Emotion Recognition System Based on Hybrid Techniques

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

***Abstract*—Emotion recognition has important applications in human-computer interaction. Various sources such as facial expressions and speech have been considered for interpreting human emotions.**

The aim of this paper is to develop an emotion recognition system from facial expressions and speech using a hybrid of machine-learning algorithms in order to enhance the overall performance of human computer communication.

For facial emotion recognition, a deep convolutional neural network is used for feature extraction and classification, whereas for speech emotion recognition, the zero-crossing rate, mean, standard deviation and mel frequency cepstral coefficient features are extracted. The extracted features are then fed to a random forest classifier.

In addition, a bi-modal system for recognising emotions from facial expressions and speech signals is presented. This is important since one modality may not provide sufficient information or may not be available for any reason beyond operator control. To perform this, decision-level fusion is performed using a novel way for weighting according to the proportions of facial and speech impressions. The results show an average accuracy of 93.22 %.

***Index Terms*—Emotion recognition, convolutional neural network, tensorflow, ADFES-BIV, WSEFEP, SAVEE.**

I. Introduction

With the widespread of computers everywhere, new tools are needed to obtain responses from interactions with humans accordingly. At present, there are various ways for extracting feedback from humans such as voice tone, body movements and heart rate, but some of these do not provide suitable or accurate feedback. Feedback provides important information that can have a positive impact in different areas; for example, developers can create a music application that adjusts the type of music played according to the detected emotions of the user [1].

Knowing the emotions of others is intuitive in human daily social life, but when it comes to computers, this is much harder [2]. Emotion can be expressed via unimodal behaviours (e.g. speech, facial expressions, text, and gestures) or bi-modal behaviours (e.g. speech and facial expressions), or they could be expressed by multimodal approaches (audio, video and physiological signals) as shown in Fig. 1 [3].

The main part of the overall impression of a message is the facial expression (55%) whereas the audio part and semantic content contribute 38% and 7%, respectively [4].

Facial emotion recognition (FER) and speech emotion recognition (SER) can be used to build human computer interfaces, which are used in different applications in different fields, including the following [5]-[9]:

a) In the medical field for disease and pain detection.

b) In the psychological field, such as lie detection, autism and depression as well as the creation of appropriate therapeutic applications.

c) In the security field, such as access control, transaction authentication, and in automated teller machine (ATM).

d) In the teaching field, such as automated tutoring systems.

e) In the entertainment field, such as computer games that react according to the player's mood.

This work extends the one presented in [10] for FER but with some changes in the pre-processing step, and merges it with the one presented in [11] for SER on the decision level using a novel way for weighting of seven emotions which are the six basic emotions defined by Paul Ekman (fear, sadness, anger, surprise, disgust and happiness) [12] besides neutral emotion, because they exist in the three datasets used in this work

The subject of this work is relatively recent. If optimum results are obtained, some very important applications will be available. However, there is no agreed-upon or ideal method for emotion recognition; therefore, this work represents a step forward.

This paper is organised 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.

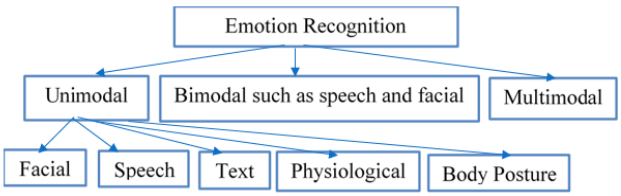

Fig. 1. Types of emotion recognition [3].

II. Related Work

In this part, we will review some of the most recent research work that was conducted for the last five years.

Mohit Shah *et al.* (2014) [13] performed a study aimed at recognizing emotion features extracted from facial expressions, language and speech from the Interactive Emotional Dyadic Motion Capture dataset. For the facial channel, multiple facial markers were extracted; for speech channel, prosodic and spectral features were extracted. Language-specific features were constructed using available transcripts with the dataset, and then replicated softmax

Manuscript received August 17, 2018; revised May 1, 2019.

