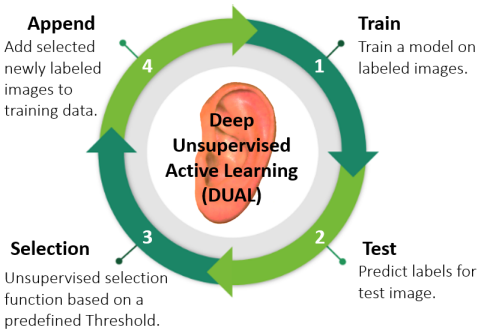
Ear Recognition Based on Deep Unsupervised Active Learning

Yacine Khaldi, Amir Benzaoui, Abdeldjalil Ouahabi, Sebastien Jacques, Abdelmalik Taleb-Ahmed

Ear Recognition Based on Deep Unsupervised Active Learning

Yacine Khaldi, Amir Benzaoui, Abdeldjalil Ouahabi, Sébastien Jacques, *Member, IEEE*, and Abdelmalik Taleb-Ahmed

Abstract —Cooperative machine learning has many appli-cations, such as data annotation, where an initial model trained with partially labeled data is used to predict labels for unseen data continuously. Predicted labels with a low confidence value are manually revised to allow the model to be retrained with the predicted and revised data. In this paper, we propose an alternative to this approach: an initial training process called Deep Unsupervised Active Learning. Using the proposed training scheme, a classification model can incrementally acquire new knowledge during the testing phase without manual guidance or correction of decision making. The training process consists of two stages: the first stage of supervised training using a classification model, and an unsupervised active learning stage during the test phase. The labels predicted during the test phase, with high confidence, are continuously used to extend the knowledge base of the model. To optimize the proposed method, the model must have a high initial recognition rate. To this end, we exploited the Visual Geometric Group (VGG16) pre-trained model applied to three datasets: Mathematical Image Analysis (AMI), University of Science and Technology Beijing (USTB2), and Annotated Web Ears (AWE). This approach achieved impressive performance that shows a significant improvement in the recognition rate of the USTB2 dataset by coloring its images using a Generative Adversarial Network (GAN). The obtained performances are interesting compared to the current methods: the recognition rates are 100.00%, 98.33%, and 51.25% for the USTB2, AMI, and AWE datasets, respectively.



Index Terms— Biometrics, ear recognition, active learning, GAN, USTB2 dataset, AMI dataset, AWE dataset.

I. Introduction

The human ear is considered a very recent modality in biometrics, and there is no commercial software based on this modality. It is considered one of the most stable human anatomical features. Compared to other modalities of the human body, especially the face, the ear does not change significantly throughout human life, whereas the face changes considerably with age. The characteristics of the face can be modified according to the cosmetic products used, the hairstyle, and the haircut. Besides, human faces change according to emotions and different facial expressions, such as sadness, joy, fear, or surprise.

Ear features are abundant, fixed, and unchanging with emotions. Unlike face identification systems, glasses, beards, or mustaches cannot obscure the ear images during the acquisition process. Most ear recognition studies have been conducted employing ear images taken in perfect conditions: ear images have ideal illumination, ears are in the exact location for each person, and ears are free of earrings, hair occlusions, or something that may obscure the ear. Actual techniques need to be developed to make ear identification practical; ear identification must prove its worth in an unconstrained context that reflects real-world circumstances. Furthermore, this technology must be portable to accommodate a substantial number of individuals and should be helpful for a wide range of people.

The supervised classification process for ear images is similar to that of the face or fingerprint recognition, or the techniques used in healthcare [1] or medical image clas-

Yacine Khaldi is with the LIMPAF Laboratory, Department of Computer Science, University of Bouira, Bouira 10000, Algeria (e-mail: y.khaldi@univ-bouira.dz).

Amir Benzaoui is with the Department of Electrical Engineering, University of Skikda, Skikda 21000, Algeria (e-mail: a.benzaoui@univ-bouira.dz).

Abdeldjalil Ouahabi is with the LIMPAF Laboratory, Department of Computer Science, University of Bouira, Bouira 10000, Algeria, and also with the iBrain, INSERM, UMR 1253, Université de Tours, 37200 Tours, France (e-mail: ouahabi@univ-tours.fr).

Sébastien Jacques is with the GREMAN UMR 7347, INSA Centre Val-de-Loire, CNRS, University of Tours, 37200 Tours, France (e-mail: sebastien.jacques@univ-tours.fr).

Abdelmalik Taleb-Ahmed is with the Université Polytechnique Hauts-de-France, 59313 Valenciennes, France, and also with the Centrale Lille, CNRS, UMR 8520-IEMN, Université de Lille, 59313 Valenciennes, France (e-mail: abdelmalik.taleb-ahmed@uphf.fr).

sification (e.g., COVID-19 detection/ classification from X-ray [2], [3] or CT [4] images). It is the process of predicting labels for input images using extracted features based on a predefined set of images/labels.

Traditionally, ear recognition has been performed using a classical machine-learning pipeline, i.e., training the model on a subset of labeled data, testing that model, and then deploying it in the real world. This technique has worked very well. Nevertheless, it is questionable whether it is possible to allow the model to acquire new knowledge by using additional images during the test phase. Such a property is desirable since it would be inspired by human cognition.

Cooperative Machine Learning (CML) [5] has been widely used to assist in decision-making, data annotation, etc. The general idea is to train an initial model on partially labeled data and use it to predict new labels for new data. Then, a human agent or corrector revises the low confidence predicted data and then retrains the model using the new predicted and corrected labels. This methodology has been used in other applications, such as accelerating the annotation of social signals [6], dynamic decision-making [7], etc.

In addition to model prediction, CML always relies on human intervention and correction, which means that we need to associate an observing human agent with the model to monitor and correct model behavior. From this point, we need to ask an important question: What if our model is accurate enough to be trusted to acquire new knowledge on its own (with a small margin of error) during the testing phase, without the help of a human agent.

This paper proposes to perform active unsupervised learning during the test phase of a trained ear recognition model. The role of the classification model is to predict the labels of test images and to classify them. Simultaneously, the role of the unsupervised active learning stage is to add some test images with their predicted labels (if the predictive confidence is larger than a predefined threshold) to the training dataset and perform additional training epochs. We call the proposed training scheme "Deep Unsupervised Active Learning" (DUAL).

We used three well-known ear datasets, namely the Mathematical Analysis of Images (AMI) dataset, the University of Science and Technology Beijing (USTB2) dataset, and the Annotated Web Ears (AWE) dataset to validate the proposed training approach. The USTB2 dataset contains grayscale images of ears. Therefore, we colorized all its images using the conditional Deep Convolutional Generative Adversarial Network (cDCGAN) [8], [9], with the resulting dataset named Colorized USTB2, or simply C-USTB2. For the classification model, we adopted VGG16 architecture. Many previous works have proven its performance. We then conducted extensive experiments to show the positive effect of using the advanced training technique over the conventional training/testing pipeline. In summary, this paper provides the following three main contributions:

1) Implementation of an original technique called Deep Unsupervised Active Learning (DUAL) in the field of ear recognition;
2) Importance of coloring ear images in grayscale instead of inputting them as is in biometric models;
3) Evaluation and discussion of the performance of the proposed technique using constrained and unconstrained ear datasets.

