IJICTR
International Journal of Information & Communication Technology Research

Volume 15- Number 4 – 2023 (25 -31)

Facial Expression Recognition Based on Separable Convolution Network and Attention Mechanism

Amir Khani Yengikand *
Department of Computer Engineering,
University of Zanjan,
Zanjan, Iran
amirkhani@znu.ac.ir

Mostafa Farrokhi Afsharyan
Department of Electrical Engineering,
Shahid Sattari Aeronautical University
Tehran, Iran
m.farrkkhi1994@gmail.com

Payam Nejati
Department of Electrical Engineering,
Malek Ashtar University of Technology,
Tehran, Iran
payamnejati@gmail.com

Received: 10 October 2022 – Revised: 10 December 2022 - Accepted: 5 February 2023

Abstract—**Facial expression recognition using deep learning methods has been one of the active research fields in the last decade. However, most of the previous works have focused on the implementation of the model in the laboratory environment, and few researchers have addressed the real-world challenges of facial expression recognition systems. One of the challenges of implementing the face recognition system in the real environment (e.g. webcam or robot) is to create a balance between accuracy and speed of model recognition. Because, increasing the complexity of the neural network model leads to an increase in the accuracy of the model, but due to the increase in the size of the model, the recognition speed of the model decreases. Therefore, in this paper, we propose a model to recognize the seven main emotions (Happiness, sadness, anger, surprise, fear, disgust and natural), which can create a balance between accuracy and recognition speed. Specifically, the proposed model has three main components. First, in the feature extraction component, the features of the input images are extracted using a combination of normal and separable convolutional networks. Second, in the feature integration component, the extracted features are integrated using the attention mechanism. Finally, the merged features are used as the input of the multi-layer perceptron neural network to recognize the input facial expression. Our proposed approach has been evaluated using three public datasets and images received via webcam**

Keywords: Facial expression recognition, Attention mechanism, Separable convolutional

Article type: Research Article

Publisher: ICT Research Institute

* Corresponding Author

[DOI: 10.61186/itrc.15.4.25]

I. INTRODUCTION

Analyzing and recognizing facial expressions using artificial intelligence methods is an interesting combination of psychology and computer science [1]. Facial expressions are one of the most natural and universal signals to convey feelings from one person to another. Facial recognition systems are used in various applications such as human-robot interaction, human behavior understanding, video game testing, etc [2].

With the rapid development of deep learning and machine learning, researchers are using these methods in various problems such as spam email detection, fraud detection, recommender systems, and social network analysis [3, 4]. In addition, deep learning techniques have been widely used in pattern recognition problems, especially facial expression recognition. The implementation of facial expressions recognition systems is very useful and beneficial to increase human-computer interaction. By applying deep learning techniques such as convolutional neural networks (CNN), recurrent neural network (RNN) in facial expression recognition, researchers have made many advances in this field [1]. However, the performance of these systems in the real world is not very acceptable, and there are challenges in the real world that should be considered in the implementation of the model [2]. One of the existing challenges is creating a balance between the accuracy and speed of model recognition in the real environment [5]. When we increase the volume or complexity of the neural network model, the accuracy of the model increases, but the recognition speed of this model in the real environment through the webcam decreases due to the large number of network weights. Increasing network weights leads to two problems. 1) First, neural network training takes a lot of time and requires powerful hardware to be used in the real environment. 2) Second, it increases the facial expression detection time. For example, using a model with high complexity on a Raspberry Pi computer causes a person's facial expression to be recognized with a long delay. Therefore, it is necessary for the proposed model to provide a balance between accuracy and detection speed in the real environment. Most researchers have used normal convolutional neural networks to implement facial emotion recognition systems [6, 7]. These networks increase calculations due to the number of multiplications between the filter and the input image. As a result, the detection speed increases. To solve this challenge and obtain acceptable accuracy, different from these works, we combined the normal convolutional and separable convolutional networks together with the residual module [8]. The main contributions of this paper are as follows:

- We proposed a model for facial expression recognition in real environment, which extracts facial features through the combination of normal and separable convolutional network with the residual module.
- To combine the extracted features, we proposed a feature composition component, which combines the features extracted from the feature extraction component by the attention mechanism to increase the accuracy of the model.
- Finally, we use multi-layer perceptron neural network to detect the emotion of the input face. In this way, the features obtained in the feature combination component are fed to the perceptron neural network. At the end of this network, the softmax function is used to recognize the input face class.

