ORIGINAL ARTICLE

Brain-Inspired Deep Networks for Facial Expression Recognition

Nafiseh Zeinali [1,*], Karim Faez [2], Sahar Seifzadeh [3,4]

[1] *Department of Electrical Computer and Biomedical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

[2] *Electrical Engineering Department, Amirkabir University of Technology Tehran, Iran*

[3] *Division of Cognitive Neuroscience, University of Tabriz, Tabriz, Iran*

[4] *Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

*Corresponding Author: Nafiseh Zeinali
Email: n.zeinali@gmail.com

Received: 15 August 2020 / Accepted: 21 September 2020

Abstract

Purpose: One of the essential problems in deep-learning face recognition research is the use of self-made and less counted data sets, which forces the researcher to work on duplicate and provided data sets. In this research, we try to resolve this problem and get to high accuracy.

Materials and Methods: In the current study, the goal is to identify individual facial expressions in the image or sequence of images that include identifying ten facial expressions. Considering the increasing use of deep learning in recent years, in this study, using the convolution networks and, most importantly, using the concept of transfer learning, led us to use pre-trained networks to train our networks.

Results: One way to improve accuracy in working with less counted data and deep-learning is to use pre-trained using pre-trained networks. Due to the small number of data sets, we used the techniques for data augmentation and eventually tripled the data size. These techniques include: rotating 10 degrees to the left and right and eventually turning to elastic transmation. We also applied deep Res-Net's network to public data sets existing for face expression by data augmentation.

Conclusion: We saw a seven percent increase in accuracy compared to the highest accuracy in previous work on the considering dataset.

Keywords: Deep Learning; Small Database; Face Expression; Res-Net Network; Self-Made Dataset.

DOI: https://doi.org/10.18502/fbt.v7i3.4619

1. Introduction

Faces are one of the most common non-verbal channels in which human transmits his inner emotional states through his face. Detecting face expression is highly applicable in various fields, such as cognitive neuroscience, behavioral sciences, and human-computer interfaces.

Rao *et al.* [1] use auto-associative neural networks to identify five facial expressions. These include anger, fear, happiness, and sadness. In their proposed system, at first, three areas, the eyes, and mouth are detected on the image. Then, to extract the facial features, a rectangular net is placed on each of these three areas. The net on the eyes consists of four rows and five columns, and the net on the mouth consists of five rows and seven columns. Then, to extract facial features, they applied the erosive morphology on the eyes area and the extension morphology on the mouth. Thus, for the three regions of the left eye, the right eye and the mouth, they obtained three vectors in size of 20, 20, and 35, respectively. For each area, they designed a face recognition system consisting of five auto-associative neural networks each of which comprising of five-face expressions. These networks consist of five layers, and their inputs are the vector properties corresponding to that region. Then, using the output of each network with its corresponding input, the normalized squared error equals with the following Equation 1:

$$e = \frac{||y - o||2}{||y||2} \quad (1)$$

Where *y* is the property vector of each network, and *o* is the output of that network. Then the degree of assurance of these five networks is obtained by the formula of c = exp (-e), and finally, the average of these assurance levels is calculated for the three models to obtain the ultimate degree of assurance of each facial expression. The face expressions with the highest reliability are the face expression of the input image. Nevertheless, their proposed system is unable to distinguish the two expressions of disgust and surprise. Most of the works conducted in the field of face recognition in sequence images use geometric features. After dividing the face into nine parts, Khanam *et al.* [2] extract eight facial motion elements as a facial feature, and they give the Mamdani model as an input to the fuzzy type-1 system. Their fuzzy system has an output with seven membership functions for six main facial expressions and one natural facial expression. However, they do not optimize the parameters of the membership functions used in their fuzzy system.

Jamshidnejad and Nordin [3] propose a type 1 fuzzy system based on a genetic algorithm for face expression detection that used twelve points on the face to extract geometric features of the face. These twelve points comprise the two middle corners of the eyes, the upper and lower eyelids of both eyes, the inner corner of the eyebrows, the corners of the mouth, and the upper and lower lips.

They extracted these 12 points manually from the images. They chose Gaussian functions for their fuzzy system membership functions. They used the bee queen genetic algorithm to optimize the parameters of these membership functions. They presented their proposed system for identifying four expressions of surprise, happiness, sadness, and anger, and did not consider two expressions of disgust and fear.

llBeigi and Shah Hosseini [4] used several appearance-based methods to extract facial features. For example, to extract the feature of the openness of the eyes, they used an edge detector. They divide the extracted features for classifying the six primary facial expressions, into main features and sub-features, and used the sub-features only when the values obtained from the main features are equal for the different facial expressions.

