I.J. Image, Graphics and Signal Processing, **2019, 10, 1-7**
Published Online October 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2019.10.01

Facial Expressions Recognition in Thermal Images based on Deep Learning Techniques

Yomna M. Elbarawy
Faculty of Science, Al-Azhar University, Cairo, Egypt
Email: y.elbarawy@azhar.edu.eg

Neveen I. Ghali
Faculty of Computers & Information Technology, Future University in Egypt, Cairo, Egypt
Email: neveen.ghali@fue.edu.eg

Rania Salah El-Sayed
Faculty of Science, Al-Azhar University, Cairo, Egypt
Email: rania5salah@azhar.edu.eg

Received: 17 June 2019; Accepted: 07 August 2019; Published: 08 October 2019

Abstract—Facial expressions are undoubtedly the best way to express human attitude which is crucial in social communications. This paper gives attention for exploring the human sentimental state in thermal images through Facial Expression Recognition (FER) by utilizing Convolutional Neural Network (CNN). Most traditional approaches largely depend on feature extraction and classification methods with a big pre-processing level but CNN as a type of deep learning methods, can automatically learn and distinguish influential features from the raw data of images through its own multiple layers. Obtained experimental results over the IRIS database show that the use of CNN architecture has a 96.7% recognition rate which is high compared with Neural Networks (NN), Autoencoder (AE) and other traditional recognition methods as Local Standard Deviation (LSD), Principle Component Analysis (PCA) and K-Nearest Neighbor (KNN).

Index Terms—Thermal Images, Neural Network, Convolutional Neural Network, Facial Expression Recognition, Autoencoders.

I. Introduction

The awareness of facial expressions allows the prediction of the human status which can facilitate the adaptation through social situations. Also in computer-human interaction area facial expressions detection is very important as in driver fatigue detection in order to prevent the accidents on roads [1].

In 1997 Y. Yoshitomi et al. [4] introduced a system for FER using thermal image processing and NN with an accuracy recognition rate 90% applied over image sequences of neutral, happy, surprise and sad faces of one female. Deep Boltzmann Machine (DBM) model was used by Sh. Wang in 2014 for emotional recognition with an accuracy rate 62.9% over the USTC-NVIE database [5]. In 2015 a 72.4% recognition rate was achieved by Nakanishi et al. [6] using a thermal dataset consists of three subjects and three facial expressions of "happy", "neutral" and "other". The introduced system uses the 2D discrete cosine transform and the nearest-neighbor criterion in the feature vector space. Elbarawy et al. [21] in 2018 used the local entropy as a feature extractor and KNN as a classifier based on the discrete cosine transform filter achieving 90% recognition rate over the IRIS thermal database [14].

In the 1980s, CNN was proposed by Y. LeCun [7] as a NN is composed of two main consecutive layers defined as convolutional and subsampling. In 2012 a deep CNN was presented by Hinton et al. [8] since then image recognition based CNN was given a wide tension.

This paper presents CNN as an effective deep learning method to recognize facial expressions in thermal images by achieving acceptable accuracy of recognition rate compared with other recognition methods as explained later in experimental results section. CNN is specifically implemented as it reduces the pre-processing time by passing data through its multiple convolutional layers and making its own data filtering layer by layer which is worthy in real time applications [1]. The proposed system is applied over the IRIS dataset which has different poses for each subject and multiple rotation degrees as well as occlusion by glasses. Although using thermal images in recognition overcomes many challenges that faces recognition through visible images as illumination [2, 3], thermal images has its own challenges to overcome as temperature, occlusion by glasses and poses which will be tackled in this research.

The remainder of the paper is structured as follows: Section II briefly introduces feature extraction and classification methods which were used here, including NN, AE and CNN. Proposed system with phases of input, pre-processing, recognition and output are in Section III. In Section IV, an evaluation of the recognition rates

under different factors as network structure and pre-processing is illustrated. Finally, conclusions and results analysis are presented in Section V.

II. Preliminaries

This section briefly debates classic neural network, AEs and CNN applied on facial expression thermal images data to recognize expressions.