II. RELATED WORKS

With the extensive growth of deep learning in various fields, CNN have been widely used by researchers in the face recognition problem. These networks have achieved good results in laboratory environments, but they have shown weaknesses in the real environment [1]. One of the important stages of the facial expression recognition system is the feature extraction component. In the last decade, researchers use deep learning methods instead of manually extracting features from data. Most researchers use CNN to solve the problem of facial expression recognition. In [6], CNN have been used for feature extraction, and finally, Support Vector Machines (SVM) has been used to classify the input images instead of using the softmax function. Yao et al. [9] proposed a new architecture called Holonet, in this architecture they combined the activation function of CRELU with the Resent network to increase the depth of the CNN without reducing the efficiency. In [10], CNN architecture is employed to obtain image features, then Restricted Boltzmann Machine (RBM) network is used to obtain more complete and higher-level features, and finally, SVM is employed to classify facial expressions. Rangulov et al. in [11] combined CNN with RNN. In this way, first, the image features are extracted by CNN and then to obtain more accurate features, the features extracted by CNN are fed to RNN. Mollahosseini et al. [5] extended the inception network by adding some regular convolutional layers to its end. Also, some researchers used pre-trained networks [7, 12]. For instance, in [13], the VGG-face network that was implemented to solve the face recognition problem is used to develop the facial emotion recognition problem. In [14], several CNNs with different structures were trained for facial expression recognition, and finally the voting method was used to determine the class of the input image.

TABLE I. SUMMARY OF THE RELATED WORKS ALONG WITH THEIR ADVANTAGES/DISADVANTAGES.

model	advantage	disadvantage
DLSVM [6]	Replacing the softmax function	complexity and costly

[DOI: 10.61186/itrc.15.4.25]

	with a support vector machine for classification, which led to improved results.	
HoloNet [9]	Use the Concatenated Rectified Linear Unit (CReLU) instead of ReLU function for reduce redundant filters and enhance the non-saturated nonlinearity in the lower convolutional layers, combine the inception and residual structure to broaden network width and introduce multi-scale feature extraction property	Decreased facial emotion detection speed in the real environment due to the use of a different network with more layers
AUDN [10]	use the computational representation MAP(Micro-Action-Patter) to obtain local appearance changes caused by facial expression.	Increasing the learning process due to the use of many layers, ineffective in the real environment
RNN-CNN [11]	Feeding the RNN network with features extracted from the CNN network for exploiting the temporal dynamics of video	overfitting of the model due to the training of two networks at the same time
[5]	simple and efficient model and combining seven datasets for network training	Need more data to train the model due to the simplicity of the model
Multi-model [13]	Proposing a multi-model using audio and image data for video-based emotion recognition in the wild	Relatively costly and complex
[14]	Multiple CNNs for feature extraction and use voting method to determine the input image class with the aim of increasing accuracy	Reduced detection speed due to the use of voting method

III. PRELIMINARIES

Before explaining the proposed approach, in this section, the normal and separable convolutional neural network is discussed and the difference between these two methods is clearly explained.

A. Convolutional neural networks

CNN is a type of feedforward neural network. This network has been successfully applied to many practical problems such as image classification, speech recognition, self-driving cars and natural language processing (NLP) [15]. A CNN model has three main parts, convolution layer, pooling layer and fully connected layers [1]. The convolutional layer is used to extract the features from the input, the pooling layer is used to reduce the dimensions of the extracted features and select the important features with the aim of reducing the processing time and finally, fully connected layers are used for input classification. The convolution layer is the core of the CNN model, which performs most of the CNN calculations. In general, the convolution operation in the CNN is divided into two main methods. In the following, we will examine these two types of convolution operations with an example.