Wisal Hashim Abdulsalam is with University of Baghdad, Iraq (e-mail: wisal.h@ihcoedu.uobaghdad.edu.iq).

models were used for recognition. The authors combined features in decision-level fusion. The results showed that using multi-modal fusion improves the accuracy up to 68.92% and showed that a turn of 1s duration can be classified in 666.65ms, which makes it appropriate for real-time implementation.

Samira Ebrahimi *et al.* (2015) [14] focused on hybrid convolutional neural networks (CNNs) and recurrent neural network (RNN) for facial expression analysis. The authors combined features from speech and facial expressions in decision-level fusion and achieved a higher classification accuracy (52.875%). They used the Acted Facial Expressions in the Wild 5.0 dataset for small-sized videos in their model training, as well as Toronto Face Database and FER2013, both for static images with the six basic emotions besides neutral emotion.

Sara Zhalehpour *et al.* (2016) [15] presented a fully automatic emotion recognition system using a video channel. For facial expressions, peak frame selection was used; for speech, mel frequency cepstral coefficients (MFCCs) and relative features based on perceptual linear prediction were extracted. They used for recognition the radial basis kernel support vector machine (SVM). Modalities were fused at the decision level, and the system showed promising results from two datasets in two languages, these are eNTERFACE and BAUM-1a.

Panagiotis Tzirakis *et al.* (2017) [16] proposed an emotion recognition system for facial and speech features in order to obtain the emotional content of different styles. This required robust feature extraction. Therefore, they utilised a CNN to use features from speech and from facial expressions using a ResNet-50 network architecture. The extracted features were fused together and fed to long short-term memory models and the system was trained on the RECOLA dataset.

Jingwei Yana *et al.* (2018) [17] proposed a multimodal emotion recognition framework via audio signals, facial landmarks and facial texture. Audio-signals were modelled with a CNN. Facial landmarks used the movement of facial muscles. A cascaded CNN and bi-directional recurrent neutral network were employed to extract the dynamic changes of facial textures. SVM and CNN were used to explore emotion related patterns. The authors fused these models at the feature level and decision level.

The main contributions noted in this work are as follows:

(a) Facial expressions and speech are a hybrid for emotion recognition using two trends of machine-learning a new one that use a custom deep convolutional neural network (DCNN) for FER and the traditional one that uses random forest (RF) for SER for seven emotions. To our knowledge, in previous works either traditional or deep learning (DL) methods were used only.

(b) Decision from facial and speech channels were integrated into decision level fusion using a novel weighted method according to the proportions of facial and speech impressions.

