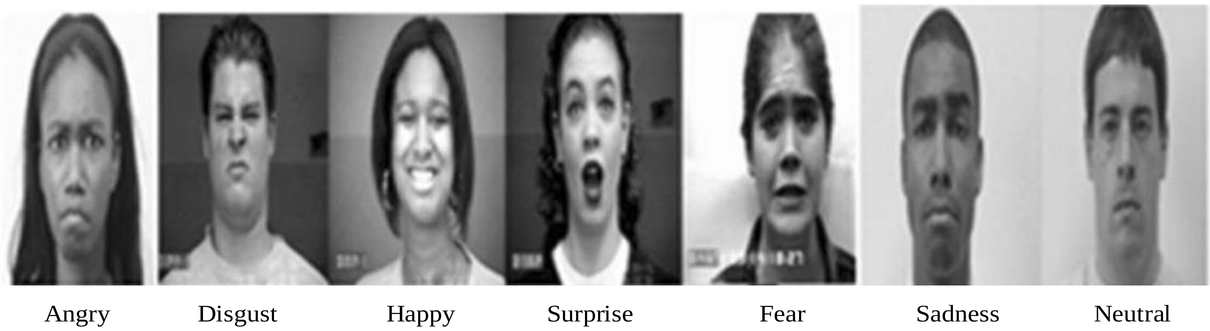


Figure 6. Example of images from the CK+ dataset with related emotions.

2.4.1 Fine-tuning

The purpose of fine-tuning is to fine-tune the parameters on the target network [26]. At this stage, some layers are kept frozen and untrainable as the initial weights and other layers remain unfrozen and trainable.

AlexNet is a pre-trained network on the large ImageNet dataset. At this point, AlexNet is fine-tuned with the target dataset. The AlexNet model has 1000 classes, while our proposed model for facial expression recognition has 7 classes. Thus the fully connected layers are removed, and the network is fine-tuned using a flatten and a dense layer with 7 neurons. In Figure 7, the fine-tuning of the proposed model is presented. As shown in this figure, the layers of conv1-conv4 are frozen and untrainable. In the proposed model, the AlexNet network weights are trained on the Large ImageNet dataset for the conv1-conv4 layers, and the conv5 and fully connected layers are fine-tuned.

3. Emotional Facial Expression Dataset

The proposed method was evaluated on the two public facial expression datasets of JAFFE and CK+. The following is a brief explanation of these datasets.

3.1. JAFFE Dataset

The Japanese Female Facial Expression (JAFFE) [20] dataset is a public domain facial expression dataset that contains 213 images of 10 different individuals in 7 main facial expressions (6 basic facial expressions + 1 natural mode). All images are the Japanese women made by Kamachi and Gyoba from the Kyushu University.

This dataset includes emotions of sadness, anger, happiness, surprise, disgust, fear, and neutral. The images are 256 x 256 gray level, in .tiff format, with no compression [20]. A sample of the images from the JAFFE dataset is shown in Figure 5.

3.2. CK+ Dataset

CK+ [21] is the extended Cohn-Kanade (CK) dataset [33]. This dataset is a complete dataset for the emotion-specified expression. The CK+ database includes 593 image sequences from 123 subjects. The images are of a combination of men and women in the range of 18-50 years. 69% are females and 31% are males. 81% are European-American, 13% are African-American, and 6% are of other races. The images are gray scale and size 640 * 490 Pixel. A sample of images from the CK+ dataset is shown in Figure 6.

4. Experimental Setup and Results

We tested the proposed model on the two benchmark datasets JAFFE and CK+. The proposed method was run on a laptop with specifications of CPU-Intel Core i7-7500U, VGA-GTX 960M (4GB), RAM-16GB DDR4. The programming language was Python3.6 and all simulations and implementations were done in the Spyder environment. The TensorFlow software and Keras library [34] were employed in order to design the DCNN model. The image processing, data augmentation, and image gradient calculation were performed using the OpenCV library. The proposed DCNN model included AlexNet Network with 5 convolution layers. We adopted a dropout technique with a rate of 0.5 to the first fully connected layer. This technique reduces the system complexity and thus the risk of overfitting by the random removal of neurons.

The proposed method was performed on the training set with "Categorical Cross-entropy" as The loss function. Learning rate = 0.001, beta1= 0.9, and beat2 = 0.95). The best model based on the validation accuracy was obtained with 200 epochs.