1) Normal Convolutional Layer

Suppose you have an image with dimensions $12 * 12 * 3$, which you want to apply 256 filters with dimensions $5 * 5 * 3$. As you can see in Figure 1, after applying convolution, the resulting output will be $8 * 8 * 256$. Equation 1 shows the calculation of the output after applying the filter on the input.

$$w_o = \frac{W - K + 2P}{s} + 1 \qquad (1)$$

where W, K denotes the dimensions of the image and the dimensions of the filter, respectively. p, s denotes padding and stride length, respectively. It should be noted that the number of output dimensions will be equal to the number of applied filters.

Let us calculate the number of multiplications that are performed in normal convolution. As mentioned above, we have 256 filters of dimensions $5 * 5 * 3$, which must be applied to the input image $8 * 8$ times, to produce an output of dimensions $8 * 8 * 256$. Therefore, the number of multiplications in this type of convolution will be equal to $256 * 3 * 5 * 5 * 8 * 8 = 1228800$.

2) Separable Convolutions Layer

The main purpose of separable convolution layer is to reduce the number of network parameters compared to the normal convolutions. The basic idea of separable convolution is to divide a convolution filter into two smaller filters. In this method, two types of convolutions are applied to the input image [16]: Depthwise Convolution and Pointwise Convolution. For a more detailed understanding, consider the example of the previous section, which we want to apply 256 filters of dimensions $5 * 5 * 3$ to the same image. As you can see in Figure 2, First, three filters with dimensions of $5 * 5 * 1$ are applied on the image. After applying this filter, the output dimensions will be $8 * 8 * 3$. Then, in the second step, in order to make the number of output dimensions equal to 256, 256 filters with dimensions $1 * 1 * 3$ are applied to the output of the previous step. Figure 3 shows this process. However, the number of multiplications in the

[DOI: 10.61186/itrc.15.4.25]

first stage (i.e. depthwise convolution) is equal to $3 * 5 * 5 * 8 * 8 = 4800$, and in the second stage (i.e. pointwise convolution), it is equal to $256 * 1 * 1 * 3 * 8 * 8 = 49152$. The total number of multiplications becomes 53,952, which is less number of multiplications compared to normal convolution. For this reason, we use this method in the proposed model to increase the recognition speed of the model.

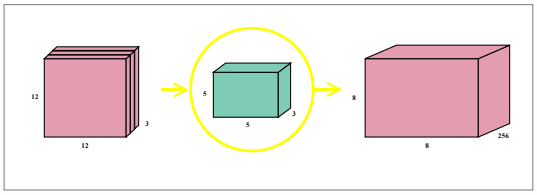


Figure 1. Normal convolution operation

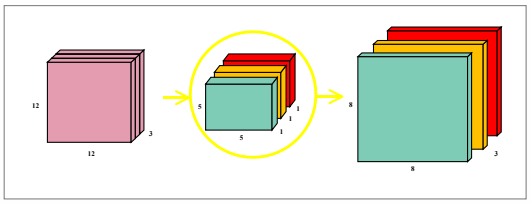


Figure 2. Separable convolution operation (Depthwise)

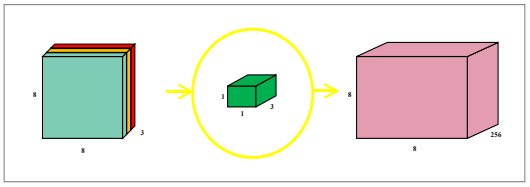


Figure 3. Separable convolution operation (Pointwise)