A. Neural Networks Based FER

NNs have two learning types, supervised and unsupervised techniques [9, 10]. This system uses the supervised technique feedforward with backpropagation training neural network [11] to train and produces a desired output as shown in Fig. 1. Number of input images $x^i$are $\sum_{i=1}^{60} x^i$ and the number of the hidden layer neurons are adjusted according to the recognition accuracy. The numbers of decision classes are 3 denoting different facial expressions in the used dataset.

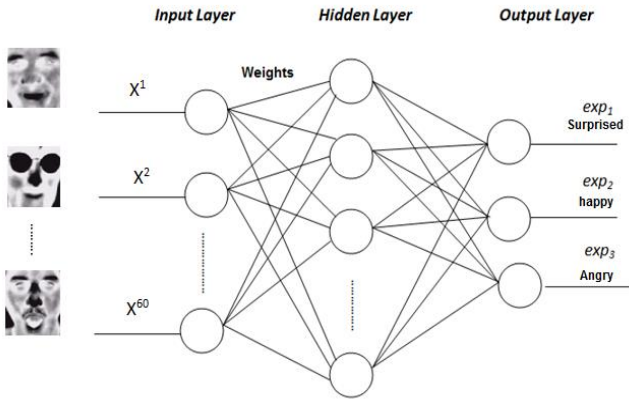


Fig.1. NN based facial expression recognition.

Scaled conjugate gradient backpropagation was used to recognize the facial expressions [12] and (1) represents the cross entropy function used for calculating the network error [13].

$$C(X|Y) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \ln\left(o\left(x^{(i)}\right)\right) + (1 - y^{(i)}) \ln(1 - o(x^{(i)})) \quad (1)$$

Where $X = \{x^{(1)}. \ldots. x^{(n)}\}$ are the set of input images in the training data, $Y = \{y^{(1)}. \ldots. y^{(n)}\}$ are the corresponding labels of input examples and $o(x^{(i)})$ is the output of the neural network given input x^i calculated as in (2)

$$o = f(\sum_{i=1}^{n} x^{(i)} w^{(i)}) \quad (2)$$

Where $w^i$is the network weight for input x^i. Neural network is introduced in algorithm 1.

Algorithm 1: Neural Network Algorithm

1) Input pre-processed thermal images.
2) Propagate forward through the network and randomly initiate weights.
3) Generate the output expressions.
4) Calculate the network error using cross-entropy Equation 1.
5) Re-adjust weights using (3).

$$\Delta w^{(i)} = rCx^{(i)} \quad (3)$$

Where r is defined to be learning rate with proposed value 0.01

6) Goto 2 until acceptable output accuracy is reached

B. Deep Autoencoder Neural Networks Based FER

Autoencoder neural networks is an unsupervised learning for features extraction. It sets the output nodes with same dimensions as the input nodes. Therefore, a training goal is created which does not depend on existing labels but on the training data itself. This made an unsupervised optimization of the full neural network [16].

As in general deep learning models, AEs read the input data images as a matrix or array of images. AEs mainly has two parts encoders $\alpha$and decoders β transited as in (4)

$$\alpha: X \to Y. \beta: Y \to X \quad (4)$$

Where X is the input vector and Y is the output one. The lower dimensional feature vector A is represented by (5).

$$A = f(\sum Wx + b) \quad (5)$$

Where W associated weight is vector with the input unit and hidden unit, b is the bias associated with the hidden unit.

Networks can use more than one AE for feature extraction. Features extracted by the first AE behaves as an input for the second AE and so on. Finally, classification is done at the softmax layer which unlike AEs its training is supervised using the training data labels. The softmax layer uses a softmax function to calculate the probabilities distribution of the images over different expressions. Architecture of AEs network is illustrated in Fig. 2.

The predicted probability for the j^{th} class given a sample vector x and a weighting vector w is given by (6).

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{n=1}^{N} e^{x^T w_n}} \quad (6)$$

Where $x^T w$ denotes the inner product of x and w.

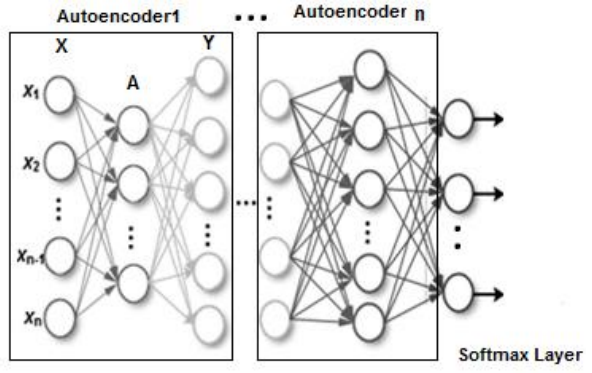


Fig.2. Autoencoders architecture of input vector X with length n.