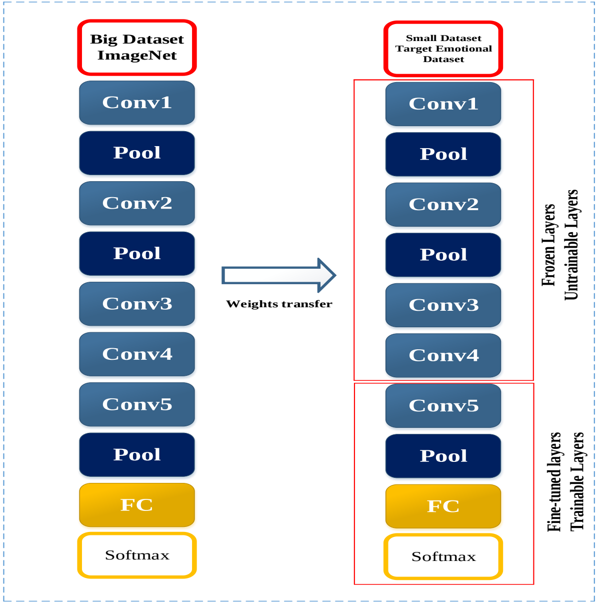


Figure 7. Block diagram of the fine-tuning proposed model.

A 10-fold cross-validation technique was used in order to evaluate the proposed model. According to this method, the database was divided into 10 non-overlapping equal segments. In ten successive independent experiments, one of the 10 segments was selected for validation, while the other 9 segments were employed for training. It guaranteed the contribution of all samples in validation, while training was performed with an acceptable amount of data [35].

Table 2 shows the fine-tuning procedure on the AlexNet-DCNN model. Considering the architecture shown in Figure 7, initially, the convolution_1 layer was frozen and the weights of this layer were untrainable. This meant that the AlexNet network weights were used for the first layer and the other layers remained trainable. The weights of these layers were randomly selected and trained. In this step, the accuracy recognition rates were obtained to be 76.25% and 78.68% for the JAFFE and CK+ datasets, respectively.

In another experiment, the convolution_1 layer and the convolution_2 layer were selected untrainable (frozen), and the remaining layers were kept trainable. In this experiment, the accuracy recognition rates were obtained at 76.40% and 79.30% for the JAFFE and CK+ datasets, respectively. In the third experiment, the first 3 layers were untrainable and the rest of the layers were trainable. In this case, the results obtained were better than the other two experiments.

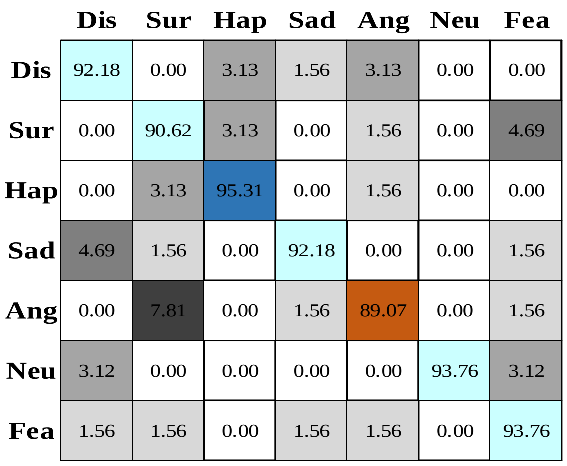
Table 2. Recognition accuracies (%) using fine-tuned AlexNet on the two datasets.

Method	JAFFE	CK+
Frozen Conv1 & Trainable other layers	**76.25%**	**78.68%**
Frozen Conv1+Conv2 & Trainable other layers	**76.40%**	**79.3%**
Frozen Conv1+Conv2+Conv3 & Trainable other layers	**80.65%**	**81.12%**
Frozen Conv1+Conv2+Conv3 +Conv4 & Trainable other layers	**83.35%**	**84.86%**

Table 3. Best Recognition accuracies (%).

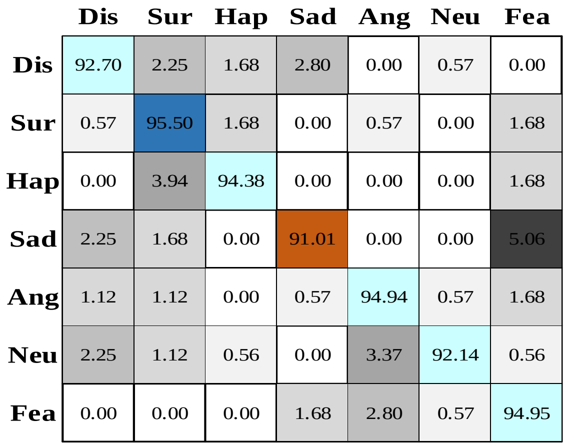
Method	JAFFE	CK+
AlexNet+Origin Image_Input	**83.35%**	**84.86%**
AlexNet+3 channel (Origin Image, G_x,G_y)_Input	**90.10%**	**91.36%**
AlexNet+3 channel (Origin Image, G_x,G_y)_Input+ Data Augmentation	**92.02%**	**93.12%**
Proposed metnod:[AlexNet+ 3 channel (Origin Image, G_x,G_y)_Input+ Data Augmentation+Dropout technique]	**92.41%**	**93.66%**