The rest of the paper is organized as follows. In Section 2, we briefly introduce previous works related to ear recognition. Then, Section 3 presents an overview of the related colorization technique. Section 4 presents our proposed training approach in detail. Finally, we conducted extensive experiments and comparisons, as explained in Section 5. Section 6 highlights the main findings of this work, as well as some research perspectives.

II. Related Work

Most biometric recognition systems based on 2D ear images consist of extracting features and comparing the extracted vector with the enrolled models. Based on this, ear recognition approaches can be classified into four different categories: holistic, geometric, local, and Deep Neural Network (DNN) approaches.

Holistic methods generate a set of basis vectors representing a subspace that respects the original set of images. In the set of basis vectors, each image of the ear can be reconstructed in the subspace. A hybrid classification system was proposed in [10] that integrates the ear shape and its algebraic features. The authors defined five coarse classes based on shape features and then applied Principal Component Analysis (PCA) or Independent Component Analysis (ICA) to extract the ear shape features for classification. Their results indicate that combining the proposed method with ICA yields promising results compared to PCA. In [11], the authors used a modular neural network to improve ear recognition performance. A 2D wavelet analysis with global thresholding was applied for image compression. The proposed system contains nine modules; each module was trained to recognize a part (helix, concha, or lobule) using a subset of training data. Another technique for automatic ear recognition that uses Haar wavelets to extract ear features and Fast Normalized Cross-correlation (FNCC) classifier was presented in [12]. In summary, holistic methods are widely used in biometric recognition systems. Nevertheless, they are sensitive to background changes and misalignment. Therefore, a small misalignment can lead to serious classification errors.

Geometric methods exploit the wealth of information contained in the geometric features of the ear, such as edge information and ear shape. In [13], the authors proposed a geometric approach where the feature vector was created by fusing the shape of the outer ear with the structural shape of the inner ear. In a recent study [14], the authors presented a geometric feature vector that considers the external helix's edge based on the minimum ear altitude line. Also, they took three measure-based characteristics to improve ear representation further. In [15], the authors used a multi-level fusion of the ear score, considering only its middle part. They extracted the outer and inner ear features in two successive steps and then fused the resulting scores before matching. Geometric methods appear to be simple to implement and of excellent

algorithmic complexity. However, their main drawback is their dependence on ear contours, which can be affected by noise or lightning.

Local methods are based on extracting features from different regions of the image, especially local orientation information, to perform biometric identification [16]. In [17], the authors combined morphological operators and Fourier descriptors to detect the ear zone automatically from gray-scale images. Secondly, they explored new ear feature extraction techniques using complex Gabor filters to extract local phase and localized orientation information and a local phase encoding by employing log-Gabor filters. In [18], the grayscale mapping technique was used to improve the contrast of ear images. The Scale-Invariant Feature Transform (SIFT) was applied as a local feature extractor, while Euclidean distance was adopted for classification. In work published in [19], the authors presented the complete version of the challenging AWE dataset. They conducted an extensive experimental analysis by testing and comparing the performance of eight local descriptors on the AWE dataset. In another work [20], the authors adopted a tunable filter bank based on a half-band polynomial of order 14 as a local feature extractor. For the matching phase, L1, L2, Cosine similarity, and Canberra distance were used to compute the distance between feature vectors. In [21], the authors presented a comparative experimental analysis on ear recognition by adopting several recent variants of the LBP descriptor. Moreover, the authors proposed another version of the LBP descriptor, called Averaged Local Binary Patterns (ALBP). They showed that the LBP operator and its variants are suitable only for ear recognition under controlled conditions, but they suffer greatly in unconstrained cases.

Deep Neural Networks (DNNs) have recently gained increased interest in pattern recognition and computer vision [22]–[24]. Various DNN models and architectures have been proposed in the literature; the most well-known are AlexNet [25], VGGNet [26], Inception [27], VGG-face [28], and SqueezeNet [29]. These models have achieved remarkable performance in several unconstrained biometric applications, including face recognition [30]. However, a significant body of work on ear biometrics uses deep learning in the literature. In [31], a comparative study between handcrafted and CNN-based models was conducted. The authors conducted a series of comparative experiments using seven handcrafted feature extractors versus four CNN-based models. Several researchers have introduced sets of deep learning models. In [32], a model set consisting of different configurations of VGG architectures was proposed. Image features are extracted using multiple VGG configurations and then averaged before being fed to a fully connected layer to predict a label. A six-layer deep CNN-based architecture for ear recognition was proposed in [33]; the authors used a constrained ear dataset to validate the performance of the proposed model. In [34], Handcrafted and learned features have been fused in the proposed two-stage framework, where a CNN-based model has been employed for landmark detection, followed by features extraction, learned and handcrafted, and finally score normalization and fusion. In a recent study [9], a generative adversarial network was used to colorize grayscale ear images to improve ear recognition performance. In the study conducted in [35], the authors tested the performance of a finely tuned VGG-face model for ear recognition. The VGG-face has the same architecture as VGG-16, except that it was pre-trained using face images. In [36], data limitation in ear recognition was addressed using methods based on few-shot learning. The authors used data augmentation to overcome the problem.

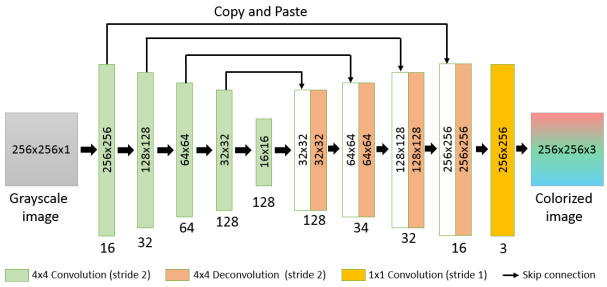


Fig. 1. U-Net architecture of the generative model.

Table I summarizes the works discussed in this section by categories, datasets used, and experimental protocols. To date, ear recognition systems are limited to feature extraction and then testing. No additional learning is provided after the training phase [35]–[37]. Researchers have exploited local, holistic, geometric, deep, and hybrid features. For classification tasks, they used statistical and sparse representation classifiers [38], as well as neural networks [39].

Until now, ear recognition systems lie within the scope of training then testing. No further learning is expected after the training phase. This limitation prohibits the model from exploiting any possible information during the test phase. We propose a method to overcome this specific limitation by employing active learning to expand the initial model knowledge base.

