TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 2, April 2021, pp. 523~530
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i2.18322

Real time ear recognition using deep learning

Ahmed M. Alkababji, Omar H. Mohammed
Department of Computer Engineering, University of Mosul, Iraq

Article Info

Article history:

Received Aug 9, 2020
Revised Oct 1, 2020
Accepted Oct 20, 2020

Keywords:

CNN
Ear recognition
Faster R-CNN
PCA
Real time

ABSTRACT

Automatic identity recognition of ear images represents an active area of interest within the biometric community. The human ear is a perfect source of data for passive person identification. Ear images can be captured from a distance and in a covert manner; this makes ear recognition technology an attractive choice for security applications and surveillance in addition to related application domains. Differing from other biometric modalities, the human ear is neither affected by expressions like faces are nor do need closer touching like fingerprints do. In this paper, a deep learning object detector called faster region based convolutional neural networks (Faster R-CNN) is used for ear detection. A convolutional neural network (CNN) is used as feature extraction. principal component analysis (PCA) and genetic algorithm are used for feature reduction and selection respectively and a fully connected artificial neural network as a matcher. The testing proved the accuracy of 97.8% percentage of success with acceptable speed and it confirmed the accuracy and robustness of the proposed system.

Corresponding Author:

Ahmed M. Alkababji
Department of Computer Engineering
University of Mosul
Mosul, Iraq
Email: ahmedalkababji72@uomosul.edu.iq

1. INTRODUCTION

Biometric has lately been getting interests in many popular media. It deals with the identification based on the behavioral or physiological characteristics of individuals. Commonly used physiological characteristics involve fingerprint, iris, face, and ear, and behavioral characteristics involve voice, gait, and signature [1, 2]. In many application areas using automatic machine learning to identify people from the image of the ear is a challenging problem. Most past research in this field has largely been focused on capturing the ear image in a controlled condition [3]. Ear recognition, as shown by a number of researchers, is a practical alternative to more common physiological characteristics such as iris, face, and fingerprint because the ear is less invasive to capture, stable over time, and does not need as much control during acquisition as other biometrics [4]. With the spread of coronavirus, ear recognition has become an important way to identify people wearing masks with no physiological contact.

In the last decade, the improvements in the capabilities of computation and the introduction of active methods to train deep neural network has led to address a wide area of challenges in computer vision. In the latest years, deep learning got a lot of attention due to its capability to learn features from data [5]. Deep learning represents learning approaches with multiple representation levels attained by making simple but nonlinear modules that transform the representation of one level into a representation of a higher level. The key issue in deep learning is that the layers of features are learned from data using a general-purpose learning procedure and not designed by human engineers [6, 7]. State of the art computer vision tasks such as object detection, semantic segmentation, and image classification achieved promising performance with convolutional neural networks (CNNs). CNNs that are trained with big amounts of data can solve tasks even for which they had not been trained. It contains two phases, automated feature learning is the first phase and classification is the other, both of them can be successfully trained in tandem through gradient descent of the error surface. CNN based ear recognition methods generally exploit an already trained object classification model, so-called a pre-trained deep CNN model [8-12]. M. Rungruanganukul [13], used deep learning for a vision-based technique for gesture classification of hand. By collecting a dataset for hand counting from 0 to 5, then build CNN from scratch to classify hand count by training it with the collected dataset. A grayscale of each image of the hand is fed to the CNN then the CNN gives one of six categories of hand, in this research for ear recognition the CNN was not used as an end-to-end recognition, instead the output of it first fully connected layer is used as feature extraction. Object detection, in general, is successfully achieved using faster region based convolutional neural (faster R-CNN). Faster R-CNN is composed of two components; for proposing candidate regions a fully convolutional region proposal network (RPN), followed by a downstream Fast R-CNN classifier. Therefore, the Faster R-CNN system is a purely CNN based method with no use of hand-crafted features (e.g., selective search that is based on low-level features). Using a single spatial pyramid pooling layer, end to end fine-tuning of a pre-trained ImageNet mode can be done using Fast R-CNN. This is the key to its better performance compared to the original R-CNN [14, 15].