IV. THE PROPOSED MODEL

In this section of the paper, we describe the details of the proposed model framework for facial expression recognition using deep learning model. Figure 4 shows our proposed approach. As it is clear, the proposed model consists of three main components: the feature extraction component, the feature composition component and the classification component. The feature extraction component receives images as input and then extracts features using a separable and normal convolutional neural network for use in the classification component. The feature composition component is included after the feature extraction component with the aim of integrating the obtained features. In this component, additive attention and normal attention mechanism are used to select more accurate features. Finally, in the classification component, the features integrated in the previous step are placed as the input of the perceptron neural network for facial emotion recognition. In the proposed model, we used a depthwise convolutional model to extract features from images to recognize the facial expression type instead of using high-volume models such as normal convolutional networks. As we argued in the preliminaries section, the depthwise convolutional network does not add many operations to the model to extract features, and this causes the model to have a faster detection speed in the real environment, because less calculations are performed in the detection mode. Also, in order not to decrease the accuracy of the model, we used normal convolution in some layers, which keeps the accuracy of the model. In the following, we explain the proposed approach in more detail.

A. Feature Extraction Component

Facial expression recognition system needs to extract sufficient features from images. In other words, this component plays a vital role in facial expression modeling [1]. In this section, we describe feature extraction through separable and normal convolutional networks along with the residual module. This network has 6 blocks, three blocks include spatially separable convolution and batch normalization. Also, three other blocks, including normal convolution, batch normalization and pooling layer. After extracting the feature by this network, the obtained features are placed directly as the input of the attention mechanisms as well as the input of the other two layers that are defined in the image. As it is clear in Figure 4, we used the idea of the residual module of the resent network. This idea actually creates a shortcut and connects the output of one layer to several layers ahead, which prevents the gradient vanishing problem.

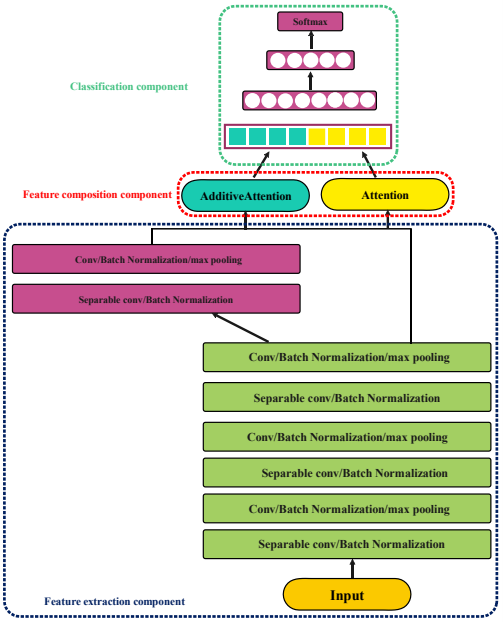


Figure 4. The proposed model

B. Feature Composition Component

One of the common strategies to combine features extracted from different networks is to use a fusion layer [3, 17]. In this paper, different from other works, we use two types of attention mechanisms to merge the features extracted by the feature extraction component. As is clear in Figure 4, the feature extraction component has two outputs. These outputs are used as inputs for the normal attention and the additive attention mechanism. The equation of the attention mechanism is the following.

[DOI: 10.61186/itrc.15.4.25]

Additive attention: This method is also known as Behdanav mechanism, this method uses the hidden layer to calculate the attention level score, this mechanism is formulated as follows.

$$f_{att}(h_i, s_j) = v_a^T \tanh(W_a[h_i; s_j]) \quad (2)$$

w_a and v_a represent the learning parameters of the attention layer. s_j and h_i are the input features of the attention layer.

Normal attention: This method is similar to the previous method, with the difference that instead of the sum of the features, it uses the multiplication of the features.

C. Classification Component

Softmax function is widely used as a classifier in the last layer of most facial expression recognition models [6]. Different from these works, we used a multilayer perceptron network before applying the softmax function. In this way, the integrated features in the feature composition component are fed to this network. The reason for using the perceptron network is to give weight to each of the features extracted by the CNN, which makes better features to be extracted. Better feature extraction will lead to better classification.