III. Colorization Using Conditional Deep Convolutional GAN

Recently, Generative Adversarial Networks have been used to tackle different challenges in ear biometrics. In [39], the authors proposed synthesizing the region of interest out of the ear image using image-to-image translation instead of feeding ear images as they are to the classifier. They trained a Pix2Pix generative adversarial network to generate a synthetic segmentation of the ear free of occlusions, hair, and non-ear pixels. Furthermore, missing parts of the ear due to different occlusions can also be synthesized using the proposed technique. Following our previous research [9], we asserted that feeding color ear images to a deep classifier yields higher performance. We used conditional deep convolutional generative adversarial networks (cDCGAN) to colorize gray-scale ear images and enhance the recognition rate. Based on this idea, we designed a custom U-Net architecture for our generative model to synthesize colorized ear images, as illustrated in Fig. 1.

The discriminative model architecture is straighter than that of the generator, as shown in Fig. 2. It is composed of five convolutional layers. The first four layers use 4×4 filters to

TABLE I
COMPREHENSIVE SUMMARY OF RELATED WORKS STUDIED

Approach	Publication	Method	Employed dataset			Evaluation Protocol
			Name	#Sub.	#Img.	
Holistic	Zhang and Mu [10]	Geometric Features + ICA	USTB-1	60	180	2 img/sub Train & 1 remaining Test
			USTB-2	77	308	3 img/sub Train & 1 remaining Test
			Private	17	102	5 img/sub Train & 1 remaining Test
	Gutierrez et al. [11]	Wavelet Transform	USTB-2	77	308	3 img/sub Train & remaining 1 Test
	Anjum et al. [12]	Haar Wavelets + FNCC	USTB-1	60	180	120 Train & 65 Test
			USTB-2	77	308	3 img/sub Train & 1 remaining Test
			IITD-1	125	493	250 Train & 243 Test (Rand)
Geometric	Mu et al. [13]	Geometrical Measures	USTB-2	77	308	3 img/sub Train & 1 remaining Test
	Lakshmanan [15]	Multi-Level Fusion	USTB-2	77	308	3 img/sub Train & 1 remaining Test
	Omara et al. [14]	Geometric measurements	UUSTB-1	60	180	60 Train & 120 Test (Rand, 10 times)
			IITD-1	125	375	125 Train & 250 Test (Rand, 10 times) (Rand)
Local	Kumar and Wu [17]	Orthogonal log-Gabor filter pair	IITD-1	125	493	250 Train & 243 Test
			IITD-2	221	793	442 Train & 351 Test
	Ghoualmi et al. [18]	SIFT	IITD-1	125	421	3 img/sub Train & remaining Test
			USTB-1	180	60	
			USTB-2	77	308	
	Chowdhury et al. [20]	Tunable Filter Bank	AMI	100	700	60% Train & 40% Test (Rand)
			IITD-1	125	493	2 img Train & remaining Test (Rand)
			UERC-17	3540	11804	2304 Train & 9500 Test
	Hassaballah et al. [21]	Completed LBP	IITD-1	125	493	2 img Train & remaining Test (Rand)
			IITD-2	221	793	
			AMI	100	700	60% Train & 40% Test (Rand)
			AWE	100	1000	60% Train & 40% Test (Rand)
Deep Neural Networks	Alshazly et al. [31]	AlexNet (Fine Tuning)	AMI	100	700	60% Train & 40% Test (Rand)
			CVLE	16	804	
	Alshazly et al. [32]	VGG	AMI	100	700	60% Train & 40% Test (Rand)
			WPUT	474	3348	
	Priyadharshini et al. [33]	CNN	IITD-2	221	793	490 img Train & 303 Test (Rand)
			AMI	100	700	600 img Train & 100 Test
	Hansley et al. [34]	CNN + HOG	UERC-17	3540	11804	2304 Train & 9500 Test
	Zhang et al. [35]	VGG-face	AWE	100	1000	60% Train & 40% Test
	Khaldi and Benzaoui [9]	DCGAN + VGG16	AMI	100	700	60% Train & 40% Test
			AWE	100	1000	
	Zhang et al. [36]	MAML + CNN	AMI	100	700	60% Train & 40% Test (Rand)
			UERC-17	3540	11804	2304 Train & 9500 Test

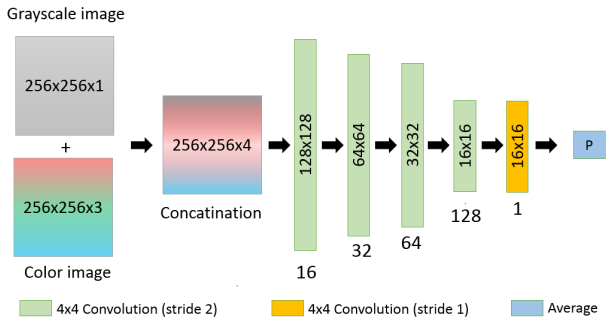


Fig. 2. Discriminative model architecture.

convolve with, sliding by two steps. Each layer is followed by batch normalization and the Leaky-ReLu activation function. The last convolutional layer has only one 4 × 4 filter, with a stride of one and activated by the Sigmoid function, to output a single scalar indicating whether an input color image is real or false. This binary classification model predicts a probability output in the range [0, 1]. In our case, it predicts the probability that an image is real or fake. This probability, called P in Fig. 2, is calculated by averaging the output patch of size 16 × 16.

The discriminator takes as input an array of 256 × 256 × 4, which is the concatenation of a gray-scale image and a real or generated color image. It is trained to maximize the probability of identifying generated color images out of real images $\log D(y|x)$. At the same time, the generator is trained simultaneously to minimize $1 - \log D(G(0_z|x))$. The final cost function V can be expressed mathematically as shown in (1) [8].

$$min_G max_D V(G,D) = \mathbb{E}_x \left[\log D(y|x))\right] + \mathbb{E}_z \left[1 - \log D(G(0_z|x))\right] \quad (1)$$

where x is the grayscale image, y is the ground truth, $G(0_z|x)$ is the mapping function representing the generator output color image of the input image x. Similarly, the discriminator is represented by the mapping function $D(y|x)$ that produces a scalar between [0,1], indicating the probability of the input being generated or not. $\mathbb{E}_x$ is the expected value over all real color images, $\mathbb{E}_z$ is the expected value over all generated color images.

$$Distance(x,\theta) = \frac{1}{3nm} \sum_{c=1}^{3} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| x_{i,j}^{c} - y_{i,j}^{c} \right\|_2^2 \quad (2)$$

As shown in (2), we aim to train the adversarial model to minimize the average Euclidean Distance on a pixel level between colorized images and ground truth images, where x is the grayscale image, y is the ground truth, θ is the corresponding image colorized by the generative model, c is the channel index, i and j are the pixel indices of the image.