In this step, the accuracy recognition rates were obtained to be 80.65% and 81.12% for the JAFFE and CK+ datasets, respectively. In the last experiment, from layer 5 onwards were trainable, while the rest of the layers were kept frozen. In this step, the accuracy recognition rates were obtained at 83.35% and 84.86% for the JAFFE and CK+ datasets, respectively. Our experiments confirmed that when the datasets had a limited numbe ther of samples, it was better to keep more layers of the AlexNet DCNN network untrainable. This reduced complexity and improved the training process. In the next stage, we tested the AlexNet model (Conv_1 to Conv_4 were frozen and the other layers were trainable) on the proposed model; is the results obtained are given in Table 3. When the network inputs were original images, the accuracy recognition rates were obtained at 83.35% and 84.86% for the JAFFE and the CK+ datasets, respectively. In the next stage, the three channels of image (Origin Image, Image Gradient X, and Image Gradient Y) were extracted similar to the RGB image as the AlexNet DCNN input. This proposed input was compatible with the AlexNet-DCNN network, and improved the recognition rate to 90.10% and 91.36% for the JAFFE and the CK+ datasets, respectively. The use of data augmentation increased the recognition rate by 1.92% and 1.76% compared to the previous state for the JAFFE and CK+ datasets, respectively. The best recognition rate was obtained based on 3 channels (Origin image, Gx, Gy) _input, data augmentation, and dropout technique. Using the proposed method, the recognition rates were obtained to be 92.41% and 93.66% for the JAFFE and CK+ datasets, respectively. Figures 8 and 9 show the confusion matrixes based on the best-proposed model for the JAFFE and CK+ datasets. In the confusion matrix, while each row denotes the true emotion, the columns denote the recognized emotions. The diagonal of the matrix shows the recognition rate of each emotion. As it can be seen in Figure 8, the highest and the lowest recognition rates for the JAFFE dataset are related to the emotions of happiness and anger with 95.31% and 89.07%, respectively. As shown in Figure 9, the highest and the lowest recognition rates for the CK+ dataset are related to the emotions of surprise and sadness with the rates of 95.50% and 91.01%, respectively. Figures 10 and 11 show the training accuracy and validation accuracies on the JAFFE and CK+ datasets per 200 epochs, respectively. As shown in the figures, the input data is well-trained to high accuracy, and the validation data is converged to the input data. The validation accuracies were obtained to be 92.41% and 93.66% on the JAFFE and CK+ datasets, respectively. Figures 12 and 13 show the training loss and validation loss on the JAFFE and CK+ datasets per 200 epochs, respectively. The error rate was obtained to be about 0.6 for the JAFFE dataset and less than 0.5 for the CK+ dataset. Training on the input data moved towards a minimal error, and the validation loss almost converged to the training loss. According to Figures 10 to 13, it can be concluded that the proposed model performed well on the JAFFE and CK+ datasets, and acceptable results were obtained. Table 4 lists some of the works on facial expression recognition using the JAFFE and CK+ dataset so that we could compare our work fairly with them. In [39] and [2], the recognition rates on the JAFFE datasets were obtained to be 93.03% and 93.43%, respectively. Despite the recognition rate of 92.41% obtained in the proposed model which was slightly lower than those, we gained the best recognition rate on the CK+ dataset in comparison with the other studies. As it can be seen in Table 4, the proposed method is competitive with the state-of-the-art methods.



	Dis	Sur	Hap	Sad	Ang	Neu	Fea
Dis	92.18	0.00	3.13	1.56	3.13	0.00	0.00
Sur	0.00	90.62	3.13	0.00	1.56	0.00	4.69
Hap	0.00	3.13	95.31	0.00	1.56	0.00	0.00
Sad	4.69	1.56	0.00	92.18	0.00	0.00	1.56
Ang	0.00	7.81	0.00	1.56	89.07	0.00	1.56
Neu	3.12	0.00	0.00	0.00	0.00	93.76	3.12
Fea	1.56	1.56	0.00	1.56	1.56	0.00	93.76

Figure 8. Confusion matrices on the JAFFE dataset; average accuracy = 92.41%.



