Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks

Wisal Hashim Abdulsalam, Rafah Shihab Alhamdani, and Mohammed Najm Abdullah

***Abstract*—Its well known that understanding human facial expressions is a key component in understanding emotions and finds broad applications in the field of human-computer interaction (HCI), has been a long-standing issue. In this paper, we shed light on the utilisation of a deep convolutional neural network (DCNN) for facial emotion recognition from videos using the TensorFlow machine-learning library from Google. This work was applied to ten emotions from the Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations (ADFES-BIV) dataset and tested using two datasets.**

***Index Terms*—Facial emotion recognition, deep convolutional neural network, TensorFlow, ADFES-BIV, WSEFEP.**

I. Introduction

Facial expressions convey emotions and provide evidence on the personalities and intentions of people's. Studying and understanding facial expressions has been a long-standing problem. The first reported scientific research on the analysis of facial expressions can be traced back to as early as 1862 to Duchenne who wanted to determine how the muscles in the human face produce facial expressions. Charles Darwin also studied facial expressions and body gestures in mammals [1]. An influential milestone in the analysis of facial expressions was the work of Paul Ekman [2], who described a set of six basic emotions (anger, fear, disgust, happiness, sadness, and surprise) that are universal in terms of expressing, and understanding them.

Emotions are a fundamental component of being human [3]. Recognising emotions has extensive applications in the area of human-computer interaction (HCI) such as affective computing, interactive video games, human-robot interaction, and medical diagnosis [4], [5]. Emotions can be expressed through unimodal social behaviours, such as speech, facial expressions, and gestures, or bimodal behaviour such as speech and facial expressions, or they can be expressed through multimodal parameters such as audio, video and physiological signals [6].

The main part of the message is the facial expression, which constitutes 55% of the overall impression [7]. Therefore facial expressions are the key mechanism for understanding emotions [8]. In this paper, we shed light on the use of a deep convolutional neural network (DCNN) for facial emotion recognition from videos using the TensorFlow machine- learning library.

This paper is organized as follows. Section II describes the related work, Section III describes the proposed method, Section IV presents the results, and Section V presents the conclusions and future work.

Manuscript received May 13, 2018; revised October 14, 2018.

Wisal Hashim Abdulsalam and Rafah Shihab Alhamdani are with Iraqi Commission for Computers & Informatics, Baghdad, Iraq; Wisal Hashim Abdulsalam is also with the Computer Science Department, the College of Education for Pure Science-Ibn Al-Haitham, University of Baghdad, Iraq (e-mail: wisal.h@ihcoedu.uobaghdad.edu.iq).

Mohammed Najm Abdullah is with the Department of Computer Engineering, University of Technology, Baghdad, Iraq.

II. Related Work

Wingenbach *et al*. (2016) [9] developed and validated a set of video stimuli portraying three levels of intensity of emotional expressions, from low to high intensity. The videos were adapted from the Amsterdam Dynamic Facial Expression Set Bath Intensity Variations (ADFES-BIV) dataset, completing a facial emotion recognition task, which included six basic emotions in addition to contempt, embarrassment and pride, which were expressed at three different intensities of expression and neutral. Accuracy rates above the chance level of responding were found for all emotion categories, producing an overall raw hit rate of 69% for ADFES-BIV. The three intensity levels were validated as distinct categories, with higher accuracies and faster responses to high-intensity expressions to intermediate-intensity expressions, which had higher accuracies and faster responses than low-intensity expressions.

In [10], Sönmez (2018) challenged the classification experiment run on the ADFES-BIV dataset. The proposed automatic system uses the sparse-representation-based classifier and reaches the top performance of 80% by considering the temporal information intrinsically present in the videos.

III. The Proposed Method

The structure of the system is shown in Fig. 1.

This work was trained on ADFES-BIV dataset. This dataset is freely available upon request for scientific research. When we downloaded it, we found that its main folder contained a subfolder named 'Practice', which was used for testing, and the remaining videos were used for training. All the videos were 1.04s only in length. The proposed system contains two phases, training and testing, for which frames were extracted from the input video; the number of frames extracted was 13 frames/s. The number of samples in the dataset was not large, rendering it inappropriate for DCNNs. To solve this problem, we considered two procedures one for the training phase and the other for the testing phase. The dataset set used in this work contains three different levels of

doi: 10.18178/ijmlc.2019.9.1.759

intensity (low, medium and high), which made other researchers perform dedicated experiments on each level using traditional machine-learning methods. In our work, because we used a DCNN, which requires a large amount of training data since the inference accuracy improves by considering more data, and because what really mattered to us here was to recognize facial emotions (e.g. to know whether a person is happy or sad, regardless of whether, for example, he is happy to a low, medium or high degree) these three levels of intensity were merged under the name of their emotion. Thus, instead of obtaining 13 frames for one emotion from one person, we obtained 39 frames for each emotion and from each person, which means three times the first number (1560).