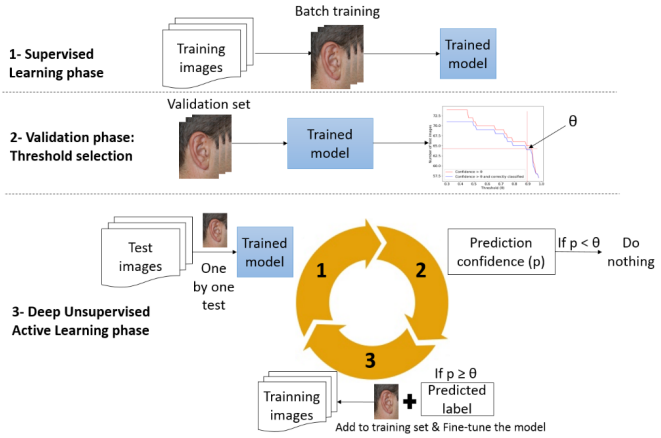


Fig. 3. Proposed Deep Unsupervised Active Learning (DUAL) scheme.

IV. Proposed Deep Unsupervised Active Learning Workfolw

The proposed DUAL training scheme consists of three consecutive training phases: a supervised training phase, a validation and hyper-parameter fine-tuning phase, and an unsupervised active learning phase. In the first phase, we performed supervised training of the classification model, i.e., using a labeled training dataset. Then, we performed a validation experiment using a limited validation set to determine the best value of the hyper-parameter θ. Finally, we performed unsupervised active learning using the test images during the test phase. Therefore, the unsupervised active learning phase is independent of the initial training dataset, i.e., when deploying a biometric model, it should be trained using only the initial labeled dataset. Then, an unsupervised active learning phase is performed using real-time test images.

During the standard testing phase, the classification model cannot obtain additional learning from the test images (i.e., the recognition rate of the model will not improve), even if it is a high recognition rate model. Therefore, we propose an alternative testing technique where a model can gain additional knowledge during the classification of test images using unsupervised active learning. We refer to this testing phase as the unsupervised active learning test phase. Using the unsupervised active learning during the test phase, images that were classified with above-threshold confidence are included in the initial training dataset before performing additional training epochs. We go through test images one by one, as shown in Fig. 3.

The classification model we adopted is based on VGG16 architecture. Improving the classification model itself is not our concern in this work. Thus, we did not consider other CNN-based architectures such as VGG19, ResNet, etc. It is sufficient to use a classification model architecture with a high recognition rate to validate the proposed method. We added a fully connected layer and a softmax output layer on top of the convolutional layers of the VGG16 model, which is pre-trained on the ImageNet dataset [40]. Table II details the global structure of the classification model. We used categorical cross-entropy as a loss function to measure the performance of our model during the training phase. The cross-entropy can be calculated using (3), where M is

TABLE II
Details of the Parameters of the Classification Model Architecture

Layer	Neurons number	Activation function	Drop
VGG16 convolutional layers			
Dense layer	1024	ReLu	50.00%
Output layer	Nbr of persons	Softmax	-

the number of classes (individuals) for each dataset, y is the binary indicator vector if label c is the correct classification for image o, and p is the vector of the predicted probability that image o belongs to class c. Cross-entropy is the simplest and most widely used cost function because it follows directly from the definition of entropy. On the other hand, we used Adam's well-known optimizer [41] to update the model weights based on the training data.

$$CrossEntropy(p, y) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3}$$

V. Experimental Analysis

To evaluate the performance of our framework, we performed a series of experiments using a set of ear images from the USTB2, AMI, and AWE ear datasets. All datasets, along with the evaluation protocol, are presented in the following sections of this paper.

A. Experimental Data

The benchmarks AMI, USTB2, and AWE ear image datasets were used in the experiments.

1) The AMI Ear Dataset: The Mathematical Analysis of Images (AMI) ear dataset [42] was created by collecting uncropped ear images from 100 subjects, seven images per person, in an indoor environment. These images are in jpg format with a resolution of 492 × 702 pixels. Each subject has seven images, six from the right ear and one image from the left ear. As shown in Fig. 4 (a), the AMI dataset already contains colored images, and we did not need to preprocess or crop the images.

2) The USTB2 Ear Dataset: The University of Science and Technology in Beijing (USTB) [43] collected four ear images under different lighting conditions for 77 subjects (students and teachers). The entire dataset contains 308 uncropped images. The first image is the frontal image of the ear under standard illumination, the second and third images are captured with +30 and −30 degree rotations, respectively, and the fourth image is taken under low illumination conditions. Fig. 4 (b) shows that the USTB2 dataset contains grayscale images with different illumination conditions. As a preprocessing step, we colorized these images using cDCGAN, as explained in Section 3, to obtain what we called Colored USTB2, or simply C-USTB2.

3) The AWE Ear Dataset: The Annotated Web Ears (AWE) [19], an unconstrained ear dataset, is considered one of the most challenging ear datasets. It contains 1000 cropped images of the left and right ear with different head rotations, genders, races, occlusions, and lighting conditions. The dataset was collected from the Internet for 100 different public figures. These images vary in size from 473 × 1022 to 15 × 29.

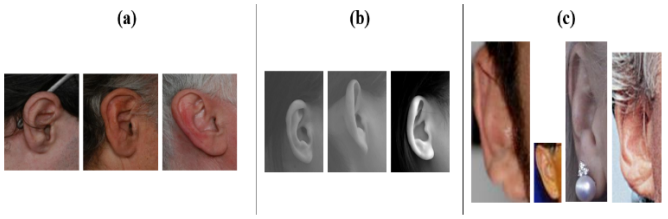


Fig. 4. Sample ear images from (a) the AMI dataset, (b) the USTB2 dataset, and (c) the AWE dataset.

Fig. 4 (c) shows representative images of different subjects from this dataset.

B. Setup

Before proceeding with the validation of the proposed learning approach and the different test phases, we colorized the USTB2 images to improve the recognition rate (RR), which is defined as the total number of correctly identified probe images divided by the total number of probe images, as shown in (4).

$$RR = \frac{Nbr\ of\ correctly\ identified\ probe\ images}{Total\ Nbr\ of\ probe\ images} \quad (4)$$

We generated a new colorized dataset called C-USTB2 using a cDCGAN model. Since we used a VGG16-based model pre-trained on ImageNet with color images, this colorization step increases the recognition rate, unlike grayscale images. To train the cDCGAN colorization model, we used the colored images from the AMI dataset, we can use any other ear dataset with colored images, and the labels do not matter because we are interested in the colors, not the labels. The model implicitly generates a corresponding grayscale image for each colored image and then generates a colorized image. To measure its performance, we computed the accuracy, and it is defined as the ratio of the number of correctly colorized pixels to the total number of pixels. A pixel is correctly colored if the difference between its Red, Green, Blue (RGB) values and the original pixel values is below a certain threshold. More precisely, the accuracy is defined mathematically in (5) [44], where x is the colorized image, y is the corresponding ground truth image, $1_{[0,\epsilon_c]}$ is the indicator function, n and m are the image dimensions, i and j are the pixel indices of the image, c is the color channel, and ϵ_c is the channel threshold.

$$Accuracy(x,y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\prod_{c=1}^{3} 1_{[0,\epsilon_c]}|x_{i,j}^{c} - y_{i,j}^{c}| \quad (5)$$

We trained the model for a total of 62 epochs (over 1000 mini-batch iterations) to achieve a distance loss of 2.41 for the generator and 1.38 for the discriminator, and a training colorization accuracy of 79.73%, as shown in Fig. 5. The resulting colorization accuracy is the highest we could achieve using only the color images from the AMI dataset. Although the result is very acceptable, we believe it can be improved using a larger set of multi-color ear datasets. Fig. 6 shows the images of C-USTB2. We can see that not only did we obtain a colorized version of USTB2, but we also successfully equalized the brightness and intensity of the images using cDCGAN.