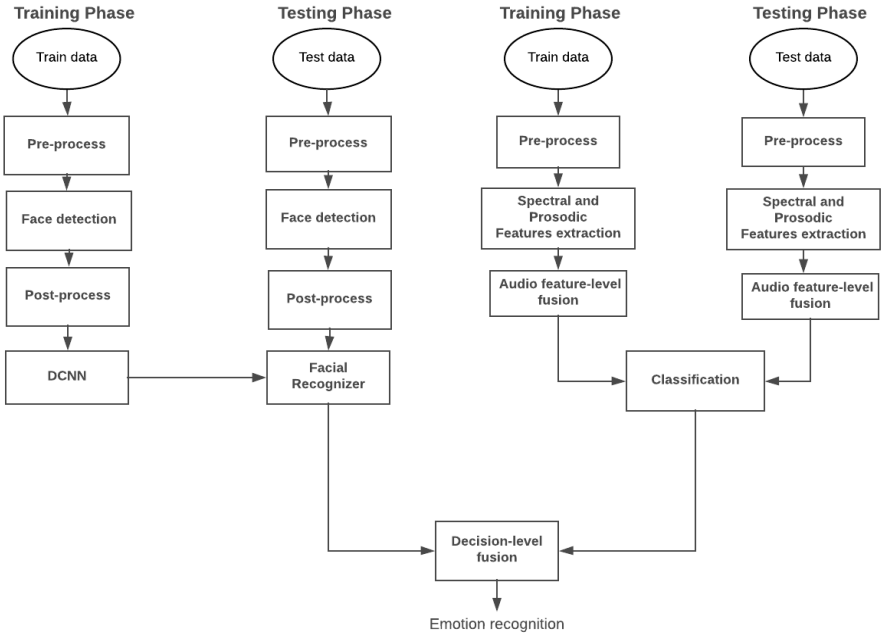

Fig. 2. General block diagram of the proposed system.

III. The Proposed Method

A general block diagram of the proposed system is illustrated in Fig. 2. The proposed system contains two parts: FER and SER. Both parts consist of data training and data testing. The initial step of both parts is to load the datasets.

A. Facial Emotion Recognition (FER)

FER consists of two phases: data training, and data testing. TensorFlow is used so that the graphics processing unit expedites the training process. Training is performed using the Amsterdam Dynamic Facial Expression Set and termed the Bath Intensity Variations (ADFES-BIV) dataset. This dataset has three levels of intensity (low, medium and high) for each emotion, if isolated experiments are performed on each level, then the limited number of the available samples will makes this dataset unsuitable for DCNN, and because it is important in this work to recognise the facial emotions (e.g. to

know whether a person is happy or sad) but not important to know his/her emotional level (happy to a low, medium, or high extend) the three levels were merged under the name of their emotion, and frames were extracted from each one, and then they were used for training the proposed model.

For both phases, the frames were extracted (13 frames per video) from the input video, and then only the last five frames from each video were considered because all the videos in this dataset begin with neutral emotion; and ends with the highest expressive emotion. With the exception of neutral video clips, the 13 extracted frames were adopted. Therefore, the total number of the extracted frames is as follows: 12actors × 3 levels × 6 emotions × 5 frames = 1080+ (12 × 3 × 1 for neutral × 13 = 468) = 1548 frames. Next, all extracted frames were converted to the grey-scale level. The Viola Jones (VJ) algorithm was applied to all video frames to detect the face region. The face region detected by the VJ algorithm was then cropped to remove the background information and obtain the actual face region in order to keep only expression specific features, which accelerate the expression recognition speed. In other words, the cropping operation was applied to the region of interest (ROI) of the image, which was extracted using the face detection box area.

Finally, this ROI is normalised by resizing the area to a size of 70 × 70 pixels. This step is necessary to shorten the processing time [18]. Fig. 3 shows an example of the steps applied to the reading video.

For the training phase, the result from previous steps entered to a DCNN that extracts features from the input without the need to a manual feature extraction. The output of the training process is a set of weights that used to speculate the test image to give it a single expression label and it is the final system output.

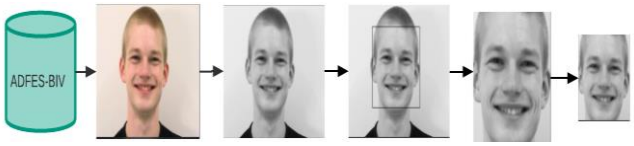

Fig. 3. Example of the steps applied to the reading video.

Fig. 4. Facial emotion recognition system.

For the testing phase, a 'Practice' folder in the ADFES-BIV dataset was used. The frames were extracted from each video in the same way as in the training phase: 1actor × 1 level × 6 emotions × 5 frames = 30 + (13 frames for neutral) = 43 frames in addition to 210 frames from the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) dataset for static facial expressions that are added completely to the test set = 253 total frames. The output is a fixed size image that use network weights learned during the training phase for the recognition of facial emotions. The details of the network can be found in [10] Fig. 4. shows the FER system.