The main features used in this system are the amount of eye openness, mouth openness, eyebrow shrinkage, and lip corner displacement; also, the sub-features they chose include mouth width, nasal wrinkles, tooth appearance detection, eyebrow slope and the thickness of upper and lower lips. They extracted 573 law of their fuzzy system empirically.

Eventually, to classify the facial expressions, they use a fuzzy type 1 Mamdani model system and optimize the parameters of the membership functions of the system with a genetic algorithm. Their proposed system is capable of detecting facial expressions in images where some facial parts (such as eyebrows or mouth) are uncertain. An example of these images is shown. However, the type 1 fuzzy system they use for

their system has less power than type 2 fuzzy systems to face with uncertain facial expressions.

In our previous study [5] we proposed the method inspired by the brain that could be fast and accurate emulation of the inferior temporal cortex with feature selective hashing to recognize animals. We used KTH database containing 1239 images in 13 classes.

In the other previous work [6] we used a combination of HMAX model as feature extractor as well as Exterm Learning Machin (ELM) as classifies to mimic brain function of object recognition and physical features in our research. However, ELM has its own limitations, such as the complexity of the whole process and algorithms.

Halder *et al.* [7] present two methods for face recognition, when the class of facial expressions is available for several persons with values of their facial features, with the help of type 2 fuzzy references. The first method, which uses type 2 interval fuzzy sets, is similar to the method presented in [8]. They propose a new evolutionary algorithm to construct quadratic membership functions for each general type 2 fuzzy set. They presented their own method for identifying the four states of anger, disgust, fear, and happiness, and did not consider the two states of surprise and sadness. Also, in this paper, no comparison is made with its corresponding type 1 fuzzy method.

Facial expressions are for conveying a person's emotional state to observers. This is an efficient and fast way to recognize facial expressions. The proposed method is based on the results obtained in two publicly available data sets. Since the graphics processing unit based on the cafe module is used to perform this experiment, the time required to extract a feature is significantly reduced. The proposed model can be called any dataset and is a general recognition term that also includes recognition in static or video images [9].

Liu *et al.* have proposed a new dynamic face recognition method. Due to two critical issues of this problem, temporal flush and rhetoric are various modeling across temporal-spatially aligned regions. To obtain a set of states, which are represented as mid-level such as a bridge between the low-level features and the high-level semantics, as in the geometrical state of facial recognition, the criterion of facial expression recognition has shown its superiority over traditional methods for video-based on face recognition. As a framework, it is quite general and is not limited to state recognition [10].

2. Preparing Data and Results

Initially, to increase existing classes, we decided to collect data for the expression we wanted to. We used 18-megapixel camera, the number of participants was 60, all backgrounds were the same, and all participants were in the same room with the same illumination. All the faces were straight. The size of the pictures was 244*244.

These expressions included six modes like the following images (Figure 1).

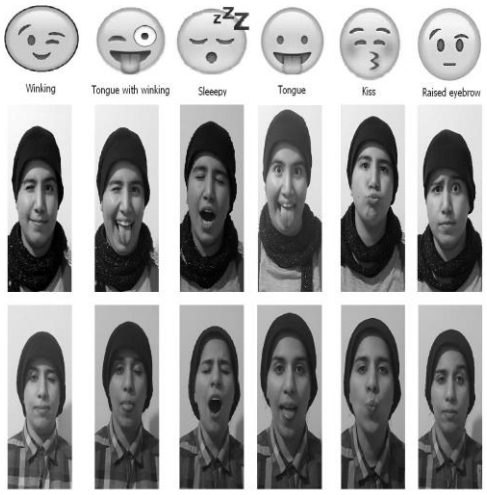

Figure 1. Six suggested modes for collecting data

But due to the inability to sum up all the expressions, eventually, four expressions were collected like the ones below; it means that we had 60 shots for each expression (Figure 2).