C. Convolutional Neural Network Based FER

CNN differs from general deep learning models as it can directly accept 2D images as the input data, so that it has a unique advantage in the field of image recognition [18]. A four-layer CNN architecture designed to be applied over the used dataset, including two convolutional layers (C1, C2) and two subsampling layers (S1, S2). Finally, a Softmax classifier is used for image classification. General network architecture is illustrated in Fig. 3.

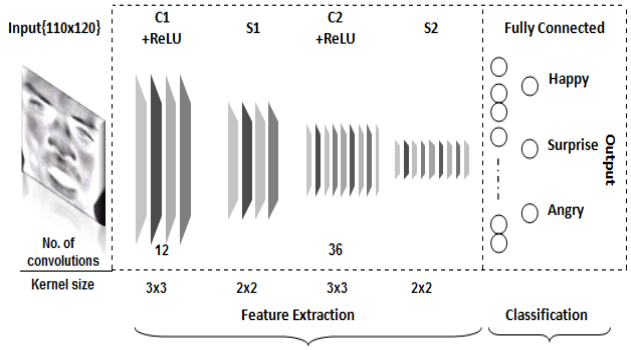


Fig.3. General convolutional neural network architecture.

Where C1 and C2 are used with different number of convolutions and kernel sizes. After each convolutional operation an additional operation is used called Rectified Linear Unit (ReLU) which is applied per pixel and replaces every negative pixel values in the feature map by zero value. Subsampling processes minimize the resolution of the functional map by max-pooling which takes the maximum response within the domestic features map size of the input (which is always the output of the convolutional layer) and reached a definite degree of invariance to deformity in the input [19]. At the fully connected layer, the output unit activation of the network made by softmax function which calculates the probability distribution of K different possible outcomes. After training, the network uses the cross entropy to indicate the distance between the experimental output and the expected output [20].

III. Facial Expression Recognition System Using CNN

A. System Overview

This section introduces CNN based facial expression recognition system. The system flow is shown in Fig. 4 where the input thermal images dataset under test is manipulated to detect face to reduce noise and unify size of images before feature extraction. CNN is implemented for feature extraction as discussed previously. From the extracted classes the analysis and accuracy is calculated. Details of each module are stated below.

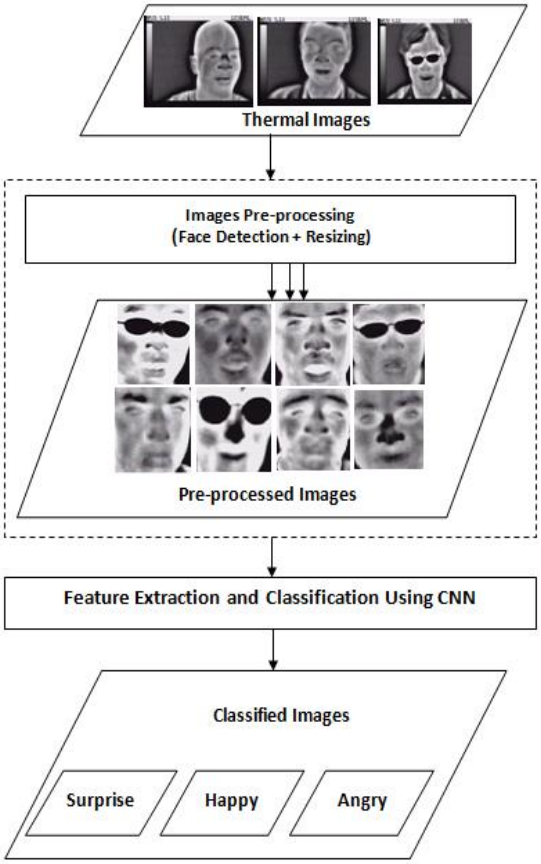


Fig.4. Scheme for facial expression recognition system

B. IRIS Dataset

The proposed system is applied over IRIS dataset [14], the standard facial expression dataset in the OCTBVS database which contain images in bitmap RGB format. The database contains thermal and visible images of 30 subject (28 males and 2 females) with size 320x240, collected by the long wave IR Camera (Thermal-Raythoen Palm IR-Pro) at the University of Tennessee having uneven illuminations and different poses. Each subject has three different expressions Surprise, Happy and Angry. Fig. 5 has samples of the visible thermal images in IRIS dataset with different rotations.