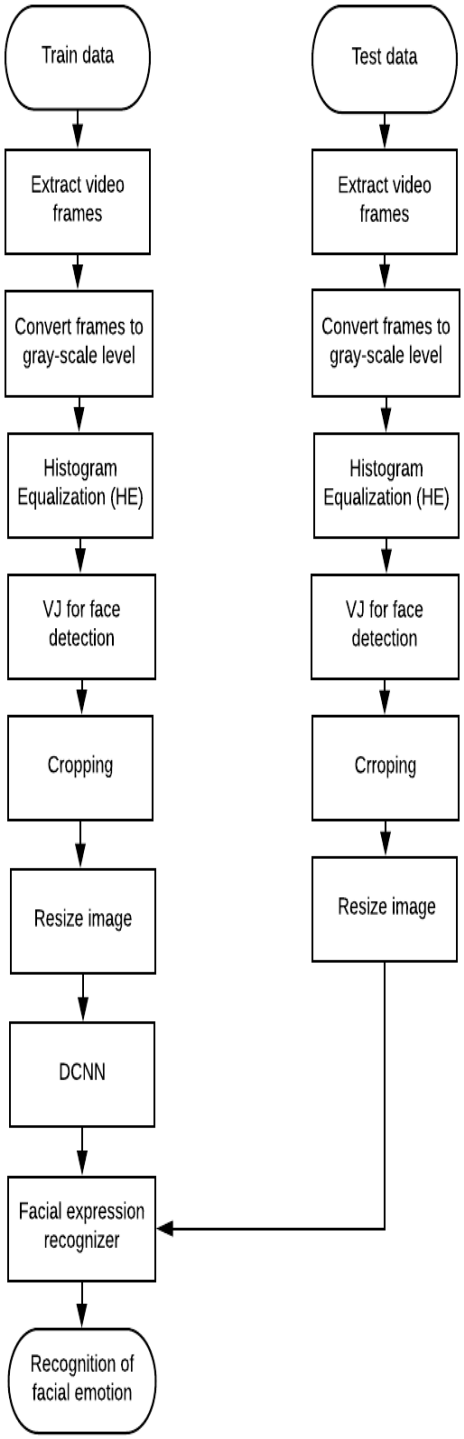

Fig. 1. The system structure.

During the testing phase, we downloaded another free dataset called the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP), which contained 210 photographs of facial expressions. All of them were merged with the frames extracted from the 'Practice' subfolder of the ADFES-BIV dataset and used for testing to check whether our system works well.

After extracting the frames from the input video, other operations were applied to obtain the facial emotions. There are ten facial expressions: six basics facial expressions (sadness, happiness, disgust, surprise, fear and anger) in addition to neutral, pride, contempt and embarrassment. All the extracted frames were converted to grey-scale and then histogram equalization (HE) was applied. After that, the Viola-Jones (VJ) algorithm was used to detect faces in these frames, followed by cropping to determine the region of interest (in other words, cropping the detected face from the whole image), and finally all the images were resized to a uniform size. All processed images were stored automatically in a folder. More details on these steps with example images will be discussed in the following subsections.

A. Preprocessing

Here, the following steps were made:

- 13 frames/s were extracted from each input video.
- All extracted frames were converted to grey-scale as shown in Fig. 2

Fig. 2. Original image converted to a grey-scale image.

- HE was applied to the frames to adjust the contrast and get an image clearer than the original one under the circumstances of different lighting conditions as shown in Fig. 3.

Fig. 3. Applying histogram equalization (HE).

B. Face Detection

- Faces in the video frames are detected using VJ algorithm [11], as shown in Fig. 4.

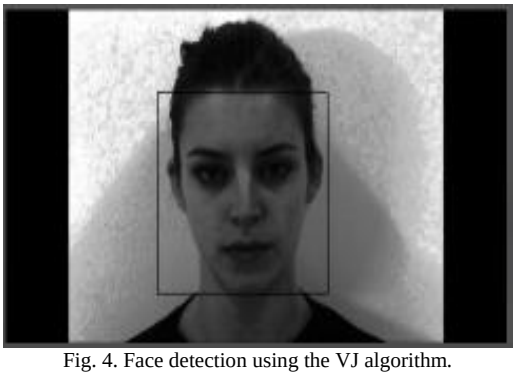

Fig. 4. Face detection using the VJ algorithm.