Methods successfully implemented in other biometric areas can be applied in ear recognition, principle component analysis (PCA) is one. It is widely used in many biometric applications, ear recognition as a biometric is not different [16]. PCA is a method used to reduce the dimensionality, dataset suffers from variations by PCA feature vectors are represented in a space of low dimensionality. In particular, the distribution of the original feature vectors (the eigenvectors) is extracted using PCA from the covariance matrix. K. Chang [17] showed by comparing eigen-ears with eigen-faces that ear and face might have similar value for biometric recognition. Genetic algorithms represent an intelligent exploitation for solving both constrained and unconstrained optimization problems based on random searching. It is the way of solving problems by simulating the process of natural selection [18]. In the genetic programming, first, a randomly assemble computational programs are chosen by implementing a group of primitive operators, which are declared as the initial population. Then through sexual reproduction this population is allowed to progress with single or pair parents chosen stochastically while biased in their fitness on the task at hand. Over time the general fitness of the population tends to improve. In the end, the final solution is the obtained individual that achieves the best performance [19]. This paper is structured as follows: section 2 presents the architecture of the proposed system. section 3 discusses the experiments and results. Finally, conclusions are mentioned in section 4.

2. SYSTEM DESIGN

Any ear recognition system is in fact a pattern recognition system. Starting with the ear image as an input, converting it to a feature vector which after comparing with database a decision can be made to whom the input image is. There are many approaches to build an ear recognition system, which could be summarized by using five main stages: ear acquisition, ear detection, feature extraction, feature reduction, and a matcher.

2.1. Ear acquisition

The main target of any ear acquisition system is to capture the human ear image. Different approaches for ear acquisition are shown in Figure 1. Scanning the ear with a scanner is one, but this approach however makes the user uncomfortable when using the system. The most commonly used approach is taking a photo of the ear, where the photo is taken and combined with previously taken photos for identifying a person [20].

2.2. Ear detection

Ear detection is a process to detect the presence of an ear in the acquired image and then locate ear position in that image, it is a very important process, and weak ear detection stage may affect the successive stages and finally lead to system failure. For a robust ear recognition system against pose and blockage, an accurate and rapid detection of the ear is most critical. Two main categories used in this field [21]:

- Template matching: detect and locate the ear using a template of ear faster R-CNN or others.
- Shape-based: enroll the ear based on finding the elliptical shape of the ear by finding edges using a hough transform (HT).

Faster R-CNN is used here as ear detection and localization. It consists of a feature extraction network followed by two sub-networks, as shown in Figure 2. The feature extraction network is typically a pre-trained CNN such as Alexnet, Inception v3, or ResNet-50. The first sub-network is an RPN trained to produce object proposals (background). The second sub-network is trained to expect the actual class of each proposal (ear) [22].

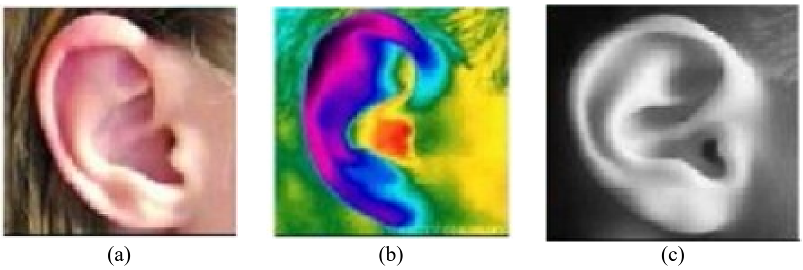


Figure 1. Different methods for ear acquisition; (a) photo an ear, (b) thermo gram of an ear, and (c) scan image of an ear