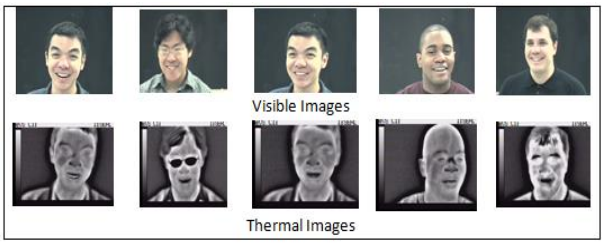


Fig.5. IRIS: different rotation sample images.

C. Images Pre-processing

The system uses 90 images (60 for training and 30 for testing) with different rotations, as well as occlusion by glasses and poses. Only poses less than 45° rotation were selected.

Image Pre-processing was done to reduce unnecessary regions in the original images through two main steps: First, face detection and extracting useful regions of the face and neglecting other images parts which hold non-essential background information using Viola-Jones algorithm [15]. Fig. 6 shows samples of detected faces with different expressions. Second step was resizing extracted faces with size 120x120 and preparing a two matrices one for training images and the other for the testing images.

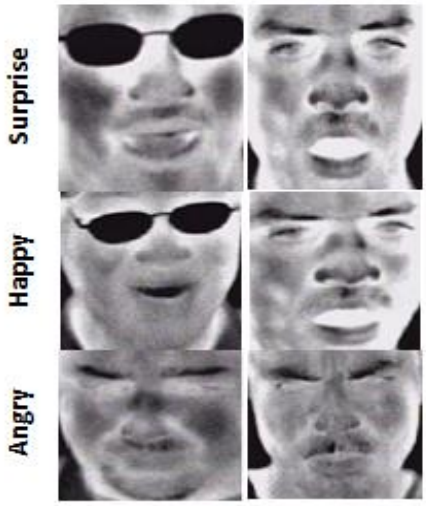
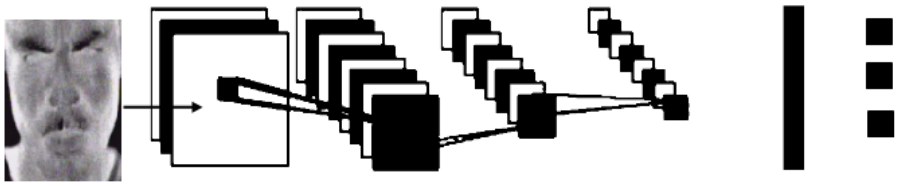


Fig.6. Sample of detected faces with different expressions.

D. Feature Extraction Based CNN

Proposed CNN applied over pre-processed images to extract features. CNN was robust for expression recognition with different number of convolutions and kernel sizes. The architecture of our CNN is illustrated in Fig. 7.



		Convolutional	Subsampling	Convolutional	Subsampling	Fully connected	Output
Image size	120x120	12x116x116	12x58x58	36x54x54	36x27x27		
Kernel size		3x3	2x2	3x3	2x2		

Fig.7. Proposed CNN architecture.

The architecture includes two convolutional layers, two sub-sampling layers and one fully connected layer. The network has an input of 120x120 grayscale images and outputs the credit of each expression. The first layer of the CNN is a convolution layer its objective is to extract elementary visual features as corners and borders of face parts, applies a convolution kernel size of 3x3 with stride of 1 horizontally and vertically and outputs 12 images with size 116x116 pixels. This layer is trailed by a sub-sampling layer that utilizes max-pooling with kernel size 2x2 at stride 2 to lessen the image to half of its size, outputs 12 images with size 58x58 pixels. Therefore, a new convolution layer performs 36 convolutions with a 3x3 kernel size and stride of 1 to guide of the past layer and trailed by another sub-sampling, with a 2x2 kernel size at stride 2, their aim is to perform the same operations as earlier, but to handle features at a lower level, perceiving contextual elements rather than borders and corners.