B. Speech Emotion Recognition (SER)

Speech is the first way of communication invented by humans [19], so it represents the second part of the proposed work. The first step is to load the dataset, dividing it into 75% for training and 25% for testing sets. Both phases comprise the following steps:

(a) First, Pre-processing is performed in order to apply a pre-emphasis filter to the signal for high frequency amplification, and then framing is used to divide the signal into a sequence of frames to analyse each frame independently, which is then represented by a single feature vector. The number of frames varies since it depends on the length of the audio track. The frame block in this work has a length of 512 with a 50% overlap between frames, and a sample frequency of 16000Hz. Finally, a Hamming window is used to eliminate the edges and excess silence periods that do not contain information.

(b) Second, feature extraction, feature fusion and classification are performed. In the current work, MFCC features were extracted, as spectral features from the frames, and then the zero crossing rate (ZCR), energy and pitch (the last two features were neglected because they did not yield good results) were extracted as prosodic features in addition to the mean and standard deviation (SD) from the original signal. Each one of them returned only a single value giving three features, whereas the first 12 MFCCs returned two-dimensional values. Therefore, vector quantisation was used to convert them into a unidimensional form and then fused them with the ZCR, mean and SD features to form a single feature vector using a direct concatenation method to enhance the effect of SER. Finally, the result was a vector of 15 features. The extracted features were then fed into RF to generate predictions.

Test speech samples were classified using a classifier and the information provided by the training model.

C. Information Fusion

In this work, decision-level fusion is used, since multiple modules can be processed asynchronously. The accuracies from both modalities are fused to generate the final accuracy using three methods: average, weighted and majority voting. The average method is computed using the following equation:

$$\text{Average} = (Af+As)/2,$$

where:

Af is the facial accuracy and

As is the speech accuracy.

For the weighted method, two coefficients are used: ω1 as the weight for facial expressions, and ω2 as the weight for speech. They are then multiplied to the results from FER and SER accuracy, as in the following formula:

$$R = \omega 1 \times Af + \omega 2 \times As,$$

where

R represents the result of applying the weighted method and ωi is the weight given to each model, with the summation of weights = 1.

The weights used followed those used by Mehrabian [4] who suggested that communication between human beings is represented in facial expressions by 55%, speech by 38% and that the remaining 7% is related to other modalities, but to our knowledge was not applied in any form of FER and/or SER. Therefore, the weight used was 0.55 for facial recognition and 0.38 for speech. The remaining weight of 0.07 was added to these ratios according to the following equations:

$$\text{Facial weight} = (0.07 \times \omega 1) + \omega 1 = 0.5885 \text{ which we round to } 0.6.$$

$$\text{Speech weight} = (0.07 \times \omega 2) + \omega 2 = 0.4066 \text{ which we round to } 0.4.$$

D. Dataset Used

Three datasets were used in this work: ADFES-BIV, WSEFEP, and Surrey Audio-Visual Expressed Emotion (SAVEE).

ADFES-BIV is an extension of the ADFES dataset expressing the six basic emotions plus three complex emotions (i.e. contempt, pride and embarrassment) besides neutral emotion. Wingenbach *et al.* [20] created this dataset by editing 120 videos played by 12 North European actors (5 females, 7 males) to add three levels of intensity. They 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.

WSEFEP is a high quality photograph of genuine dataset facial expressions for the six basic emotions besides neutral with 210 high quality photographs of 30 subjects [21].

SAVEE dataset is an English audio-visual emotional dataset from four English male subjects for the six basic emotions beside neutral emotion. The dataset contains 120 utterances per actor (15 utterances for each of the six basic emotions plus 30 for neutral) yielding 480 sentences in total. 60 markers were recorded on the faces of the subjects to extract the visual features [22].

The three datasets used are available free of charge for research purposes.

IV. Results

Table I shows the confusion matrix of the seven emotions by merging the WSEFEP dataset with the testing frames from the ADFES-BIV dataset.