C. Postprocessing

- The face region detected by VJ algorithm was cropped to obtain an actual face region, as shown in Fig. 5.

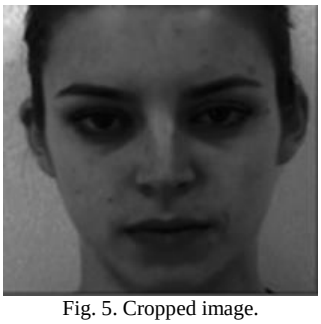

Fig. 5. Cropped image.

- All cropped frames were then resized to a uniform size, which is 70x70 pixels. This step is necessary to shorten the processing time, as shown in Fig. 6.

Fig. 6. Resizing the image.

Fig. 7 shows how the original video image that we started with will be after applying the preprocessing, face detection and postprocessing steps.

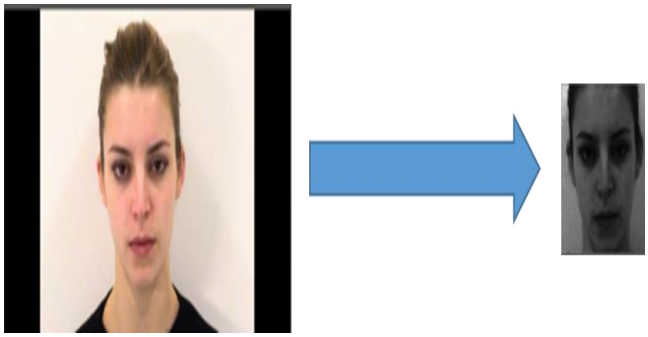

Fig. 7. Example of an original video image and how it will be after applying the preprocessing face detection and postprocessing steps.

After the previous steps are completed, the images are stored in a folder to use them later by the DCNN. Fig. 8 shows examples of these processed images.

D. DCNN Structure

Fig. 8. Examples of processed images stored in a folder.

DCNNs are trained using the popular TensorFlow machine-learning library. The network consists of three convolutional layers: the first layer had 16 filters, the second layers had 32, and the last layer had 64 filters, each one with a kernel size of 5x5, all with the same padding. We added a max-pooling layer with a 2x2 filters and stride of 2 after each layer. The three convolutional layers are followed by a fully connected (FC) linear layer with 2048 neurons and rectified linear unit (ReLU) activation and 50% dropout to improve the results of the model. A second FC layer was created with 1024 neurons. Softmax activation was used to derive probabilities for a particular example on each label, one for each target class 0–9 (corresponding to the number of expressions used), with linear activation. Batch normalizations were applied after all convolutional layers and the first FC layer to be less careful about initialisation. The network was trained from scratch using stochastic gradient descent with a specific batch size, with momentum set to 0.9. For gradient descent optimization algorithms, adaptive moment estimation was used.

Throughout the training stage, the selection of the parameters was based on trial and error within a range of recommended values in previous studies. However, these recommendations were applied to different types of datasets than what used in this paper and this meant that the results obtained from the experiments were different from what was reported in the literature. The results of all experiments performed in this study were obtained by using LENOVO Laptop with the following properties and software installed:

- Processor: Intel® CoreTM i7-7700 HQ.
- CPU @ 2.8 GHz.
- RAM 16 GB.
- GPU 4GB.
- 64-bit Operating System, X64 based processor.
- Windows 10 professional.

- OpenCV (2.4.8), Visual C++ under Visual Studio 2013 Integrated Development Environment (IDE), Python (3.5), Java (8) programming language under NetBeans (8.2) IDE, and TensorFlow-GPU (1.1).

TensorFlow is used to enable us to use GPU that is faster than CPU by ten times [12], the training process for 1000 epochs finished in only four hours. Other studies mentioned the time their experiments took, as in [13], while in this study the time was not explained because it depends on the properties of the hardware and software. The structure of the DCNN used in this work for the training data is shown in Fig 9.