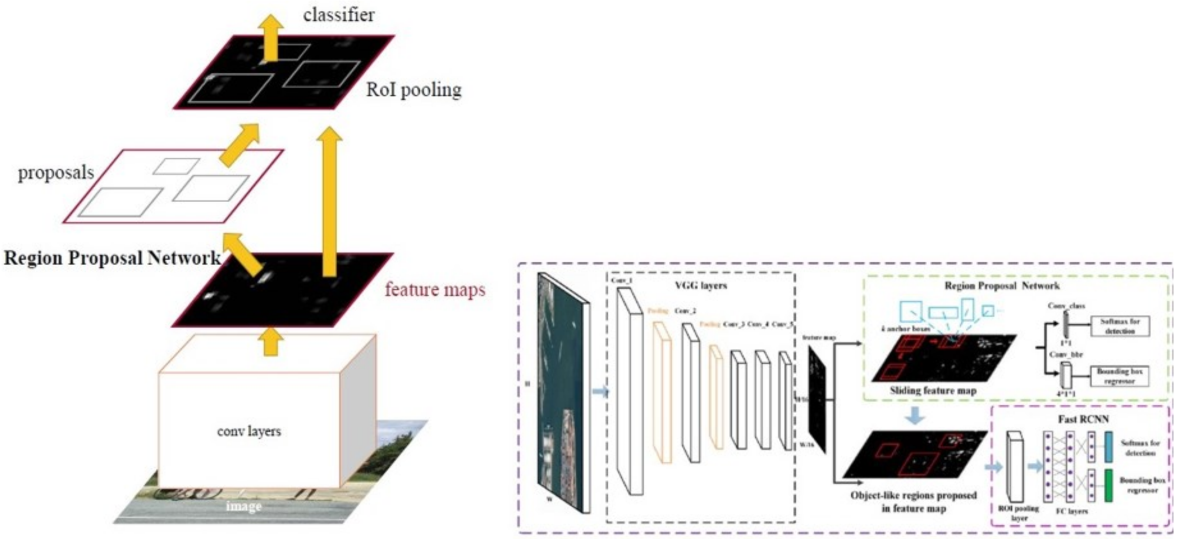


Figure 2. Faster R-CNN is a single, unified network for object detection

2.3. Feature extraction

Feature extraction is the most important phase of any biometric system; it supplies the system with the important characteristics of an ear that needs to be classified or to be matched with another ear. There are many approaches to find the feature vector of the ear such as [23]:

- Geometric features based techniques: such as Iannarelli's system, which is based on 12 measurements taken on ear photographs and Voronoi diagrams is modeled as an adjacency graph as illustrated in Figure 3.
- Appearance based techniques: which use either local or global ear image presence in recognition. Independent component analysis (ICA) and PCA fall under this category.
- Wavelet transformation: discrete Haar wavelet transform to extract the textural features of the ear.
- Force field transformation: force field based techniques purposed by D. J. Hurley [24], as shown in Figure 4.
- 3D ear shape: three-dimensional has more flexibility to problems that occur in two-dimensional data, such as illumination and pose.
- Deep learning: is considered a new/trend topic used in ear recognition, the deep neural network is supplied with an image of the ear, then useful features are automatically generated.

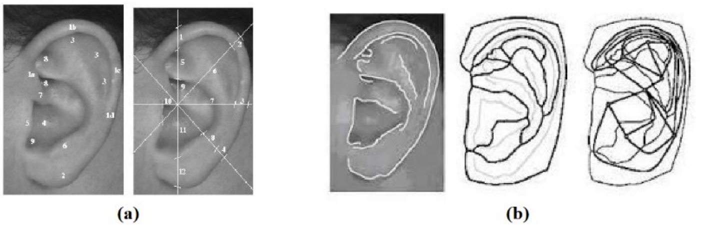


Figure 3. Geometric features based. (a) Iannarelli’s system, (b) Voronoi diagrams

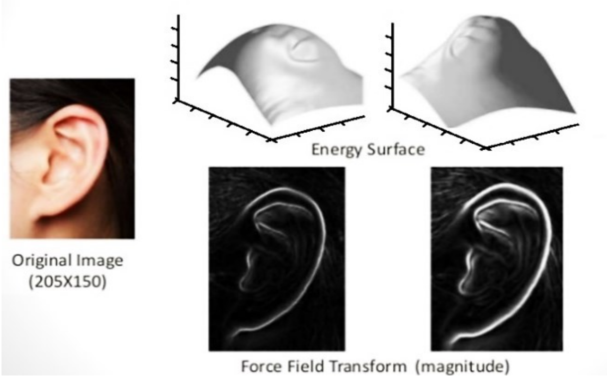
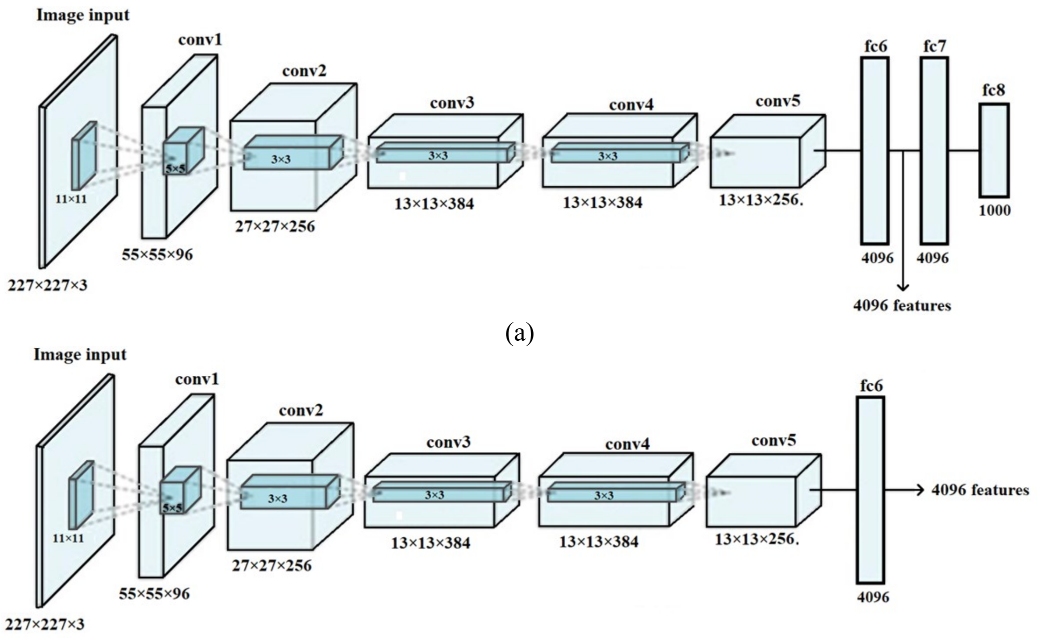