	Dis	Sur	Hap	Sad	Ang	Neu	Fea
Dis	92.70	2.25	1.68	2.80	0.00	0.57	0.00
Sur	0.57	95.50	1.68	0.00	0.57	0.00	1.68
Hap	0.00	3.94	94.38	0.00	0.00	0.00	1.68
Sad	2.25	1.68	0.00	91.01	0.00	0.00	5.06
Ang	1.12	1.12	0.00	0.57	94.94	0.57	1.68
Neu	2.25	1.12	0.56	0.00	3.37	92.14	0.56
Fea	0.00	0.00	0.00	1.68	2.80	0.57	94.95

Figure 9. Confusion matrices on the CK+ dataset; average accuracy = 93.66%.

Table 4. Comparison of the average recognition accuracy (%) with the state-of-the-art works.

Research work	Average Recognition Accuracy (%)		Method
	JAFFE	**CK+**	
Huang *et al*.[36]	85.15	85.07	SRC+ Raw pixel
Ying *et al*.[37]	89.7	86.65	SRC+ Raw LBP
Du *et al*.[38]	89.45	90.72	M-CRT
Sumaidaee *et al*.[39]	93.03	90.62	Gabor
Ding *et al*.[2]	93.43	-	Histogram + TFP
Xie *et al*. [40]	-	92.06	FRR-CNN
Proposed Method	92.41	93.66	AlexNet DCNN

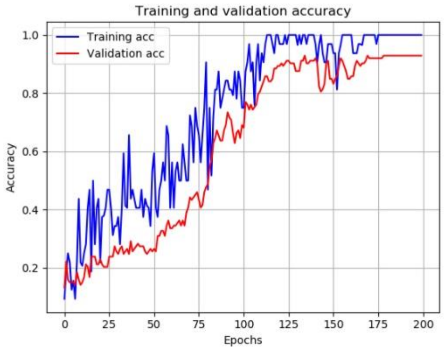


Figure 10. Training accuracy and validation accuracy on JAFFE dataset per epoch.

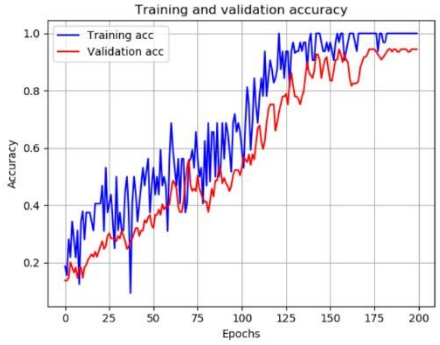


Figure 11. Training accuracy and validation accuracy on CK+ dataset per epoch.

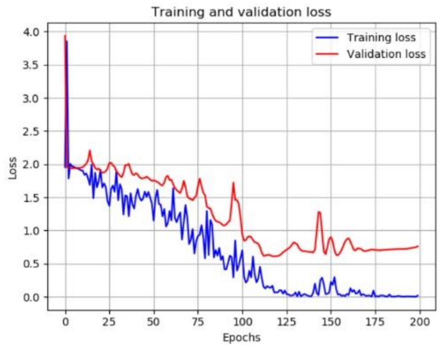


Figure 12. Training loss and validation loss on JAFFE dataset per epoch.

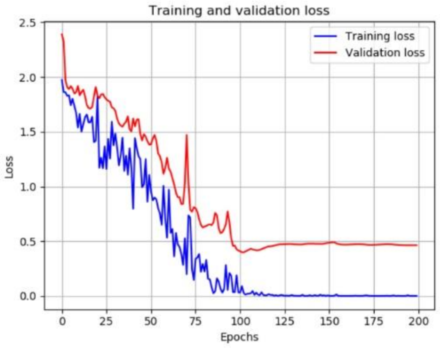


Figure 13. Training loss and validation loss on CK+ dataset per epoch.

5. Conclusions and Future Works

In this paper, we proposed a new facial expression recognition (FER) system using the pre-trained

AlexNet DCNN. To this end, the dataset samples should be adapted to the network input.

Here, we used the original image in addition to its horizontal and vertical gradients as the red, green, and, blue channels of an RGB image for the DCNN input. The following conclusions can be drawn from our experiments.