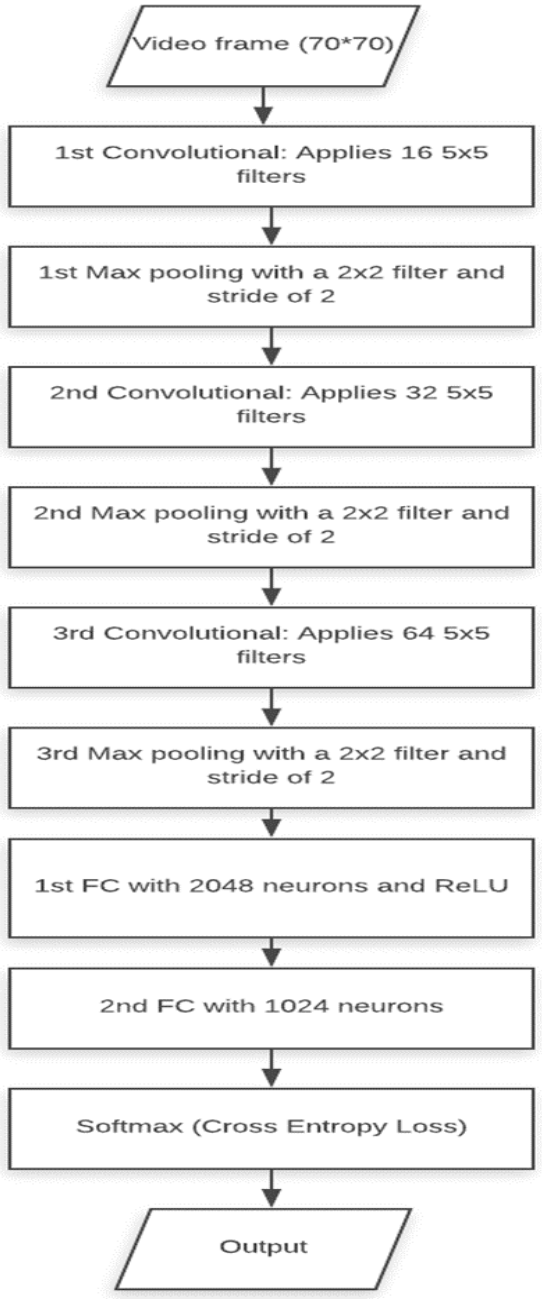

Fig. 9. DCNN structure.

E. Dataset Used

Two datasets were used in this work, ADFES-BIV and WSEFEP.

ADFES-BIV is an extension of the ADFES dataset, which was first introduced by Van der Schalk *et al*. [14]. ADFES is acted by 12 North European subjects (five females, seven males) and 10 Mediterranean actors (five females, five males) expressing the six basic emotions plus the three complex emotions of contempt, pride and embarrassment, in addition to neutral. Wingenbach *et al*. [9] created the ADFES-BIV dataset by editing the 120 videos played by the 12 North European actors to add three levels of intensity. They also created three new videos, displaying the same emotion at three different degrees of intensity -low, medium and high-, for a total of 360 videos. Every tape of ADFES-BIV starts with a neutral expression and ends with the highest expressive frame. The label of the video provides information on the acted emotion as well as its level of intensity (i.e. low, medium and high).

WSEFEP dataset contains 210 high-quality photographs of genuine facial expressions of 30 individuals [15].

IV. Results

We run a 10-classes experiment working with nine emotions plus the neutral faces; that is, our training and the test sets both contain the neutral face.

Table I shows the recognition accuracy that we obtained with each emotion for the testing phase using only the videos in the 'Practice' subfolder of the ADFES-BIV dataset.

TABLE I: The Recognition Accuracy with Each Emotion for Testing Phase Using 'Practice' Subfolder of the ADFES-BIV Dataset Only

Emotion	Recognition accuracy (%)
Happiness	100
Sadness	100
Anger	100
Surprise	100
Disgust	100
Fear	100
Neutral	100
Pride	100
Contempt	100
Embarrassment	100

Then WSEFEP dataset was used completely for testing by merging it with the testing frames from the first dataset used, to check if the system works well with other dataset. The total number of testing of facial frames for 10 emotions become 340, the number of correct recognized are 319 and the number of misrecognized are 21, then the recognition rate accuracy is 95.12%. Table II shows the recognition accuracy with each emotion for testing phase.

There are 21 misrecognized facial emotion images; (5) for disgust, (4) for fear, (4) for sad, (4) for surprise, and (4) for anger. Fig. 10 shows examples of misrecognized emotions. Fig. 10 A shows a disgust emotion, but misrecognized as surprise. Fig. 10 B shows a fear emotion, misrecognized as neutral. Fig. 10 C is a sad emotion, but misrecognized as disgust. Fig. 10 D is a surprise emotion, but misrecognized as

fear. Fig. 10 E is an anger emotion, but misrecognized as sad.