V. Result and Experiments

In this section, we perform facial expression classification on real data sets. We also present the results of experiments to evaluate the performance of the proposed model. We implemented our own model with Python and its libraries such as keras, pandas, numpy, openCV. We used the colab environment to train our model. The accuracy metric is used to evaluate the model. The Accuracy score is calculated by dividing the number of correct predictions by the total prediction number. This metric is formulated as follows.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{All\ Sample} \quad (3)$$

A. Datasets

Generally, two types of datasets can be used for facial expression recognition: 1) The datasets collected in the laboratory environment. 2) The datasets collected from the real environment. With the aim of obtaining high accuracy, we used three public datasets namely Fear2013[1], CK+[2], Jaffe[3] and trained the proposed model with these three datasets.

Fear2013

FAIR 2013 is a large-scale and unrestricted dataset which has been used in many papers and ICML competitions. This dataset is automatically collected by the Google image search API. The number of images in this dataset is 35,887, which resized to 48*48 pixels after rejecting incorrectly labeled frames and adjusting the cropped region. This dataset is not balanced, as it contains images from 7 different modes, with distributions of Angry (4,953), Disgust (547), Fear (5,121), Sad (6,077), Happy (8,989), Surprise (4,002), and Neutral (6,198).

CK+

The Cohn–Kanade (known as CK+) database is the most extensive controlled laboratory database for emotion recognition and action unit. It includes 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age with a diversity of genders and heritage. In this dataset, each video shows a face change from a neutral expression to a peak expression, which was recorded with 640x490 or 640x480 pixels. Out of these videos, 1000 are labelled with one of seven expression classes.

Jaffe

The Japanese Female Facial Expression (JAFFE) is a relatively small dataset containing 219 images from 10 Japanese females. Each woman has 4-3 images with each of seven facial expressions.

B. Parameter setting

Python programming language and Keras library are used to conduct the experiments. The number of epochs for learning the models is set to 20 for all datasets. The batch size is adjusted to 32, 64 and 16 for Fer2013, CK+ and Jaffe, respectively. In order to control the overfitting problem, the limited area of dropout ratio is selected from the range of [0.1, 0.2, 0.3] using a greedy search strategy. We used the *Relu* activation function for each layer. Also, we used Adam optimizer [18] and learning rate of 0.001 to optimize the model.

C. Accuracy

We compared the proposed model with two types of deep learning models. Table 1 shows the results for three datasets. As you can see, our model is more accurate than the other two models. Figure 5 shows the confusion matrix generated for the three datasets. As you can see, in all three confusion matrices, the diagonal line is thicker, which shows that the classification of the test data is done well. In Figure 5, the confusion matrix shows several misclassifications, for example the sad state is predicted as fear or the angry state is classified instead of disgust, this is due to the similarity of these classes to each other. The results

[1] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

[2] https://www.consortium.ri.cmu.edu/ckagree/

[3] http://www.kasrl.org/jaffe.html

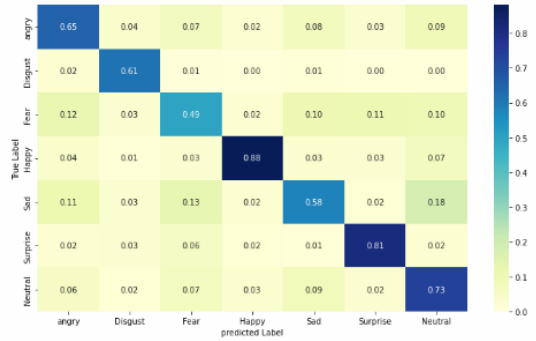
[DOI: 10.61186/itrc.15.4.25]

of CK+ and Jeff dataset are better than Fair2013. The reason is that we first trained the model with the Fear2013 dataset and then saved the weight of the model and trained it again with two datasets (i.e. CK+ and Jaffe) and finally evaluated the model with the test data of these two datasets. For this reason, the results of these two datasets are better than Fear2013.