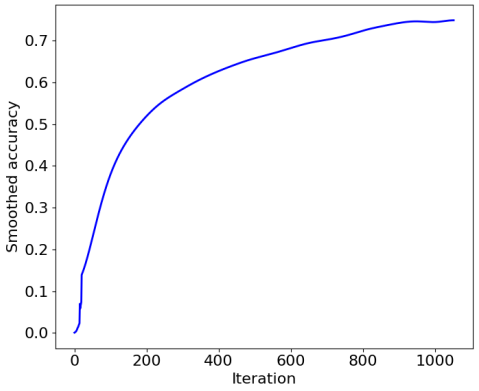


Fig. 5. Accuracy of the training colorization of the cDCGAN model.

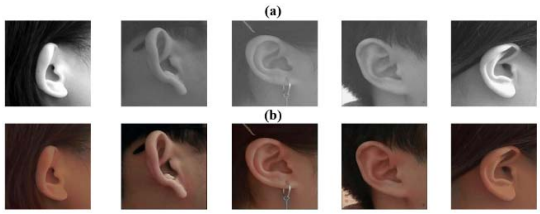


Fig. 6. Coloring of USTB2 images with cDCGAN (C-USTB2), (a) original images of USTB2, (b) colored images of C-USTB2.

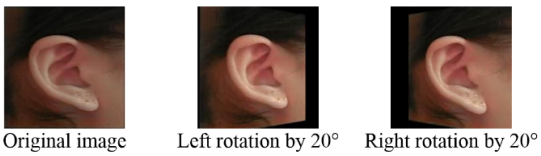


Fig. 7. Augmentation of the training dataset.

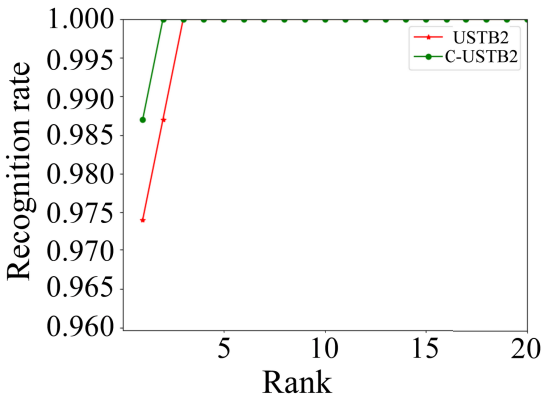


Fig. 8. Cumulative matching characteristic curves for USTB2 and C-USTB2 datasets.

As shown in Fig. 7, we increased the training set by generating two additional images for each training image. One image was rotated 20° to the left, and the other was rotated 20° to the right. We used 60% of the images for training and 40% for testing in the case of AMI. For C-USTB2, we used three images for training and one image for testing. The train/test distribution of the AWE datasets is predefined by their owners, 600 images for training and 400 images for the test.

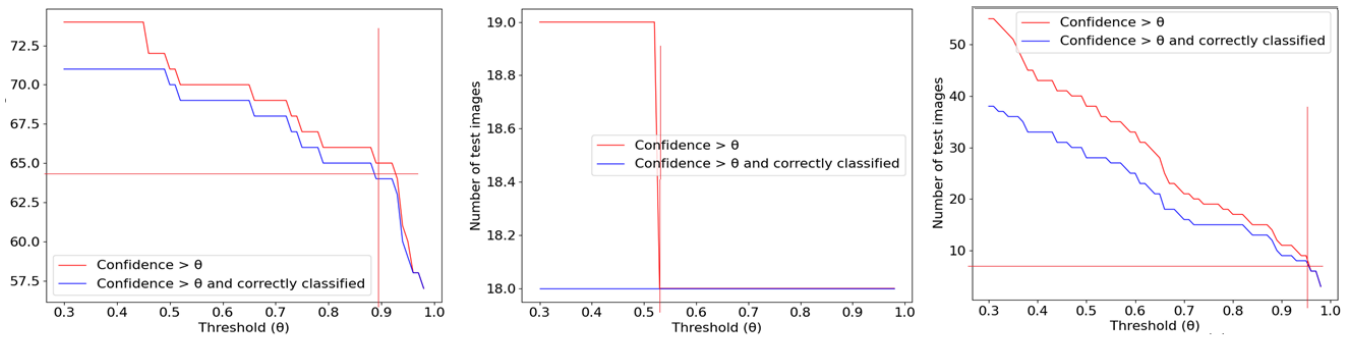


Fig. 9. Number of images correctly identified with a given θ for (a) the AMI dataset, (b) the C-USTB2 dataset, and (c) the AWE dataset.

C. Experiment #1

To highlight the positive effect of coloring the USTB2 dataset on the recognition rate, we used the VGG-based pre-trained classification model with the USTB2 and C-USTB2 datasets with the same configuration. The recognition rate we obtained using the C-USTB2 dataset is 98.70%, while it was 97.40% using the USTB2 grayscale dataset. Fig. 8 shows the Cumulative Matching Characteristic (CMC) curves for both scenarios. As expected, coloring the USTB2 dataset improved the recognition rate; this may be due to a VGG model pre-trained on color images (ImageNet), so it is best to also use color images in the following steps.

TABLE III
RESULTS OF THE SUPERVISED LEARNING AND THE DUAL SCHEME FOR AMI, C-USTB2, AND AWE DATASETS

Method	AMI Rank-1	C-USTB2 Rank-1	AWE Rank-1
Supervised Learning	96.00%	98.70%	49.25%
Deep Unsupervised Active Learning	98.33%	100.00%	51.25%

TABLE IV
SUPERVISED-LEARNING TEST STATISTICS

Test images	AMI	C-USTB2	AWE
Nbr of test images with confidence $\geqslant \theta$	132	57	19
Nbr of test images with confidence $\geqslant \theta$ and correctly classified	131	57	17

D. Experiment #2

In this experiment, we tested and evaluated the effectiveness and impact of the DUAL scheme. The confidence threshold θ is a hyper-parameter that is chosen to trigger the fine-tuning process. The best value can be easily selected by performing a single supervised learning test phase using a validation set and monitoring the number of correctly classified images with a confidence value higher than a defined θ. For that, we used a quarter of the test set from each dataset as a validation set. Visualizing the data gives a clearer perspective to decide. Fig. 9 shows the relation between θ values and the number of correctly identified test images with confidence greater than or equal to θ. We used the best estimation to be located where the vertical difference between the two curves is minimal, considering that we want to maximize the number of images with confidence greater than θ.