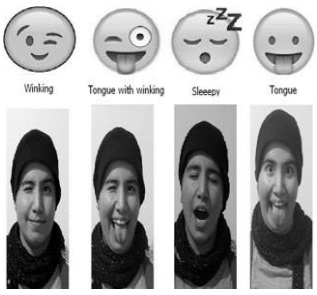

Figure 2. The final four states that our database contains. Examples of the initial images which have been collected

2.1. Steps to Prepare the Dataset

The collected images were raw, and collected images need to be processed to be ready to be used. These processes include: 1- Detecting a person's face in a photo; 2- Cropping the photo and extracting the face; 3- Matching the photo size; 4- Turning it to a gray photo.

For doing necessary preprocesses on dataset open CV Library is used, which is an open project source for data preprocessing, and we used it with Python language. In the first phase, which is face recognition, we used this library to identify faces in images. And the face is cut from the original image, and the length and height of the photo are 350, and in the end, the image becomes gray. Finally, images like the ones below are sent to the network for learning and testing (Figure 3).

Figure 3. Images are related to the blinking expression after the pre-processing steps

2.2.Parameters of the Training Phase

The implementation of this part is also used in Python language using a deep PyTorch framework. We used the Pre-Trained ResNet model (ResNet 18), and considering that the ImageNet dataset was included face data and that our data sets were small, we only re-weighted the last layer, i.e., fully connected, to adequate it to our workload.

The initial learning rate is also 0.01, and we gradually decrease the education rate, which is 0.001 every seven epochs.

Table 1. The value of parameters for network training

Parameters	Value
Momentum	0.9
Learining Rate	0.01
Epochs	20
Loss Function	Cross Entropy Loss

You can see all the network parameters in Table 1. In this section, we used the seven primary modes of the Jaffe dataset. It is trained on the central processing unit.

The best accuracy recorded during training is 85%. Due to using the pre-training model, we have achieved this with 20 epochs. Accuracy and Error Diagram of test and training data in the training phase are shown in the diagram below (Figure 4). Checking for accuracy and error is to ensure that the model chosen on our data is the appropriate model. The holistic schema of the proposed work is illustrated in Figure 4.

According to the description given in previous discussions, the full flowchart of this procedure is shown in Figure 4.

After the training phase, it is time to check the results. The first criterion (Table 2) we consider is the confusion matrix for each class. This matrix gives us a complete overview of how the network works in each class (Figure 6-7)

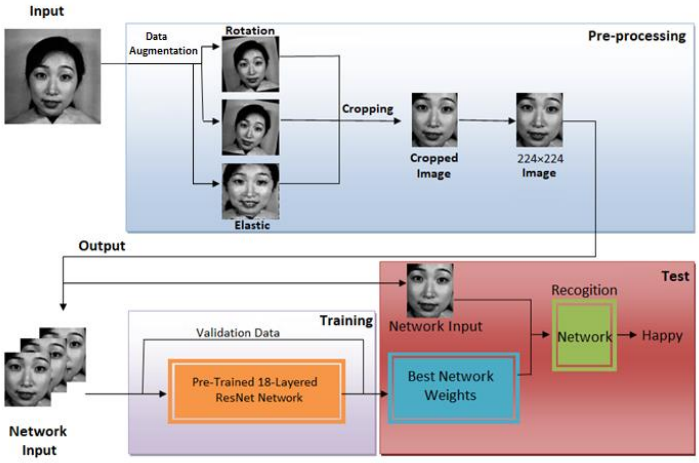

Figure 4. Our system architecture for recognizing human face expressions

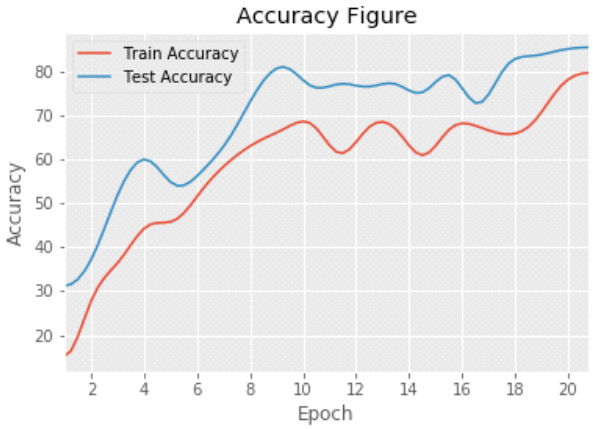

Figure 5. The accuracy of the test and training data sets in the network training phase are displayed in each phase