TABLE I: Confusion Matrix for the Seven Emotions Using ADFES-BIV and WSEFEP Datasets

Emotion	Happy	Sad	Anger	Surprice	Disgust	Fear	Neutral
Happiness	35	0	0	0	0	0	0
Sadness	0	31	0	1	2	1	0
Anger	0	1	31	2	1	0	0
Surprise	0	0	1	31	1	1	1
Disgust	0	1	0	2	30	1	1
Fear	0	1	0	0	0	31	3
Neutral	0	0	0	0	0	0	43

TABLE II: Different Performance Criteria for the Testing Phase Using the 'Practice' Subfolder of the ADFES-BIV Dataset in Addition to the WSEFEP Dataset

Emotion	**Accuracy (%)**	**Recall (%)**	**Specificity (%)**	**Precision (%)**	**F-measure (%)**
Happiness	100	100	0	100	100
Sadness	97.23	91.18	1.83	88.57	89.86
Anger	98.02	96.88	1.8	88.57	92.54
Surprise	96.44	86.11	1.84	88.57	87.32
Disgust	96.44	88.24	2.28	85.71	86.96
Fear	97.23	91.18	1.83	88.57	89.86
Neutral	98.02	89.58	0	100	94.5
Average (%)	97.25	91.88	1.37	91.43	91.58

TABLE III: Accuracy with Each Emotion for the Testing Phase Using an RF Classifier

Emotion	**Accuracy (%)**
Happiness	80.3
Sadness	87.87
Anger	94.34
Surprise	88.79
Disgust	83.67
Fear	84.4
Neutral	90.92

TABLE IV: Different Performance Criteria for the Testing Phase Using the SAVEE Dataset

Parameters	**Fusion Method (%)**
Specificity	4.17
Accuracy	87.18
Recall	78.54
Precision	78.75
F-measure	78.05

Different performance criteria are shown in Table IV.

Different performance criteria are shown in Table II.

For SER, Table III shows the accuracy with each emotion for the testing phase obtained from applying feature fusion with an RF classifier on the SAVEE dataset.

The simplest fusion scheme is decision-level fusion, which relies on the assumption that different modules are independent of each other. In this method, each module is classified separately, and the output of each module is integrated to obtain the global decision of the expressed emotion. Three different methods are tested including majority voting. However, majority voting does not consider the confidence in the single decisions as some modalities (facial expressions in this work) may be better suited for the recognition of certain emotions. Therefore, other schemes have been explored, such as average, which yields good results, but the proposed weighting method yields better results and takes into account the psychological view. The final accuracies obtained from applying the three methods are shown in Table V.

TABLE V: FINAL ACCURACY OBTAINED FROM APPLYING DECISION LEVEL FUSION USING DIFFERENT METHODS

Method	Accuracy (%)
Voting	97.25
Average	92.22
Weighting (proposed)	93.22

Finally, it is difficult to compare between the proposed hybridisation system and the published studies, as they did not use the same datasets that were used in this work. In addition, they hybridised both channels either using only DL methods (which yield good accuracy but require more time and data for training), or using only traditional methods (which require less data and time for training but have a lower accuracy), whereas in the proposed work a mix between DL methods (DCNN for FER) and traditional machine-learning methods (RF for SER) was applied to balance the amount of data, time and accuracy required. More about a comparison of machine learning and DL can found in [23].

V. CONCLUSIONS AND FUTURE WORK

In this paper, a novel weighting method based on a psychological view was used, which enhanced the results of the proposed system.

Future work should attempt to combine our technique with other modalities, such as physio-recognition and try to cover the missing 7% to make decisions more confident with a higher accuracy.

REFERENCES

[1] H. G. Valero, *Automatic Facial Expression Recognition*, School of Computer Science, University of Manchester, 2016.

[2] Q. Yao, "Multi-sensory emotion recognition with speech and facial expression," Umi Dissertations Publishing, University of Southern Mississippi, 2014.

[3] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial emotion recognition: A survey," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 7, no. 11, p. 9, November 2018.

[4] A. Mehrabian, "Communication without words," *Communication Theory*, pp. 193-200, 2008.