TABLE II: THE RECOGNITION ACCURACY WITH EACH EMOTION FOR TESTING PHASE USING 'PRACTICE' SUBFOLDER OF THE ADFES-BIV DATASET IN ADDITION TO WSEFEP DATASET

Emotion	Recognition accuracy (%)
Happiness	100
Sadness	90.7
Anger	90.7
Surprise	90.7
Disgust	88.4
Fear	90.7
Neutral	100
Pride	100
Contempt	100
Embarrassment	100

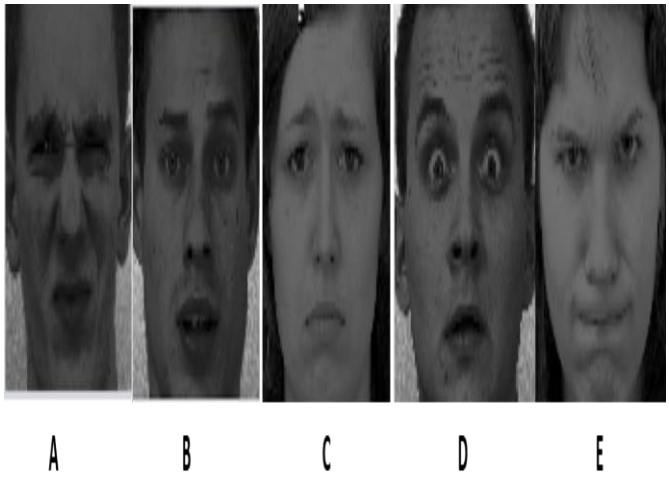

Fig. 10. Examples of misrecognized emotions.

A comparison with the previous studies in which the ADFES-BIV dataset was used is shown in Table III. Note that, in previous studies, dedicate experiments were performed for each level of intensity for nine emotions, except the neutral class because it does not have the three levels of intensity. Therefore, we took the average of the results obtained in these studies for the three experiments after adding their result with the neutral class.

TABLE III: COMPARISON WITH THE PREVIOUS STUDIES IN WHICH THE ADFES-BIV DATASET WAS USED

Related work	Recognition accuracy (%)
Wingenbach *et al*. (2016) [9]	68.9
Sönmez (2018) [10]	75.6
The proposed work using 'Practice' subfolder only of ADFES-BIV dataset	100
The proposed work using 'Practice' subfolder of ADFES-BIV dataset in addition to WSEFEP dataset	95.12

As shown in the table above, the accuracy of the proposed system is better than other related work. The detailed for each emotion accuracy with the related work using both the 'Practice' subfolder of ADFES-BIV dataset in addition to WSEFEP dataset appears in Table IV and Fig. 11.

TABLE IV: COMPARISON WITH THE PREVIOUS STUDIES FOR EACH EMOTIONS

Emotion	Wingenbach *et al*. (2016) [9] accuracy (%)	Sönmez (2018) [10] accuracy (%)	The proposed work
Happiness	84.6	86	100
Sadness	79.3	67	90.7
Anger	74.6	94.6	90.7
Surprise	92.3	83.3	90.7
Disgust	65	97.3	88.4
Fear	61.6	61	90.7
Neutral	89	36	100
Pride	42.3	91.6	100
Contempt	35	50	100
Embarrassment	64.6	89	100