Table 2. Several criteria for image classification in a 10-class system

	Precision	Recall	F1-score	Support
Anger	0.69	0.82	0.75	11
Disgust	0.80	0.44	0.57	9
Fear	0.69	0.69	0.69	13
Happy	0.69	0.75	0.72	12
Neutrl	0.53	0.82	0.64	11
Sadnes	0.78	0.58	0.67	12
Sleepy	0.95	0.83	0.88	23
Surpre	0.73	0.73	0.73	11
Tongue	0.90	0.93	0.92	29
Winkig	0.88	0.88	0.88	26
Avg/Tl	0.80	0.79	0.79	157

2.3. Data Augmentation

In order to train our network with existing public data sets, we need to augment the data due to the low volume of data. There are several techniques for augmenting data. The methods used are elastic transmation and photo rotation. Elastic transmation: In this type of transmation, all pixels of the image are displaced by a factor in the horizontal and vertical axis. Here we have done the transmation from Sigma 4 and Alpha 34 values, as you can see in the figure. Rotation: Normally, we rotate the photo 10 degrees to the right and equally to the left (Figure 8).

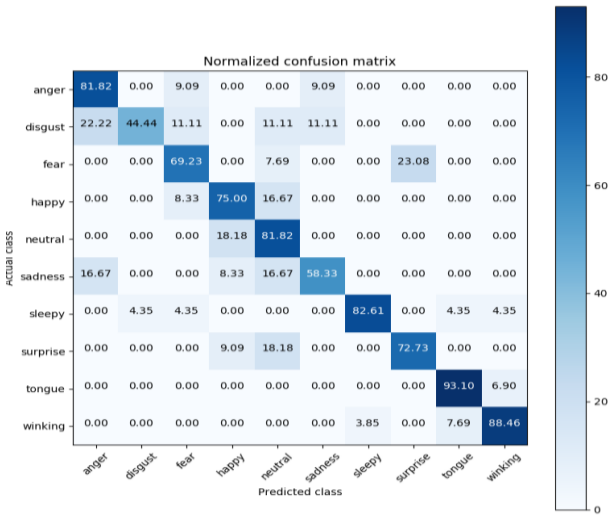

Figure 6. The normalized confusion matrix for 10 classes

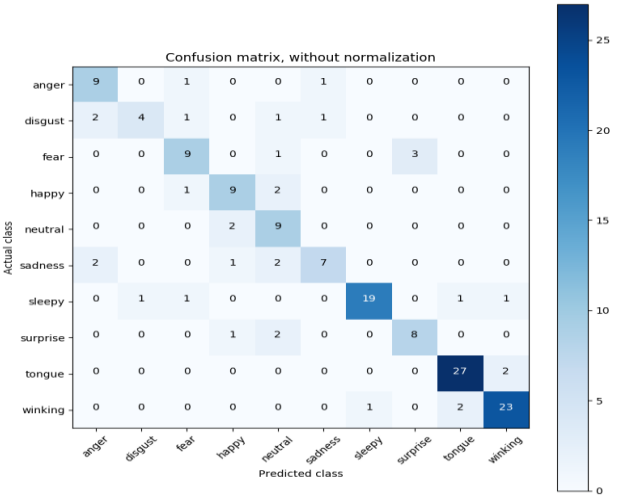

Figure 7. Confusion matrix for 10 classes without normalization

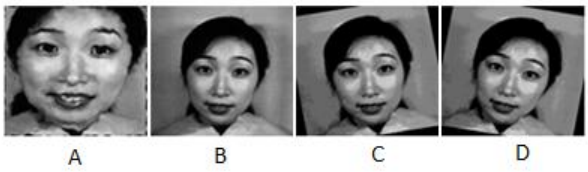

Figure 8. A. elastic transmation; B. without transmation; C. 10 degrees' rotation to the left; D. 10 degrees' rotation to the right

2.4. Increase Network Accuracy

So we used Jaffe data sets, and by applying data augmentation techniques, we tripled the size of the data sets and eventually trained the network with the created data sets. You will see the results below. The only difference between pre-training is the number of

training courses. Due to data augmentation we train the network in more than 100 courses (Figures 9-10).