Figure 4. Ear features based on force field transformation

The power of using deep neural network to extract the features of ears, the features extracted automatically without a scripted rule [25]. The network supplied with appropriate images of ear and its targets by the times (training phase) the network will be learned how to extract the useful features. Another advantage of using deep neural network over traditional machine learning is that the performance in deep learning is increased as more amount of data are supplied for the training phase, but the same is not true in the tradition machine learning. The deep CNN is used here as feature extraction because as mentioned above deep learning choose automatically the appropriate features without human bias, here a sufficient dataset is used of ears of different persons, orientation, and constant in various environment, also the computational power with reasonable duration is available for this task, if the dataset of ears is limited (which is not) or using 3D modeling, conventional computer vision techniques may be better to be used [26].

Deep neural network requires a lot of data and computational processing power. Training a deep network could take several hours, days, or even weeks this is a huge problem to deal with. Fortunately, there are many pre-trained deep neural networks such as GoogleNet, Alexnet, VGG19, and many others, these pre-trained neural networks are trained with very large dataset up to million training images with 1000 categories. So instead of building a deep neural network from scratch, the pre-trained network can be used after making some modifications, the process of modifying a pre-trained network is called transfer learning. Alexnet is used in this system. Alexnet is a CNN that is trained on more than a million images from the ImageNet database. The network, as shown in Figure 5, is 8 layers deep and can classify images into 1000 object categories, such as pencil, mouse, and keyboard. As a result, the network has learned rich feature representations for a wide range of images. The image at the network input is of 227-by-227 size [27]. The transfer learning of Alexnet to fit into the ear recognition is done by supplying it with images of ears for different persons. Through the training phase, the network supplied by images of ears with a target person that belongs to him, so the network after several iterations and epochs learn how to extract useful feature to give the high percentage of correct classification (above 95%), after that, this network, as shown in Figure 6, is ready to e used as feature extraction to supply other stages of the system. Here the training dataset of 65 different persons, images of 10 persons ear are taken from our environment each person consider as one category, and 55 persons ear images are taken from AMI ear database [28] as single category (strangers/unknown), so the network classifies each image ear of 11 categories/classes,, gives 4096 features for each ear.



(a)

(b)

Figure 5. (a) Taking the features from the Alexnet CNN network, (b) The network used as a feature extractor

2.4. Feature reduction and selection

Mapping the feature vector from a higher into a lower dimension is feature reduction. Many tools could be used such as linear discriminant analysis (LDA), ICA, PCA, or genetic algorithm [27]. CNN gives 4096 features which is a huge number of features to be stored or processed in the other stages of the system, PCA is a statistical tool used to transfer the features from high dimensional correlated features to lower uncorrelated features. PCA here reduces the features from 4096 to 400 features. 400 features to be compared with others, is also considered large, it can be further reduced by excluding the unimportant features, but how the unimportant features can be chosen? So the feature selection stage is needed.