The estimation of the best value of the variable will differ depending on several criteria, including the type of dataset used, constrained or unconstrained, the type of classification models itself, and its basic recognition rate. In our experiment, we chose 0.89, 0.52, and 0.95 as threshold values for the AMI, C-USTB2, and AWE datasets, respectively.

From Table III, it is clear that the proposed DUAL approach has significantly improved the model recognition rate for all datasets; DUAL has a higher recognition rate than supervised learning. For the AMI dataset, the recognition rate increased from 96.00% to 98.33%. For the C-USTB2 dataset, the recognition rate increased to 100.00% using the DUAL scheme. The same is true for the AWE dataset, where the recognition rate increased by 2%. At this point, and unlike the traditional supervised classification process, the proposed DUAL scheme allowed the model to gain additional information beyond that of the initial training phase. The accuracy varies between experiments due to the datasets used, not the model itself. In the general case, constrained datasets give higher accuracy, unlike unconstrained datasets. The same goes for our study, we obtained less accuracy using unconstrained datasets, but this does not affect the study's main objective.

Table IV shows the number of test images (i.e., which the model attempts to learn during the test phase) correctly classified with confidence greater than or equal to θ. Those test images with their predicted labels improved the recognition rate by using them to perform extra fine-tuning epochs. The model re-trained itself using new data for the AMI dataset: 131 new images with correctly assigned labels. In the second scenario using the C-USTB2 dataset, the model predicted correct labels for 57 images with confidence greater than θ. Hence, it re-trained itself with entirely correct information during the test phase. For the AWE dataset, the DUAL scheme using the selected threshold was able to actively train the model with 19 new images, in which 17 images were correctly identified. Nevertheless, although 10.52% of the new data was misclassified, the recognition rate was augmented due to the higher amount of correctly classified data used for active learning.

Fig. 10 shows the cumulative match characteristic (CMC) curves of the supervised learning and the DUAL learning scheme.

We measured the number of correctly recognized images during the DUAL test phase compared to the supervised learning test phase. It is evident in Fig. 11 that the model was improved using particular test images and the

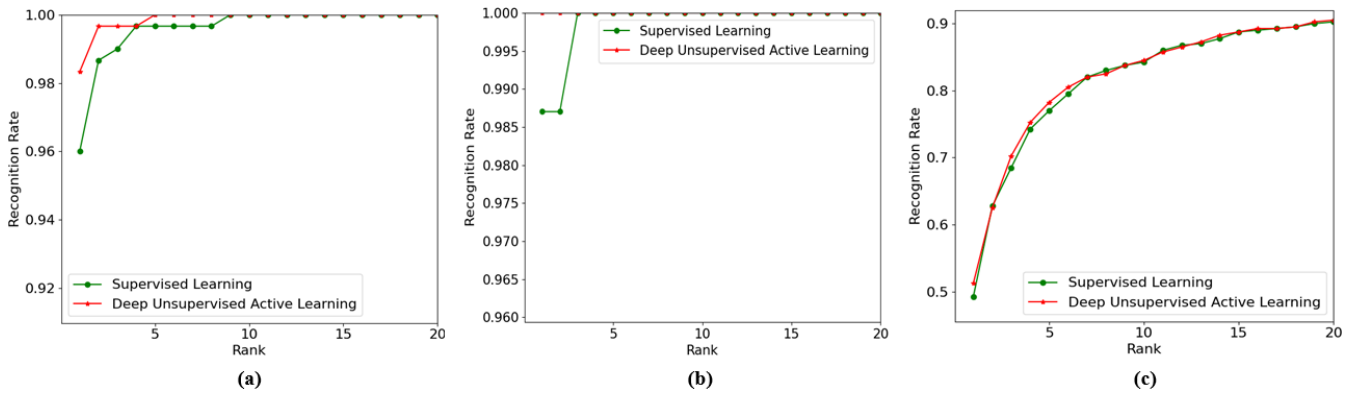


Fig. 10. Cumulative matching characteristic curves for (a) the AMI dataset, (b) the C-USTB2 dataset, and (c) the AWE dataset.

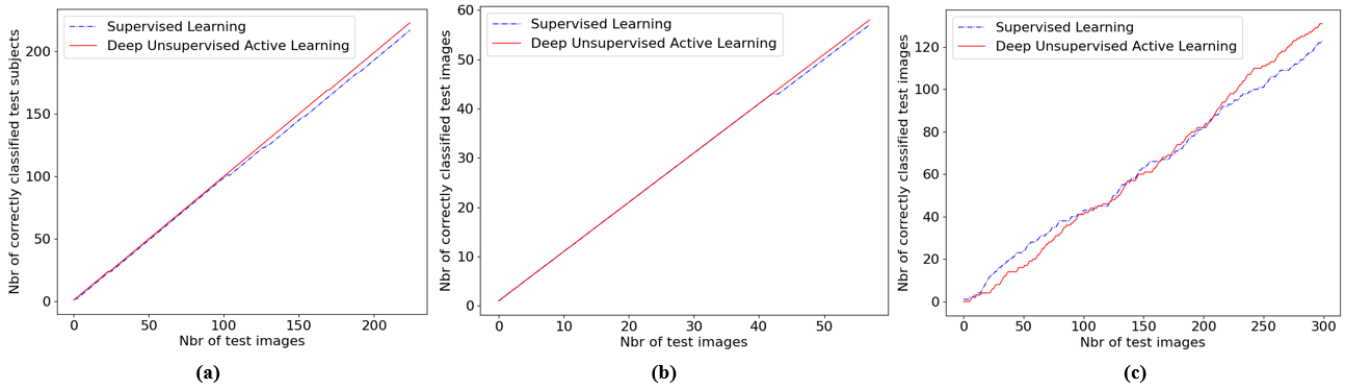


Fig. 11. Number of correctly predicted labels during the test phase for (a) the AMI database, (b) the C-USTB2 database, and (c) the AWE dataset.