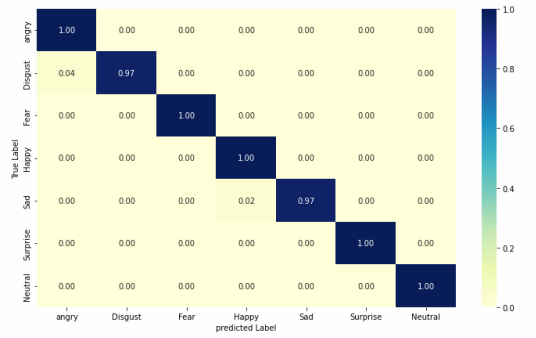
TABLE II. COMPARISON OF RESULTS WITH OTHER METHODS

Datasets	Fer2013	CK+	Jaffe
Liu et al.[19]	59.21	91.74	70.64
Samsani et al. [20]	60.86	92.33	71.28
Dapogny et al. [21]	61.13	92.97	73.00
Happy et al.[22]	62.00	94.09	74.21
Jung et al.[23]	62.11	96.64	75.27
Our model	65.32	97.01	79.19

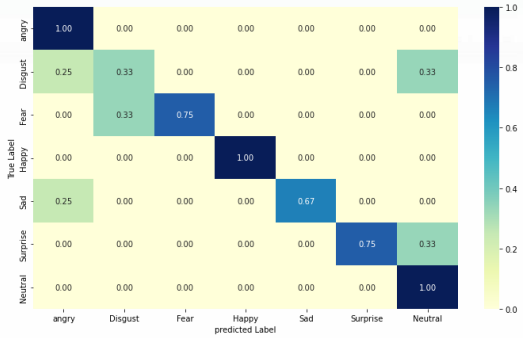
Also, we evaluated the proposed model through webcam. In Figure 6, you can see face recognition and then facial expression recognition. For this purpose, we first used the OpenCV library for face recognition, and then used the proposed model for facial emotion recognition. Our model recognized facial expressions with acceptable speed and accuracy through webcam. Therefore, it can be concluded that our proposed approach was able to balance the accuracy and speed of detection and achieve better results.



(a) Confusion matrix of Fear2013



(b) Confusion matrix of CK+



(c) Confusion matrix of Jaffe

Figure 5. The confusion matrices of proposed model

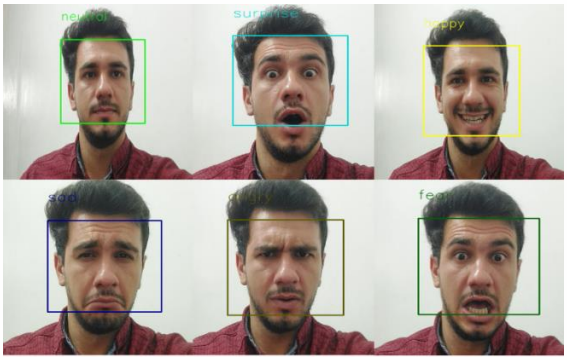


Figure 6. The results of facial emotion recognition through webcam

VI. CONCLUSION

In this paper, we proposed a deep learning model for facial expression recognition. That the image features are extracted by convolutional networks with normal and separable layers along with the idea of the residual module of the Resnet network. Finally, by combining the features extracted by the attention mechanism, and feeding it to the multilayer perceptron neural network and using the softmax layer, the class of the input image is determined. The aim of this paper was to present a model that can achieve acceptable accuracy with a lower detection speed without increasing the network weights too much. One of the limitations of our work is that we used western and eastern datasets to train the model. In the future, we can use Iranian dataset to increase the accuracy of the model. Also, for future works, we can research about emotion recognition on data with more number of emotions and combining it with emotional speech combination. There are some approaches like RNN, RBM that can be used in the future to improve the results.

REFERENCES

[1] Li, S. and W. Deng, Deep facial expression recognition: A survey. IEEE transactions on affective computing, 2020.