Genetic algorithm is used here as feature selection, it mimics the biological evolution in nature, there is a population of something, only the ones with good specifications will survive and produce offspring. Fitness function is used by the genetic algorithm used to evaluate each one of the population, make a crossover between two of the population by using a selection method, and make mutation; over time the new population will be closed to the optimum solution. Here in this system, each one in the population is represented by 400 binary variables, the fitness function is to give the most accuracy (of the matching classifier) with the least number of features used in. this stage reduces the feature to only 53 features. Genetic Algorithm is implemented, to calculate the similarity between two ears (ear1 and ear2) to figure out if it belong to the same person, as follows:

- Create randomly a binary vector(V) of size 400.
- Multiply the feature vector of each ear FE1_1 and FE2_1 by the binary vector(V), where FE1_1 and FE2_1 are vectors of 400 size (taken from PCA result for each ear) to produce another two vectors FE1_2 and FE2_2 for ear1 and ear2 respectively, as the following:
- Feature vector of ear1 FE1_2(i)= FE1_1(i) * V(i)
- Feature vector of ear2 FE2_2(i)= FE2_1(i) * V(i)
  Where i is an index from 1 to 400
- Merge the two vectors FE1_2 and FE2_2 to make a total vector (FET) of 800 length FET = [FE1_2 FE2_2]
- For all chosen two ears as a pair in dataset (train set used for the matcher) calculate the accuracy for the matcher, as following:

Matcher_result(j) = classify(FET(j),NET), j=1 to M

where where M is the number of ears pairs, classify is a function, gives 1 if the two ears belong to the same person else give -1, and NET is a trained fully connected neural network, then calculate accuracy= No. correct classified/total No. classified pair

- Calculate the number of active variables (VN) = ∑ v(i), where i=1 to 400, active variable equal to 1 if feature vector is used else 0.
- Create a minimize fitness function = [(1/Accuracy)*400]+VN, so the genetic algorthim will find the minimum number of features (VN) with maximum accuracy result by maniputing the binary vector (V).
- After many generations the mimimize fitness function reachs its mimuma point, by using the final binary vector (V), the number of active features is extracted and other features are ignored, in this research the number of active features is 53 selected out of 400.

2.5. Matching

In similarity matching the biometric system figures out if two instances (two set features) belong to the same person. Support vector machine (SVM), artificial neural network (ANN), and K-nearest-neighbors (KNN) could be used as a matcher. An ANN of two fully connected layers as shown in Figure 6, is used as ear matching, it has 106 input as a features vectors of two ears and give one output 1 for similarity or -1 for non-similarity. Figure 7 summarizes the proposed system of this research.

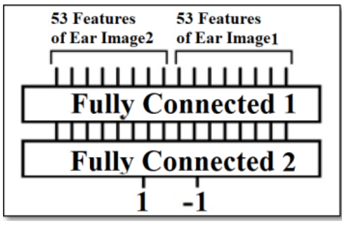


Figure 6. ANN as a matcher for two ears features

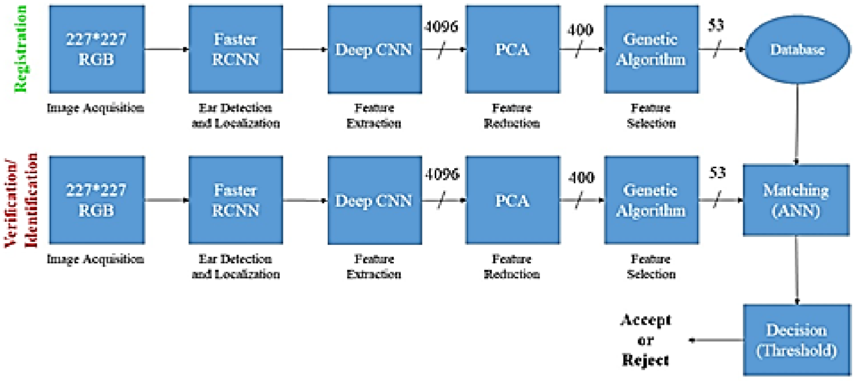


Figure 7. Block diagram of the proposed ear recognition system

3. IMPLEMENTATION AND TESTING

In the proposed system two datasets are used, Dataset1 for ear detection and localizing and dataset2 for feature extraction and matching. A sample of dataset1 shown in Figure 8 is used for training the faster R-CNN, it consists of 534 images with their annotation box ([x, y, width, hight]). 400 images for training and 134 for testing. The accuracy achieved is 97%. Dataset2 is used for training the CNN, it consists of ear images acquired from 65 persons, images of 10 persons are taken from our environment and 55 images are taken from AMI Ear Database as strangers (unknown).