E. Output Module

The output module holds classification results and recognition rates which are calculated as explained earlier in section II. The general output consists of three classes, one for each expression. In case of applying CNN the outputs were given to a fully connected hidden layer that classifies images using the softmax function earlier in (6).

The same process is done for NN and AE as feature extraction and classification techniques to be compared with CNN.

IV. Experimental Results

Three models were applied over the selected data to show how to overcome the challenge of thermal images as temperature, occlusion by glasses and poses.

First, NN model applied over the selected data and different network structures were used. Table 1 shows

recognition accuracy of applying neural network with multiple different number of neurons 4, 6, 8, 10, 12 and 14. At 8 neurons testing recognition accuracy was the best which gives a 93.3% recognition rate.

Table 1. Neural network recognition accuracy.

No. of neurons	Accuracy (%)
4	73.3
6	80
8	**93.3**
10	86.7
12	83.3
14	76.7

Second model applied AE neural networks as a feature extraction and classification method. Applying different network structures could cause a great impact on the recognition rates hence a variant number of hidden layers were tested. Two level structured network were used and testing results are in Table 2. The maximum recognition rate was made when number of the hidden neurons was 16 and 32 for first level and 8 for the second, with testing accuracy 90%. Table 3 includes processing time for applying each structure which implicates that time is directly proportional to the number of hidden layers in each level.

Table 2. Recognition accuracy (%) using AEs.

H1 \ H2	**8**	**16**	**32**
8	76.7	86.7	86.7
16	**90**	80	80
32	**90**	83.3	86.7

The third applied model used a CNN with two convolutional layers. Convolutional layer one (C1) applied with different number of features map (4, 8,12 and 16) with size 3x3. The second convolutional layer (C2) applied with different number of features map (12, 24 and 36) with size 3x3 and both layers trained with 50 epochs, obtained results are shown in Table 4 with the highest recognition rate being 96.7% in testing. Table 5 shows the processing time for each applied CNN structure.

Tables 3 and 5 are holding the processing time of using AEs and CNN respectively and this time is strongly related to the hardware specification which the proposed system uses. This study used a system has 64 bit operating system with 4GB RAMs and processor speed 2.20 GHz.

Table 3. AEs processing time in seconds.

H1 \ H2	**8**	**16**	**32**
8	145.13	154.2	148.68
16	235.34	241.05	240.5
32	370.53	376.38	371.3

Table 4. Recognition accuracy (%) using CNN.

C1 \ C2	**12**	**24**	**36**
4	93.3	**96.7**	93.3
8	**96.7**	90	**96.7**
12	**96.7**	**96.7**	**96.7**
16	**96.7**	**96.7**	93.3

Table 5. CNN processing time in seconds.

C1 \ C2	**12**	**24**	**36**
4	41.1	44.5	49.7
8	47.08	50.8	57.1
12	55	60.06	66.06
16	63.45	68.6	77.8

Table 6. Overall accuracy results for recognition.

Method	Accuracy (%)
NN	93.3
AE	90
CNN	96.7

The overall results appear in Table 6 imply that using CNNs for expressions recognition in thermal images achieve high recognition rate with 96.7% in less time compared with other recognition methods (AE and NN), since it is easier to train with the pooling operations for down sampling and its many fewer parameters than stacked AEs with the same number of hidden units. Average processing time of deep recognition methods is illustrated in Fig. 8. Table 7 clarifies the confusion matrix of using the proposed architecture of CNN with accuracy 96.7%. Also illustrates the True Positive (TP) and False Negative (FN) rates respectively, where all images with happy and angry expressions are recognized with 100% accuracy. Surprised faces recognized with 90% TP rate, the other 10% FN rate confused with happiness expression.

Table 7. Confusion matrix of CNN with accuracy (96.7%).