First, since face wrinkles convey important emotional cues, the gradients of the original image containing the image details are not only useful in providing a fit-size input for AlexNet but also enrich the network input in terms of the important emotional features of the image. In other words, considering the network input as limited space for providing the essential information for the system; in a FER system, the optimum network input contains as much emotional information as possible. Our experiments showed that the gradients of an image could significantly increase the accuracy of FER when they add to the original image.

Secondly, a large number of training variables could lead to over-fitting. Thus, in the pre-trained networks, the number of trainable layers should be chosen according to the size of the dataset. Furthermore, some methods such as data augmentation and dropout technique could improve the system performance by providing the supporting data and prevent it from over-fitting.

As a future work, since our experiments show that the most effective part of the system is input preparation providing invaluable data for the network, we plan to investigate transformations that highlight the emotional cues of the image, while removing the irrelevant features from the input image. Ideally, an optimal network input contains as much as possible emotional information, while it is free from other information.

References

[1] A. Majumder, L. Behera, and V.K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE transactions on cybernetics.*, vol. 48, pp. 103-114, Jan 2016.

[2] Y. Din, Q. Zhao, B. Li, and X. Yuan, "Facial expression recognition from image sequence based on LBP and Taylor expansion," *IEEE Access.*, vol. 5, pp. 19409-19419, August 2017.

[3] JY. Jung, SW. Kim, CH. Yoo, WJ. Park, and S.J. Ko, "LBP-ferns-based feature extraction for robust facial recognition," *IEEE Transactions on Consumer Electronics*., vol. 62, pp. 446-453, November 2016.

[4] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cGAN Based Facial Expression Recognition for Human-Robot Interaction," *IEEE Access*., vol. 7, pp. 9848-9859, January 2019.

[5] M. Z. Uddin, M.M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*., vol. 5, pp. 4525-4536, March 2017.

[6] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on pattern analysis and machine intelligence*., vol. 27, pp. 699-714, March 2005.

[7] A. Panning, A.K. Al-Hamadi, R. Niese, and B. Michaelis, "Facial expression recognition based on haar-like feature detection," *Pattern Recognition and Image Analysis*., vol. 18, pp. 447-4452, Sep 2008.

[8] C. Shan, S. Gong and P.W. McOwan. "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*"., vol. 27, pp. 803-816, Dec 2009.

[9] W. Liu, Y. Wang, and S. Li, "LBP feature extraction for facial expression recognition," *Journal of information & computional science*., vol. 8, pp.412-421, March 2011.

[10] L. Wang and k. Wang, R. Li, "Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition," *IET Computer Vision*., vol. 9, pp. 655-652, Oct 2015.

[11] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*., vol. 49, pp. 4516-4527, May 2011.

[12] B. Yang, J. Cao, R. Ni, and Y. Zhang. "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*., vol. 6, pp. 4630-40, Dec 2017.

[13] B.F. Wu and C.H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*., vol. 6, Feb 2018.

[14] J.H. Kim, B.G. Kim, P.P. Roy, and D.M. Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*., vol. 7 Jan 2019.

[15] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*., vol. 21, pp. 211-220, June 2018.

[16] Z Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing.*, vol. 317, pp. 50-57, November 2018.

[17] M. Garcia and S. Ramirez, "Deep Neural Network Architecture: Application for Facial Expression Recognition," *IEEE Latin America Transactions*., vol. 18, pp. 1311-1319, May 2020.

[18] Y. Yan, Y. Huang, S. Chen, C. Shen, , "Joint Deep Learning of Facial Expression Synthesis and Recognition," *Computer Vision*., Feb 2020

[19] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*., pp. 1097-105, Sep 2012.

[20] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," *Proceedings Third IEEE int conference on automatic face and gesture recognition: IEEE*., pp. 200-5, 1998.

[21] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops: IEEE*., pp. 94-101, 2010.

[22] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*., vol. 19, pp. 2816-31, June 2017.

[23] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Processing*., vol. 12, pp. 713-21, Feb 2018.

[24] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE Journal on Selected Areas in Communications*., vol. 37, pp. 1389-401, March 2019.

[25] U. Côté-Allard, C.L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*., vol. 27, pp. 760-71, January 2019.

[26] C. Deng, Y. Xue, X. Liu "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*., vol. 57, pp. 1741-54, November 2018.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv*., September 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*., 2016. p. 770-8.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*., 2016. p. 2818-26.

[30] G. Huang, Z, Liu, L Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks,". *Proceedings of the IEEE conference on computer vision and pattern recognition*., 2017. p. 4700-8.