Each person has 100 images, in the train set 60 images are taken for each class with 60 grayscales of that image so 120 for each class, total=120*11=1320. In validation set 20 images is taken for each class with 20 grayscales of that image, so 40 for each class, total=40*11=440. In the test set 20 images are taken for each class with 20 grayscales of that image, so 40 for each class, total=40*11=440. In MATLAB 2018b environment, the proposed system is implemented. The deep learning toolbox, computer vision system toolbox, optimization toolbox, and statistics and machine learning toolbox are used to build the program. All simulated experimentation used a personal computer with Core i7-8550U CPU 2 GHz, 16 GB DDR4 of memory, 2 GB GDDR5 (GeForce MX130) running under the windows 10 operating system. Table 1 shows the accuracy of classification for each set where CNN which later be used as feature extraction.

After the statistical test on the dataset, a sensitivity of 97.27% and specificity of 99.3% is achieved. Time for matching (database of 33 ear images features) requires 76 ms and 15 ms for feature extraction, ear detection, and localization requires 100 ms, the total time is equal to 191 ms so the system could be used in real-time application. Figure 9 shows some of the results in a real-time application, and the algorithm identified the persons in Figures 9 (a), 9 (b) and 9 (c) as the following: Omar, Mustafa and Mohammed with confidence 100%, 77%, and 99% respectively.

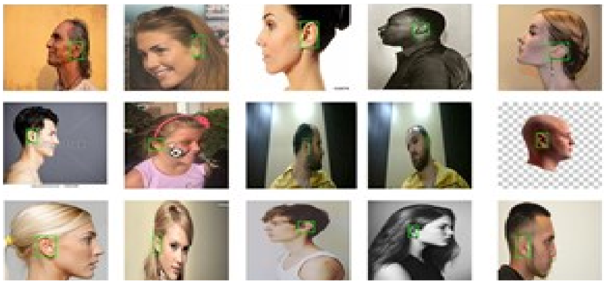


Figure 8. A sample of dataset1

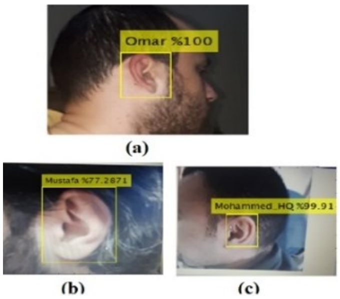


Figure 9. Results from the real time application

Table 1. The results of classification accuracy

Variable	Train set	Validation	Test set
CNN	100%	98.64%	97.73%
Matcher	100%	98.37%	95.82%

4. CONCLUSION

In this research, we have shown how to build an ear recognition system using deep learning. Using Faster R-CNN to increase the speed for ear detection and as a result, for the ear recognition system, CNN proved to be a powerful and accurate tool for feature extraction, PCA as a well-known dimensionality-reduction method succeeded in feature reduction, on the other hand, feature selection by genetic algorithm improved the quality of these features. In the end, by choosing ANN as a matcher a robust, high performance, fast ear recognition system is constructed. This combination led to an accurate recognition system that can be used in real-time applications with high accuracy.

REFERENCES

[1] A. Anwar, K. Ghany, H. Elmahdy, "Human Ear Recognition Using Geometrical Features Extraction," *Procedia Computer Science*, vol. 65, pp. 529-537, 2015.

[2] I. Omara, X. Wu, H. Zhang, Y. Du3, and W. Zuo, "Learning pairwise SVM on deep features for ear recognition," *IET Biometrics*, vol. 7, no. 6, pp. 341-346, 2017.

[3] Ž. Emerši˘c, A. Kumar S. V., B. S. Harish, W. Gutfeter, and J. N. Khiarak, "The unconstrained ear recognition challenge," in *Proceedings of the IEEE International Joint Conference on Biometrics*, Denver, CO, USA, October 1-4, 2017.

[4] E Hansley, M Segundo, S Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215-223, 2018.

[5] E. P. Ijjina and K. M. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," *Pattern Recognition*, vol. 59, pp. 199–212, 2016.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature,* vol. 521, pp. 436–444, 2015.

[7] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang, "Biometric Recognition Using Deep Learning: A Survey", *arXiv*. 2019.