	Surprise	**Happy**	**Angry**	**TPR**	**FNR**
Surprise	9	1	0	90%	10%
Happy	0	10	0	100%	0%
Angry	0	0	10	100%	0%

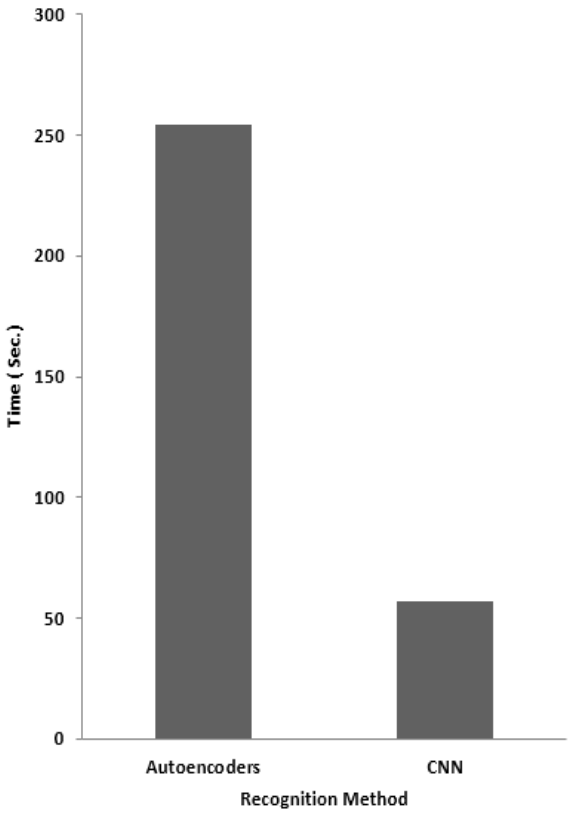


Fig.8. Average elapsed time of recognition methods based deep learning.

In real systems the subject image can be taken from a far distance which will make it hard to recognize. In order to add another challenge to the proposed system here, data augmentation was used to simulate the earlier difficulty by using a set of scaled images (horizontally and vertically). All training images were scaled by a randomly selected factor from the range vector [1 2]. Applying the same earlier CNN structure over the augmented images 10 times running for training and testing the average recognition rate was 70.9%.

V. Conclusion

Generally, experience and continuous testing is reliable to get the best network structure for a particular classification task. This paper holds two main approaches of conducted results, the first approach uses traditional NN for feature extraction and classification achieving 93.3% recognition rate. Second approach applies deep learning techniques as AEs and CNNs over the selected data. AEs has the longest processing time and lowest recognition rate with 90%.

A standard neural network with the same number of features as in CNN has more parameters, resulting an additional noise during the training process and larger memory requirements. CNN used the same features across the image in different locations at the convolution layer, thus immensely reducing the memory requirement. Therefore, the implementation of a standard neural system identical to a CNN will be permanently poorer.

The proposed CNN architecture as a deep supervised learner of features, detects facial expressions in thermal images with high recognition accuracy and less time compared with other deep learning model AE achieving 96.7%. In future, other architectures may be experimented to produce a higher accuracy.

References

[1] S. Naz, Sh. Ziauddin and A. R. Shahid, "Driver Fatigue Detection using Mean Intensity, SVM and SIFT", International Journal of Interactive Multimedia and Artificial Intelligence, In press, pp. 1 - 8, 2017.

[2] F. Z. Salmam, A. Madani and M. Kissi, "Emotion Recognition from Facial Expression Based on Fiducial Points Detection and Using Neural Network", International Journal of Electrical and Computer Engineering (IJECE), Vol. 8(1), pp. 52-59, 2018.

[3] Y. Wang, X. Yang and J. Zou, "Research of Emotion Recognition Based on Speech and Facial Expression", TELKOMNIKA (Telecommunication, Computing, Electronics and Control), Vol. 11(1), pp. 83-90, 2013.

[4] Y. Yoshitomi, N. Miyawaki, S. Tomita and S. Kimura, "Facial expression recognition using thermal image processing and neural network", 6th IEEE International Workshop on Robot and Human Communication, Sendai, Japan, pp. 380- 385, 1997.

[5] Sh. Wang, M. He, Z. Gao, Sh. He and Q. Ji, "Emotion recognition from thermal infrared images using deep Boltzmann machine", Front. Comput. Sci., Vol. 8(4), pp. 609-618, 2014.