[5] N. Kauser and J. Sharma, "Automatic facial expression recognition: A Survey based on feature extraction and classification techniques," pp. 1-4.

[6] K. A. Bozed, O. Adjei, and A. Mansour, "Detection of facial expressions based on morphological face features and minimum distance classifier," in *Proc. 14th International Conference on Sciences and Techniques of Automatic Control & Computer Engineering*, pp. 487-493.

[7] K. J. Kantharia and G. I. Prajapati, "Facial behavior recognition using soft computing techniques: A survey," in *Proc. 2015 Fifth International Conference on Advanced Computing Communication Technologies*, pp. 30-34.

[8] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing*, vol. 30, no. 10, pp. 774-784, 2012.

[9] J. S. Devi, S. Yarramalle, and S. P. Nandyala, "Speaker emotion recognition based on speech features and classification techniques," *International Journal of Image, Graphics and Signal Processing*, vol. 6, no. 7, p. 61, 2014.

[10] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial emotion recognition from videos using deep convolutional neural networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, p. 6, 2019.

[11] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Speech Emotion Recognition Using Minimum Extracted Features," in *Proc. 1st Annual International Conference on Information and Sciences (AiCIS)*, pp. 58-61.

[12] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205-221, Dec, 2003.

[13] M. Shah, C. Chakrabarti, and A. Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Proc. 2014 IEEE International Symposium on Circuits and Systems*, pp. 754-757.

[14] S. E. Kahou *et al.* Recurrent neural networks for emotion recognition in video. [Online]. Available: http://www.professeurs.polymtl.ca/christopher.pal/RNN-emotions-kahou.pdf

[15] S. A. Zhalehpour, Z. Erdem, and C. Eroglu, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image and Video Processing*, vol. 10. no. 5, PP. 827-834, 2016.

[16] P. Tzirakis *et al.*, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[17] J. Yan *et al.*, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27-35, 2018.

[18] H. D. Najeeb, "Artificial neural network for tiff image compression," *Ibn Al-Haitham Journal for Pure and Applied Science*, vol. 30, no. 1, pp. 246-261, 2017.

[19] B. J. Al-Khafaji, "Proposed speech analyses method using the multiwavelet transform," *Ibn Al-Haitham Journal for Pure And Applied Science*, vol. 27, no. 1, pp. 394-401, 2017.

[20] 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*, vol. 11, no. 12, p. E0168891, 2016.

[21] M. Olszanowski *et al.*, "Warsaw set of emotional facial expression pictures: A validation study of facial display photographs," *Frontiers in Psychology*, vol. 5, p. 1516, 2015.

[22] S. Haq, P. J. Jackson, and J. Edge, "Speaker-Dependent Audio-Visual Emotion Recognition," in *Proc.* AVSP 2009, pp. 53-58.

[23] S. R. Shakya, C. Zhang, and Z. Zhou, "Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data," *International Journal of Machine Learning and Computing*, vol. 8, no. 6, 2018.

Wisal Hashim Abdulsalam currently works at the Computer Science Department, College of Education for Pure Science-Ibn Al-Haitham, University of Baghdad. She received her Ph.D. degree from the Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics in 2019. She obtained her master's degree from the same institute in 2012 and her B.Sc. degree from the

Computer Science Department, College of Education for Pure Science (Ibn Al-Haitham), the University of Baghdad in 2003. She published six research papers in national, international journals and conferences.

Rafah Shihab Alhamdani is the Dean of the Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq. 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 economics (operations research) in 1997 all from Baghdad University. She has published many research papers in national and international journals and conferences, as well as books in various fields.

Mohammed Najm Abdullah currently works at the Department of Computer Engineering, University of Technology, Baghdad, Iraq. He 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 electronics and communication engineering from the same college in 1989 and his Ph.D. degree in 2002 in electronics and communication engineering from the University of Technology. He published many research papers in national and international journals and conferences, as well as books.