[8] Amaia Salvador, Xavier Giro-i-Nieto, Ferran Marques, and Shin'ichi Satoh, "Faster r-cnn features for instance search," in Proceedings of the I*EEE Conference on Computer Vision and Pattern Recognition Workshops*, Las Vegas, NV, USA, June 26-July 1, 2016.

[9] Pedro Galdamez, William Raveane, and Angélica González, "A brief review of the ear recognition process using deep neural networks," *Journal of Applied Logic*, vol. 24, no. A, pp. 62-70, 2016.

[10] F Eyiokur, D Yaman, H Ekenel, "Domain adaptation for ear recognition using deep convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 199-206, 2018.

[11] W Raveane, P Galdámez, and M González Arrieta, "Ear Detection and Localization with Convolutional Neural Networks in Natural Images and Videos," *Processes*, vol. 7, no. 7, 457, 2019.

[12] M. Bizjak, P. Peer, and Ž. Emeršič, "Mask R-CNN for Ear Detection," 42nd *International Convention on Information and Communication Technology, Electronics and Microelectronics, (MIPRO)*, 2019.

[13] Rungruanganukul, M., and Siriborvornratanakul, "Deep learning-based gesture classification for hand physical therapy interactive program," *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Posture, Motion and Health, pp. 349-358, 2020.

[14] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection," in Proceedings of the *14th European Conference on Computer Vision*, Amsterdam, The Netherlands, October 11-14, 2016, pp. 443-457.

[15] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in Proceedings of the *12th IEEE International Conference on Automatic Face & Gesture Recognition*, Washington, DC, USA, May 30-June 3, 2017, pp. 650-657

[16] A. F. Abate, M. Nappi, D. Riccio, and S. Ricciardi, "Ear recognition by means of a rotation invariant descriptor," In Proceedings of the 18th *IEEE International Conference on Pattern Recognition*, Hong Kong, China, August 20-24, 2006, pp. 437-440.

[17] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160-1165, 2003

[18] L. Shao, L. Liu, and X. Li, "Feature Learning for Image Classification via Multiobjective Genetic Programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1359-1371, 2014.

[19] F. Mahmud, Md. Enamul Haque, S. Tauhid Zuhori, and B. Pal, "Human Face Recognition Using PCA based Genetic Algorithm," in Proceedings of the *International Conference on Electrical Engineering and Information and Communication Technology*, Dhaka, Bangladesh, October 9, 2014.

[20] N. K. Abdel Wahab, E. Essa Hemayed, and M. Bahaa Fayek, "HEARD: An automatic human EAR detection technique," in *Proceedings of the International Conference on Engineering and Technology*, Cairo, Egypt, October 10-11, 2012, pp. 1-7.

[21] Y. Zhang and Z. Mu, "Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks," *Symmetry*, vol. 9, no. 4, p. 53, 2017.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[23] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26-39, 2017.

[24] D.J. Hurley, M.S. Nixon, and J.N. Carter, "Automatic ear recognition by force field transformations," *IEE Colloquium on Visual Biometrics*, London, UK, March 2-2, 2000, pp. 1-5.

[25] BB. Benuwa, Y. Zhan, B. Ghansah1, D. Keddy Wornyo, and F. Banaseka Kataka, "A Review of Deep Machine Learning," *International Journal of Engineering Research in Africa*, vol. 24, pp. 124-136, 2016.

[26] O. Mahony, *et al.*, "Deep learning vs. traditional computer vision," In *Science and Information Conference*, Springer, 2019, pp. 128–144.

[27] L. Ghoualmi, A. Draa, and S. Chikhi, "An efficient feature selection scheme based on genetic algorithm for ear biometrics authentication," in Proceedings of the *12th International Symposium on Programming and Systems*, Algiers, Algeria, April 28-30, 2015.

[28] CTIM, "AMI Ear Database". [Online]. Available: http://ctim.ulpgc.es/research_works/ami_ear_database

BIOGRAPHIES OF AUTHORS

Ahmed Maamoon Alkababji, received the B.Sc., M.Sc. and PhD degree in Electrical Engineering from University of Mosul, Iraq in 1994. 1996 and 2007. Currently, he is assistant professor at computer engineering department in University of Mosul, Iraq.

Omar Hatif Mohammed, received the B.Sc. and M.Sc. degree in Computer Engineering Technology from Northern Technical University, Iraq in 2011 and 2016. His current research interests include pattern recognition and signal processing Currently he is studying PhD in computer engineering in University of Mosul, Iraq