[2] Pramerdorfer, C. and M. Kampel, Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903, 2016.

[3] Yengikand, A.K., et al. Deep representation learning using multilayer perceptron and stacked autoencoder for recommendation systems. in 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2021. IEEE.

[DOI: 10.61186/itrc.15.4.25]

[4] Ahmadian, S., et al., A social recommender system based on reliable implicit relationships. Knowledge-Based Systems, 2020. 192: p. 105371.

[5] Mollahosseini, A., 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). 2016. IEEE.

[6] Tang, Y., Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239, 2013.

[7] Xu, M., et al. Facial expression recognition based on transfer learning from deep convolutional networks. in 2015 11th International Conference on Natural Computation (ICNC). 2015. IEEE.

[8] Targ, S., D. Almeida, and K. Lyman, Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029, 2016.

[9] Yao, A., et al. HoloNet: towards robust emotion recognition in the wild. in Proceedings of the 18th ACM international conference on multimodal interaction. 2016.

[10] Liu, M., et al., Au-inspired deep networks for facial expression feature learning. Neurocomputing, 2015. 159: p. 126-136.

[11] Rangulov, D. and M. Fahim. Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network. in 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS). 2020. IEEE.

[12] Ng, H.-W., et al. Deep learning for emotion recognition on small datasets using transfer learning. in Proceedings of the 2015 ACM on international conference on multimodal interaction. 2015.

[13] Kaya, H., F. Gürpınar, and A.A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing, 2017. 65: p. 66-75.

[14] Khorrami, P., T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? in Proceedings of the IEEE international conference on computer vision workshops. 2015.

[15] Kamilaris, A. and F.X. Prenafeta-Boldú, Deep learning in agriculture: A survey. Computers and electronics in agriculture, 2018. 147: p. 70-90.

[16] Chollet, F. Xception: Deep learning with depthwise separable convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[17] Yengikand, A.K., M. Meghdadi, and S. Ahmadian, DHSIRS: a novel deep hybrid side information-based recommender system. Multimedia Tools and Applications, 2023: p. 1-27.

[18] Kingma, D.P. and J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[19] Liu, M., et al. Au-aware deep networks for facial expression recognition. in 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). 2013. IEEE.

[20] Samsani, S. and V.A. Gottala. A real-time automatic human facial expression recognition system using deep neural networks. in Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018. 2020. Springer.

[21] Dapogny, A., K. Bailly, and S. Dubuisson, Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. International Journal of Computer Vision, 2018. 126: p. 255-271.

[22] Happy, S. and A. Routray, Automatic facial expression recognition using features of salient facial patches. IEEE transactions on Affective Computing, 2014. 6(1): p. 1-12.

[23] Jung, H., et al. Joint fine-tuning in deep neural networks for facial expression recognition. in Proceedings of the IEEE international conference on computer vision. 2015.

Amir Khani Yengikand received his B.Sc. degree in Computer Engineering from Amirkabir University of Technology, Arak, Iran, in 2018 and M.Sc degree in Computer Engineering from Zanjan University, Zanjan, Iran, in 2021. His research interests includes Data Science, Deep Learning, Recommender Systems, Computer Vision, Natural Language Processing and Software Security. Also, he is interested in solving real-world Machine Learning problems.

Mostafa Farokhi Afsharyan received his B.Sc. and M.Sc. degrees in Electrical Engineering from Shahid Sattari Aeronautical University, Tehran, Iran, in 2015 and 2019 respectively. His research interests include Telecommunications, Cognitive Science, Signal and Image Processing and Machine Learning.

Payam Nejati received his B.Sc. degree in Electrical Engineering, Electronic, from Malek Ashtar University of Technology, Tehran, Iran in 2014. His currently a student of Electrical Engineering, Telecommunications at Malek Ashtar University. His research interests include Electromagnetic Theory, Signal and Image Processing, Radar Applications and Deep Learning.

[DOI: 10.61186/itrc.15.4.25]