[6] Y. Nakanishi, Y. Yoshitomi, T. Asada et al., "Facial expression recognition using thermal image processing and efficient preparation of training-data", Journal of Robotics, Networking and Artificial Life, Vol. 2(2), pp. 79-84, 2015.
[7] Y. Lecun, "Generalization and Network Design Strategies", Pfeifer, Schreter, Fogelman and Steels (eds)'Connectionism in perspective', Elsevier, 1989.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks". Advances in Neural Information Processing Systems (NIPS), pp. 1106-1114, 2012.
[9] X. Liao and J. Yu, "Robust stability for interval Hopfield neural networks with time delay", IEEE Transactions on Neural Networks, Vol. 9(5), pp. 1042–1045, 1998.
[10] J. J. Hopfield, "Neurons with graded response have collective computational properties like these of two-state neurons", Proceedings of the National Academy of Sciences, USA, Vol. 81(10), pp. 3088–3092, 1984.
[11] R. Amardeep and Dr. K T. Swamy, "Training Feedforward Neural Network with Backpropogation Algorithm", International Journal of Engineering and Computer Science, Vol. 6(1), pp. 19860-19866, 2017.
[12] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning", Neural Networks, Vol. 6(4), pp. 525-533, 1993.
[13] G. E. Nasr, E.A. Badr and C. Joun, "Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand", Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, Florida, USA, pp. 381-384, 2002.
[14] University of Tennessee: IRIS thermal-visible face database: http://vcipl-okstate.org/pbvs/bench/, last accessed Jul. 2018.
[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, Kauai, HI, USA, 2001.
[16] U. Schmid, J. Günther and K. Diepold, "Stacked Denoising and Stacked Convolutional Autoencoders. An Evaluation of Transformation Robustness for Spatial Data Representations", Technical Report, Technische Universität München, Munich, Germany, 2017.
[17] B. Leng, S. Guo, X. Zhang, and Z. Xiong, "3d object retrievalwith stacked local convolutional Autoencoder", Signal Processing, Vol. 112, pp. 119–128, 2015.
[18] K. Shan, J. Guo, W. You, D. Lu and R. Bie, "Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural-Network Structure", Proceeding of IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, pp. 123-128, 2017.
[19] Y. Yang, J. Yang, N. Xu and W. Han, "Learning 3D-FilterMap for Deep Convolutional Neural Networks", Proceeding of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.
[20] A. Ruiz-Garcia, M. Elshaw, A. Altahhan and V. Palade, "Stacked Deep Convolutional Auto-Encoders for Emotion Recognition from Facial Expressions", International Joint Conference on Neural Networks (IJCNN), Anchorage, Alaska, pp. 1586-1593, 2017.
[21] Y. M. Elbarawy, R. S. El-sayed and N. I. Ghali, "Local Entropy and Standard Deviation for Facial Expressions Recognition in Thermal Imaging" Bulletin of Electrical Engineering and Informatics, Vol. 7(4), pp. 580-586, 2018.

Authors' Profiles

Yomna M. Elbarawy received her B.Sc. and M.Sc. degrees in computer science from the Faculty of Science, Al-Azhar University, Cairo, Egypt in 2008 and 2014 respectively. She is currently a Ph. D. student at the same university. Her research areas are social networks analysis, computational intelligence, machine learning and deep learning technologies.

Rania Salah El-Sayed is lecturer in the department of Mathematics & Computer Science, Faculty of science, Al-Azhar University, Cairo. Egypt. She received her Ph.D and M.Sc in Pattern Recognition and Network Security from Al-Azhar University in 2013 and 2009 respectively. Her B.Sc degree in Math & Computer Science was received in 2004 from Al-Azhar University. In 2012, she received CCNP security certification from Cisco. Her research interests include pattern recognition, machine learning & network security.

Neveen I. Ghali received her B.Sc. from the Faculty of Science, Ain Shams University, Cairo, Egypt. Finished her M.Sc. and Ph.D. degrees in computer science from Faculty of Computers and Information, Helwan University, Cairo, Egypt in 1999 and 2003 respectively. She is currently a Professor in computer science and vice dean, Faculty of Computers and Information Technology in Future University in Egypt. Her research areas are artificial intelligence, computational intelligence and machine learning applications.

How to cite this paper: Yomna M. Elbarawy, Neveen I. Ghali, Rania Salah El-Sayed, " Facial Expressions Recognition in Thermal Images based on Deep Learning Techniques", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.11, No.10, pp. 1-7, 2019.DOI: 10.5815/ijigsp.2019.10.01