TABLE V
A COMPARISON OF RANK-1 OF THE PROPOSED APPROACH WITH OTHER REPRESENTATIVE METHODS

Category	Publication	Year	Method	USTB2	AMI	AWE
Geometric Methods	Mu et al. [13]	2004	Geometrical Measures on Edge Images	85.00	–	–
	Lakshmanan [15]	2013	Multi-Level Fusion	99.20	–	–
Holistic Methods	Zhang and Mu [16]	2008	ICA	92.20	–	–
	Gutierrez et al. [11]	2010	Wavelet Transform	97.50	–	–
	Anjum et al. [12]	2011	Haar Wavelets + Fast Normalized Cross Correlation	96.10	–	–
Local Methods	Ghoualmi et al. [18]	2016	SIFT	94.79	–	–
	Emeršič et al. [19]	2017	POEM	–	–	49.60
	Chowdhury et al. [20]	2018	Tunable Filter Bank	–	70.14	–
	Hassaballah et al. [21]	2019	Completed LBP	–	73.71	49.60
Deep Learning Methods	Zhang et al. [35]	2018	VGG-face	–	–	50.00
	Alshazly et al. [32]	2019	VGG-13-16-19 ensemble	–	97.50	–
	Alshazly et al. [31]	2019	AlexNet (Fine-tuning)	–	94.50	–
	Zhang et al. [36]	2019	MAML + CNN	–	93.96	–
	Khaldi and Benzaoui [9]	2020	DCGAN + VGG16	–	96.00	50.53
	Priyadharshini et al. [33]	2020	CNN	–	96.99	–
	Our proposed method	**2021**	**VGG16 + DUAL**	**100.00**	**98.33**	**51.25**

corresponding correctly predicted labels by the DUAL scheme to gain additional knowledge. This process positively affected the recognition possibility of the rest of the test images. For the AMI dataset, retraining the model around test image number 100 has improved the likelihood of correctly predicting the rest of the images. Likewise, the DUAL scheme in the C-USTB2 and AWE scenarios, as of test image number 42 and 230, respectively, improved the overall recognition rate for the remainder of the test images.

E. Comparison of Rank-1 Recognition Rate

Table V compares the Rank-1 recognition rate between the proposed training approach and the recent and well-known approaches that used the AMI, USTB2, or AWE datasets. From Table V, the DUAL scheme showed the best results compared to the state-of-the-art methods by pushing the model performance to its limits. Although all approaches have advantages and disadvantages, in this work, we focused on the important advantage of the proposed approach over the others, namely the possibility of gaining new knowledge during the testing phase instead of relying only on what was learned during the learning phase. From our point of view, this is a very important feature to be acquired by artificial intelligence systems in general.

Although it should be kept in mind that the proposed DUAL scheme requires more processing time and memory space,

performing continuous active learning during the testing phase may be difficult, especially in real-time. Future research can find better and faster ways for the model to acquire information in new training images, as in active learning, in less time and without losing or negatively affecting the acquired knowledge. On the other hand, the presence of the initial training data is mandatory for the proposed technique. However, these difficulties could be overcome in the future through more research.

VI. Conclusion

This study aims to determine the feasibility of active learning in the field of ear recognition and biometrics in general. We have proposed a machine learning technique called Deep Unsupervised Active Learning (DUAL), by which a biometric model can acquire new knowledge continuously after the training phase. Based on this, a biometric model uses the test images that have been classified with a confidence value above a pre-defined threshold to perform additional learning epochs. We then validated this property by conducting in-depth experiments using the constrained AMI and C-USTB2 ear datasets and the unconstrained AWE dataset to measure the recognition rate under supervised and DUAL learning. The Rank 1 recognition rates are 100.00%, 98.33%, and 51.25% for the C-USTB2, AMI, and the challenging AWE datasets. The proposed method combines the power of supervised and unsupervised learning. It takes the prior gained knowledge and expands it using unsupervised learning. This hybrid property makes our proposed model more efficient than supervised or unsupervised models.

These preliminary results lead to the following conclusions:

1) Test images contain a significant amount of information, and they are left untapped by classification models;
2) Image classification models can be improved beyond the training phase;
3) The proposed DUAL scheme can be used to train the model during the test phase and enhance its performance;
4) When the performance of a method is presented, it is essential to identify the dataset used correctly [45];
5) The fact that the AWE dataset is "noisy" results in relatively poor performance. Hence the need for image preprocessing, such as denoising the background noise [39] and texture data [47], [48] by first and second-generation wavelets [49], [50] and multiresolution analysis [51].

Under conditions where no pre-treatment was performed, we compared our training approach to new cutting-edge models; the result is particularly instructive: the proposed technique significantly improves model performance. Although the results obtained are mainly satisfactory, potential improvements can be made: image preprocessing by brightness corrections, denoising, super-resolution and geometrical transformation, fine-tuning parameters such as learning rate and the number of convolutional layers and filters, and the use of a variable threshold θ, etc.

We are convinced that these results will significantly impact image classification, biometrics, and ear recognition applications such as security, forensics, access control, and medical applications.

References

[1] L. P. Etter, E. J. Ragan, R. Campion, D. Martinez, and C. J. Gill, "Ear biometrics for patient identification in global health: A field study to test the effectiveness of an image stabilization device in improving identification accuracy," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–9, Dec. 2019.

[2] V. Madaan *et al.*, "XCOVNet: Chest X-ray image classification for COVID-19 early detection using convolutional neural networks," *New Gener. Comput.*, pp. 1–15, Feb. 2021, doi: 10.1007/s00354-021-00121-7.

[3] S. Liang *et al.*, "Fast automated detection of COVID-19 from medical images using convolutional neural networks," *Commun. Biol.*, vol. 4, no. 1, pp. 1–13, Dec. 2021.

[4] P. Silva *et al.*, "COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100427.

[5] M. Dong and Z. Sun, "On human machine cooperative learning control," in *Proc. IEEE Int. Symp. Intell. Control (ISIC)*, Oct. 2003, pp. 81–86.

[6] J. Wagner, T. Baur, Y. Zhang, M. F. Valstar, B. Schuller, and E. André, "Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora," 2018, *arXiv:1802.02565*. [Online]. Available: http://arxiv.org/abs/1802.02565

[7] D. Vidhate and P. Kulkarni, "Cooperative machine learning with information fusion for dynamic decision making in diagnostic applications," in *Proc. Int. Conf. Adv. Mobile Netw., Commun. Appl.*, Aug. 2012, pp. 70–74.

[8] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[9] Y. Khaldi and A. Benzaoui, "A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions," *Evolving Syst.*, May 2020, doi: 10.1007/s12530-020-09346-1.

[10] H. Zhang and Z. Mu, "Compound structure classifier system for ear recognition," in *Proc. IEEE Int. Conf. Autom. Logistics*, Sep. 2008, pp. 2306–2309.

[11] L. Gutierrez, P. Melin, and M. Lopez, "Modular neural network integrator for human recognition from ear images," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–5.

[12] A. Tariq, M. A. Anjum, and M. U. Akram, "Personal identification using computerized human ear recognition system," in *Proc. Int. Conf. Comput. Sci. Netw. Technol.*, Dec. 2011, pp. 50–54.

[13] Z. Mu, L. Yuan, Z. Xu, D. Xi, and S. Qi, "Shape and structural feature based ear recognition," in *Advances in Biometric Person Authentication*. Berlin, Germany: Springer, 2004, pp. 663–670.

[14] I. Omara, L. Feng, Z. Hongzhi, and Z. Wangmeng, "A novel geometric feature extraction method for ear recognition," *Expert Syst. Appl.*, vol. 65, pp. 127–135, Dec. 2016.