[31] N. Ketkar, "Deep Learning with Python," Springer, 2017.

[32] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th international conference on machine learning (ICML-10).*, 2010. p. 807-14.

[33] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat No PR00580): IEEE*., 2000. p. 46-53.

[34] F. Chollet, "Deep Learning with Python," Springer, 2018.

[35] A. Harimi, A. Shahzadi, A.R. Ahmadifard, and K. Yaghmaie, "Classification of emotional speech using spectral pattrn features," *Journal of AI and data mining*., vol. 2, pp. 53-61, 2014.

[36] M.W. Huang, Z.W. Wang, and Z.L. Ying, "A new method for facial expression recognition based on sparse representation plus LBP," *2010 3rd International Congress on Image and Signal Processing: IEEE*., 2010. p. 1750-4.

[37] Z.L. Ying, Z.W. Wang, and M.W. Huang, "Facial expression recognition based on fusion of sparse representation," *International Conference on Intelligent Computing: Springer*., 2010. p. 457-64.

[38] L. Du and H. Hu, "Modified classification and regression tree for facial expression recognition with using difference expression images," *Electronics Letters*., vol. 53, pp.590-592, April 2017.

[39] S. Al-Sumaidaee, S. Dlay, "Facial expression recognition using local Gabor gradient code-horizontal diagonal descriptor," *IET International Conference on Intelligent Signal Processing.*, Novomber 2015.

[40] S. Xie and H Hu, "Facial expression recognition with FRR-CNN," *Electronics Letters.*, February 2017.

تشخیص احساس چهره با استفاده از گرادیان تصویر و شبکه عصبی کانولوشنی عمیق

محمدرضا فلاح زاده [۱]، فرداد فرخی [۱]، علی حریمی [۲،*] و رضا صباغی ندوشن [۱]

[۱] دانشکده فنی و مهندسی ،دانشگاه آزاد اسلامی، واحد تهران مرکزی، تهران، ایران.

[۲] دانشکده فنی و مهندسی ،دانشگاه آزاد اسلامی، شاهرود، ایران.

ارسال ۲۰۲۰/۰۷/۱۵؛ بازنگری ۲۰۲۰/۱۲/۱۳؛ پذیرش ۲۰۲۱/۰۱/۰۳

چکیده:

تشخیص احساس از روی چهره انسان (FER) یکی از اساسی‌ترین روش‌های تعامل انسان با ماشین می‌باشد. از اینرو در سال‌های اخیر توجه بسیاری از محققان را به خود جلب کرده است. در این مقاله به منظور تشخیص احساس یک روش جدید بر اساس شبکه عصبی کانولوشنی عمیق (DCNN) ارائه شده است. با آگاهی از توانایی DCNN ها در یادگیری ویژگی‌ها و کلاس‌بندی تصاویر، هدف از این تحقیق طراحی یک ورودی سازگار با شبکه AlexNet-DCNN از روی تصاویر چهره می‌باشد. روش پیشنهادی شامل ۴ مرحله است. در گام اول سه کانال تصویر شامل تصویر اصلی، تغییرات چهره در راستای محورx (گرادیان) و تغیرات چهره در راستای محور y به عنوان ۳ کانال R,G,B شبیه یک تصویر RGB به عنوان ورودی سازگار با شبکه AlexNet-DCNN استخراج می‌شود. در گام دوم جهت افزایش مقدار داده‌ها بزرگنمایی تصاویر همراه با شیفت به چپ و راست و چرخش تصاویر استخراج می‌شنود. سپس از یک شبکه AlexNet-DCNN که بر روی پایگاه داده Image-Net آموزش دیده شده‌اند جهت یادگیری ویژگی‌ها استفاده شده است. در گام آخر تطبیق و تنظیمات دقیق شبکه AlexNet بر روی مدل پیشنهادی انجام گرفته است. میانگین دقت نرخ تشخیص بر روی پایگاه های داده JAFFE و +CK به ترتیب ۹۲/۴٪ و۹۳/۶۶٪ به دست آمده است. نتایج آزمایش‌ها بر روی دو پایگاه داده رایج و عمومی احساس، عملکرد امیدوار کننده روش پیشنهادی را نشان می‌دهد که می‌تواند در بالا بردن دقت سیستم‌های FER موثر باشد.

کلمات کلیدی: تشخیص حالت چهره، شبکه عصبی کانولوشنی عمیق، گرادیان تصاویر، شبکه AlexNet-DCNN.