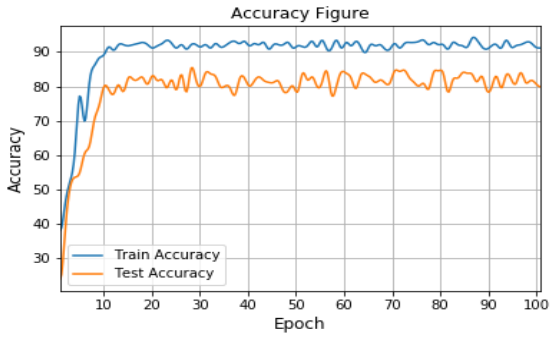

Figure 9. The accuracy diagram in the training phase of the network

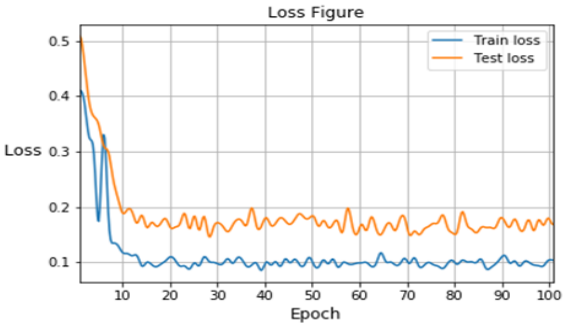

Figure 10. The error diagram in the training phase network

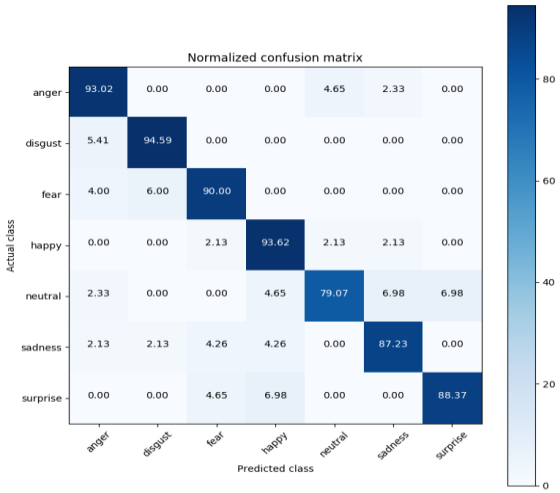

Figure 11. The normalized confusion matrix for seven expressions on the Jaffe dataset

The horizontal axis is for the values predicted by the system, and the vertical axis is for the actual label of the images. Higher values in the primary diameter indicate that the system is functioning correctly and that the predicted label is correct. The numbers in this table are expressed as a percentage, which means that the results are expressed as a percentage, over the whole image, for each expression.

Based on the results, we see that by using transfer learning, we can achieve high accuracy in less time in the intended task, which is image classification.

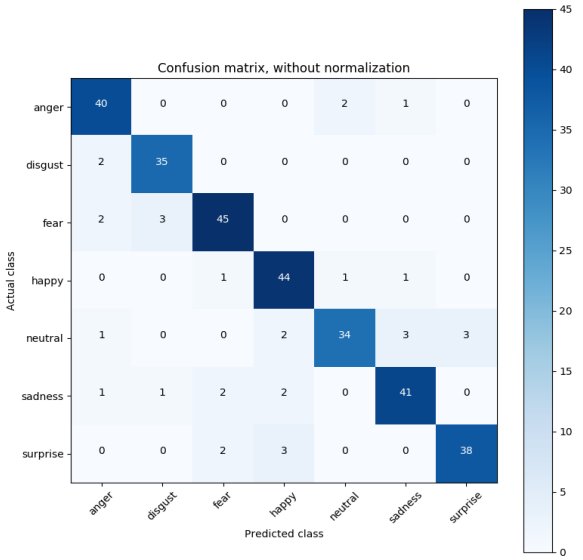

Figure 12. The confusion matrix for 7 expressions on the Jaffe dataset

The horizontal axis is for the values predicted by the system, and the vertical axis is for the actual label of the images. Higher values in the primary diameter indicate that the system is functioning correctly and that the predicted label is correct.

Table 3. Other criteria expressed for classification. Comparing results

	Precision	Recall	F1-score	Support
Anger	0.87	0.93	0.90	43
Disgust	0.90	0.95	0.92	37
Fear	0.90	0.90	0.90	50
Happy	0.86	0.90	0.90	47
Neutral	0.92	0.85	0.85	43
Sadness	0.89	0.88	0.88	47
Surprise	0.93	0.90	0.90	43
Avg/Total	0.89	0.89	0.89	310

The numbers in Table 3 are expressed as a percentage, which means that the results are expressed as a percentage, over the whole image, for each expression.