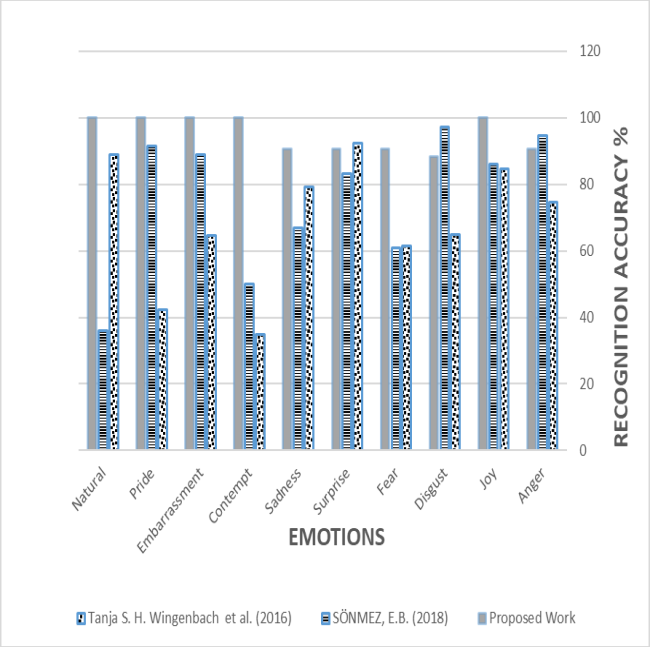

Fig. 11. Comparison with the previous studies for each emotions.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an algorithm for video-based emotion recognition with no manual design of features using a DCNN. The model considered the visual modality only and achieved an excellent recognition rate for the 10 used emotions.

Future work should attempt to combine our technique with other modalities such as audio modality, including working with other datasets.

REFERENCES

[1] S. Eleftheriadis, "Gaussian processes for modeling of facial expressions," Department of Computing, 2016, Imperial College London, p. 174.

[2] P. Ekman, *Darwin, Deception, and Facial Expression*, Annals of the New York Academy of Sciences, 2003, vol. 1000, no. 1, pp. 205-221.

[3] S. Brave and C. Nass, "Emotion in human–computer interaction," *Human-Computer Interaction*, 2003, p. 53.
[4] A. Houjeij *et al*., "A novel approach for emotion classification based on fusion of text and speech," presented at 2012 19th International Conference on Telecommunications (ICT), 2012, IEEE.
[5] P. R. Khorrami, *How deep learning can help emotion recognition*, in *Electrical & Computer Eng*., 2017, University of Illinois, Urbana-Champaign, p. 92.
[6] C. Maaoui, F. Abdat, and A. Pruski, "Physio-visual data fusion for emotion recognition," *IRBM*, 2014, vol. 35, no. 3, pp. 109-118.
[7] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, 2014.
[8] R. KalaiSelvi, P. Kavitha, and K. Shunmuganathan, "Automatic emotion recognition in video," presented at 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), 2014, IEEE.
[9] T. S. Wingenbach, C. Ashwin, and M. Brosnan, "Correction: Validation of the amsterdam dynamic facial expression set–bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions," *PloS One*, 2016, vol. 11, no. 12, p. e0168891.
[10] E. B. SÖNMEZ, "An automatic multilevel facial expression recognition system," *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 2018, vol. 22, no. 1, pp. 160-165.
[11] S. Al-Sumaidaee *et al*., "Facial expression recognition using local Gabor gradient code-horizontal diagonal descriptor," in *Proc. 2nd IET International Conference on Intelligent Signal Processing 2015 (ISP)*, 2015.
[12] O. Yadan *et al*., *Multi-gpu Training of Convnets*, 2013.
[13] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," presented at 2014 IEEE International Symposium on Circuits and Systems (ISCAS), 2014.
[14] J. Schalk *et al*., "Moving faces, looking places: Validation of the amsterdam dynamic facial expression set (ADFES)," *Emotion*, 2011, vol. 11, pp. 907-920.
[15] M. Olszanowski *et al.*, "Warsaw set of emotional facial expression pictures: a validation study of facial display photographs," *Frontiers in Psychology*, 2015, vol. 5, p. 1516.

Wisal Hashim Abdulsalam is a Ph.D. candidate at the Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq. She obtained her master's degree from the same institute and her B.Sc. degree from the Computer Science Department at the College of Education for Pure Science-Ibn Al-Haitham, University of Baghdad. She has published three research papers in national and international journals.

Rafah Shihab Alhamdani is the dean of the Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq. She has published more than 41 research papers in national and international journals and conferences.

She obtained her B.Sc. degree in agricultural economics in 1975, M.Sc. degree in agricultural economics (operations research) in 1986 and Ph.D. degree in Economic (operation research), in 1997 all from Baghdad University. She published eight books in various fields.

Mohammed Najm Abdullah received his B.Sc. degree in 1983 in electrical engineering from the College of Engineering, University of Baghdad. He received his M.Sc. degree in electronic and communication engineering from the same college in 1989 and his Ph.D. degree in 2002 in electronic and communication engineering from the University of Technology. Now, He currently works at the Department of Computer Engineering, University of Technology, Baghdad, Iraq. His areas of interest are air-borne computers, DSP software and hardware, PC interfacing, e-learning, information systems management, and wireless sensor networks. He published 42 research papers in national and international journals and conferences, as well as 13 books.