[15] L. Lakshmanan, "Efficient person authentication based on multi-level fusion of ear scores," *IET Biometrics*, vol. 2, no. 3, pp. 97–106, Sep. 2013.

[16] I. Adjabi, A. Ouahabi, A. Benzaoui, and S. Jacques, "Multi-block color-binarized statistical images for single-sample face recognition," *Sensors*, vol. 21, no. 3, p. 728, Jan. 2021.

[17] A. Kumar and C. Wu, "Automated human identification using ear imaging," *Pattern Recognit.*, vol. 45, no. 3, pp. 956–968, 2012.

[18] L. Ghoualmi, A. Draa, and S. Chikhi, "An ear biometric system based on artificial bees and the scale invariant feature transform," *Expert Syst. Appl.*, vol. 57, pp. 49–61, Sep. 2016.

[19] Ž. Emeršić, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017.

[20] D. P. Chowdhury, S. Bakshi, G. Guo, and P. K. Sa, "On applicability of tunable filter bank based feature for ear biometrics: A study from constrained to unconstrained," *J. Med. Syst.*, vol. 42, no. 1, pp. 1–20, Jan. 2018.

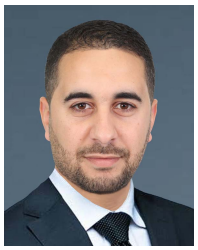
[21] M. Hassaballah, H. A. Alshazly, and A. A. Ali, "Ear recognition using local binary patterns: A comparative experimental study," *Expert Syst. Appl.*, vol. 118, pp. 182–200, Mar. 2019.

[22] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep learning-based sign language digits recognition from thermal images with edge computing system," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10445–10453, May 2021.

[23] A. Ouahabi and A. Taleb-Ahmed, "Deep learning for real-time semantic segmentation: Application in ultrasound imaging," *Pattern Recognit. Lett.*, vol. 144, pp. 27–34, Apr. 2021.

[24] A. S. Alharthi, S. U. Yunas, and K. B. Ozanyan, "Deep learning for monitoring of human gait: A review," *IEEE Sensors J.*, vol. 19, no. 21, pp. 9575–9591, Nov. 2019.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[27] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2015, pp. 1–12.
[29] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360
[30] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020.
[31] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus CNN features for ear recognition," *Symmetry*, vol. 11, no. 12, p. 1493, Dec. 2019.
[32] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, p. 4139, Sep. 2019.
[33] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, "A deep learning approach for person identification using ear biometrics," *Appl. Intell.*, vol. 51, no. 4, pp. 2161–2172, 2020.
[34] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, May 2018.
[35] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, May 2018.
[36] J. Zhang, W. Yu, X. Yang, and F. Deng, "Few-shot learning for ear recognition," in *Proc. Int. Conf. Image, Video Signal Process. (IVSP)*, Feb. 2019, pp. 50–54.
[37] Z. Wang, J. Yang, and Y. Zhu, "Review of ear biometrics," *Arch. Comput. Methods Eng.*, vol. 28, no. 1, pp. 149–180, 2021.
[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[39] Y. Khaldi and A. Benzaoui, "Region of interest synthesis using image-to-image translation for ear recognition," in *Proc. Int. Conf. Adv. Aspects Softw. Eng. (ICAASE)*, Nov. 2020, pp. 1–6.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[42] *The AMI Ear Dataset*. Accessed: Mar. 26, 2021. [Online]. Available: http://ctim.ulpgc.es/research_works/ami_ear_database
[43] *The USTB Ear Dataset*. Accessed: Mar. 26, 2021. [Online]. Available: http://www1.ustb.edu.cn/resb/en/news/news3.htm
[44] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," in *Articulated Motion Deformable Objects*. Cham, Switzerland: Springer, 2018, pp. 85–94.
[45] A. Mimouna *et al.*, "OLIMP: A heterogeneous multimodal dataset for advanced environment perception," *Electronics*, vol. 9, no. 4, p. 560, Mar. 2020.
[46] W. Raveane, P. L. Galdámez, and M. A. G. Arrieta, "Ear detection and localization with convolutional neural networks in natural images and videos," *Processes*, vol. 7, no. 7, p. 457, Jul. 2019.
[47] A. Ouahabi, "Multifractal analysis for texture characterization: A new approach based on DWT," in *Proc. 10th Int. Conf. Inf. Sci., Signal Process. Their Appl. (ISSPA)*, May 2010, pp. 698–703.
[48] D. Meriem, O. Abdeldjalil, B. Hadj, B. Adrian, and K. Denis, "Discrete wavelet for multifractal texture classification: Application to medical ultrasound imaging," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 637–640.
[49] A. Ouahabi, "A review of wavelet denoising in medical imaging," in *Proc. 8th Int. Workshop Syst., Signal Process. Their Appl. (WoSSPA)*, May 2013, pp. 19–26.
[50] S. Ahmed, Z. Messali, A. Ouahabi, S. Trepout, C. Messaoudi, and S. Marco, "Nonparametric denoising methods based on contourlet transform with sharp frequency localization: Application to low exposure time electron microscopy images," *Entropy*, vol. 17, no. 5, pp. 3461–3478, May 2015.
[51] A. Ouahabi, *Signal and Image Multiresolution Analysis*, 1st ed. London, U.K.: Wiley, 2012.

Yacine Khaldi received the M.Sc. degree in computer sciences from the University of Ouargla, Algeria, in 2017. He is pursuing the Ph.D. degree with the University of Bouira, Algeria. His current research interests include biometrics and machine learning.



Amir Benzaoui received the M.Sc. degree in computer sciences from Annaba University in 2011 and the Ph.D. degree in electronics from Guelma University in 2015. He is an Associate Professor with the University of Skikda, Algeria. His research interests include biometrics and machine learning.

Abdeldjalil Ouahabi is a Full Professor with the University of Tours, France. His research interests include image and signal processing, biomedical engineering, and machine learning. Prof. Ouahabi has authored over 170 published articles in these areas. He is a member of the editorial board of several WoS journals. He is currently a guest editor of four SI of WoS journals. He was a recipient of many awards, including the Outstanding Reviewer Award from *Knowledge-Based Systems* (Elsevier, 2018) and *Measurement* (Elsevier, 2016), and the Best Paper Award from IEEE Instrumentation and Measurement Society in 1999.

Sébastien Jacques (Member, IEEE) has been an Associate Professor with the University of Tours, France, since 2012. His research interests include smart electricity management systems, machine learning, and reliability of electronic systems.

Abdelmalik Taleb-Ahmed has been a Full Professor with the Université Polytechnique Hauts-de-France, Valenciennes, France, since 2004. His research interests include segmentation, classification, data fusion, pattern recognition, computer vision, and machine learning, with applications in biometrics, video surveillance, autonomous driving, and medical imaging. Prof. Taleb-Ahmed has coauthored over 85 peer-reviewed articles and co-supervised 30 graduate students in these areas of research.