Table 4. Comparing our results with previous works on the Jaffe dataset

Method	Accuracy
Dynamic face recognition method [10]	55.87
Deep belief network [11]	86.00
Convolution Neural Networks [12]	82.10
ResNet network with triple data augmentation	89.00

3. Conclusion

Initially, considering what had been done in facial recognition, we decided to augment the expressions. We wanted to gather six more expressions, which finally, the gathered data set consisted of four expressions. We considered three expressions with seven basic expressions (Figures 11-12). Finally, we selected ten face expressions for classification. We described the results with the ResNet pre-training network. Then we applied our target network on a public data set called Jaffe. Of course, due to the very small number of data sets, we used some techniques to augment the data and eventually tripled the size of the data. We used three techniques, ten degrees rotation to the left and right, and finally the elastic transmation. Of course, due to working on a low data area on the deep architecture we chose, we needed to make some changes in the data size and parameters of the network for the network. Lastly, we used the concept of transfer learning, which means that we used a network that had already been trained on large data sets that included face images. The considering network was the pre-training ResNet network. Finally, we saw a 7 percent increase in accuracy compared to the highest accuracy in other previous work on this data set.

3.1. Future Works

Consequently, it can be argued that another way to improve network accuracy in less counted data other than augmenting data or using an architecture that is not deep, can be through transfer learning, a deep network with large data sets appropriate and trained for our task**.**

In this study, data is gathered from regularly trained image data sets that can network training steps be done in the future using video, in which case we can achieve the ultimate goal of most systems, which is to design a prompt system in facial recognition. Of course, with so much work being done in this area that can be guaranteed to be very accurate, even with low data, it is better to do prompt research on the functions of the face expressions in the real world. For example, for face recognition using face expressions, we can research facial recognition for people with psychological disorders like anxiety-based ones or developmental disorders such as autism.

References

1- Rao, K. S., Saroj, V. K., Maity, S. and Koolagudi, S. G. “Recognition of emotions from video using neural network models”, *Expert system with application*, vol.38, pp. 13181-13185, 2011.

2- Khanam, A., Shafiq, M. Z., and Akram, M. U. “Fuzzy based facial expression recognition”, *Congress on Image and Signal Processing*, vol. 1, pp. 598-602, 2008.

3- Jamshidenezhad, A. and Nordin, M. J. “An adaptive learning model base genetic for facial expression recognition”. *International Journal of Physical Sciences*, vol. 7, pp. 619-623, 2012.

4- Ilbeygi, Mahdi, and Hamed Shah-Hosseini. “A Novel fuzzy facial expression recognition system base on facial feature extraction from color face images”, *Engineering Applications of Artifical Intelligence*, vol. 25, pp. 130-146, 2012.

5- Seifzadeh S, Faez K. A cortex-like model for animal recognition based on texture using feature-selective hashing. *Paper presented at: The 2014 Iranian Conference on Intelligent Systems, Bam, Iran (ICIS)*; Feb 4-6, 2014.

6- Seifzadeh, Sahar, Mohammad Rezaei, and Omid Farahbakhsh. "A Computational Visual Neuroscience Model for Object Recognition." *Journal of Advanced Medical Sciences and Applied Technologies*, vol. 2, no.4, pp. 313-320, 2016.

7- Halder, A., Konar, A., Mandal, R., Chakraborty, A., Bhowmik, P., Pal, N. R., and Nagar, A.K. "General and interval type-2 fuzzy face-space approach to emotion recognition", *IEEE Transactions on System, Man and Cybernetics: System*, vol. 43, pp. 587-605, 2013.

8- Konar A., Chakraborty, A., Halder, A., Mandal, R., and Janarthanan, R. "Interval type-2 fuzzy model for emotion recognition from the facial expression", *Perception and Machine Intelligence*, pp. 114-121, 2012

9- VeenaMayya, Radhika M. Pai, ManoharaPai M. M. "Automatic Facial Expression Recognition Using DCNN", *Procedia Computer Science*, vol. 93, pp. 453 – 461, 2016.

10- M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionless on the spatiotemporal manifold for dynamic facial expression recognition, *in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1749–1756, 2014.

11- P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, *in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1812, 2014.

12- Andre Teixeira Lopes,.Edilson de Aguiar,. Alberto F. De Souza., Thiago Oliveira-Santos, "Facial Expression Recognition with Convolution Neural Networks: Coping with Few Data and the Training Sample Order", *Pattern Recognition*, vol. 61, pp. 610-628, 2017.