



# MDFNet: an unsupervised lightweight network for ear print recognition

Oussama Aiadi<sup>1</sup> · Belal Khaldi<sup>1</sup> · Cheraa Saadeddine<sup>1</sup>

Received: 20 October 2021 / Accepted: 30 May 2022 / Published online: 18 June 2022  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

In this paper, we propose an unsupervised lightweight network with a single layer for ear print recognition. We refer to this method by MDFNet because it relies on gradient Magnitude and Direction alongside with responses of data-driven Filters. At first, we align ear using Convolution Neural Network (CNN) and Principal Component Analysis (PCA). MDFNet starts by generating a filter bank from training images using PCA. This is followed by a twofold layer, which comprises two operations namely convolution using learned filters and computation of gradient image. To prevent over-fitting, a binary hashing process is done by combining different filter responses into a single feature map. Then, we separately construct histograms for each of gradient magnitude and direction according to the feature map. These histograms are then normalized, using power- $L_2$  rule, to cope with illumination disparity. Several fusion rules are evaluated to combine the two histograms. The main novelty of MDFNet lies in its simple architecture, effectiveness, the good compromise between processing time and performance it provides along with its high robustness to occlusion. We conduct extensive experiments on three public datasets namely AWE, AMI and IIT Delhi II. Experimental results demonstrate the effectiveness of MDFNet, which achieves high recognition rates (82.5%, 97.67% and 98.96%, respectively), and outperformed several state of the art methods with a high robustness to occlusion. Experiments revealed also the actual need for considering ear alignment.

**Keywords** Biometrics · Unconstrained ear · Recognition · CNN · Deep learning · Unsupervised learning

## 1 Introduction

Nowadays, the need for biometric systems that are capable to reliably recognize persons is quite ostensible. Emerging applications of biometrics include self-driving, smartphone security, prevention of human trafficking, e-voting, crime investigation, etc. Moreover, with the development of smart cities, cloud and mobile computing, biometric systems have become increasingly demanded. For instance, biometrics can be used to enhance the quality of service in a smart city. Typical examples include ubiquitous healthcare, privacy

protection, access control, screening, traveling facilitation and law enforcement (Barra et al. 2018). In order to perform identification, different physiological and behavioural traits including face, iris, ear, gait and signature can be utilized. Each of these traits should satisfy certain constraints in order to be eligible for identification tasks. The main constraints are measurability, permanence, uniqueness, universality and collectability.

In recent years, ear has attracted considerable and growing interest in biometric community due to its inherent characteristics compared to other traits. Unlike face, ear is insusceptible to facial expressions, emotions, glasses, make-up and it keeps unchanged from 8 to 70 years old (Kamboj et al. 2021). Nowadays, with outbreak of COVID-19, people wearing masks for prevention purposes. Hence, it is very challenging for face identification systems to recognize persons. Additionally, ear is a non-intrusive trait because it can typically be acquired without informing the concerned person. Another appealing property of the ear is its high distinctiveness even in the case of identical twins (Emeršič et al. 2017a; b). Figure 1 presents the anatomy of human ear.

✉ Oussama Aiadi  
aiadi.oussama@univ-ouargla.dz

Belal Khaldi  
khaldi.belal@univ-ouargla.dz

Cheraa Saadeddine  
cheraa.saadeddine@univ-ouargla.dz

<sup>1</sup> LINATI Laboratory, Department of Computer Science and Information Technology, University of Kasdi Merbah, 30000 Ouargla, Algeria

- 1- Helix
- 2- Crus of antihelix
- 3- Antihelix
- 4- Anti tragus
- 5- Cavum conchae
- 6- Foseta
- 7- Crus of helix
- 8- Tragus
- 9- Lobe



**Fig. 1** Anatomy of the human ear

In the literature, several attempts have been made to develop a reliable automatic system for ear recognition. They can be categorized into three groups. Works (Hansley et al. 2018; Khaldi and Benzaoui 2020a, b; Khaldi and Benzaoui 2020a, b) within the first category are concerned with pre-process ear image to facilitate the subsequent steps in the recognition pipeline. With impressive performance of deep learning on different vision tasks (Samai et al. 2018; Hamrouni et al. 2021; Kumar et al. 2021; Zhang et al. 2021), deep-based approaches (Emeršič et al. 2018; Khaldi and Benzaoui 2020a, b) have been integrated to improve the lucidity of ear print. The second type of methods try to handle a crucial step in recognition process namely features learning. As for the third category of methods (Dodge et al. 2018; Alshazly et al. 2019a, b; Kacar and Kirci 2019; Omara et al. 2021), it aims at designing learning models that can discriminate different classes successfully. In spite of the huge amount of works devoted to cope with ear recognition problem, we believe that there is still room for improvement. What is actually coveted is a method capable to perform accurate identification, yet, in a reasonable time.

In this paper, we put forward a new unsupervised lightweight single layer-based network for human ear description and recognition. This method is denoted by MDFNet, as it integrates gradient information, namely Magnitude and Direction, together with feature maps produced using data-driven Filters. To sum up, contributions of the current paper can be summarized as follows.

- We propose an unsupervised lightweight network with a single-layer (MDFNet) for faithful ear recognition.
- Compared to certain state-of-the-art methods, which are computationally expensive, MDFNet provides a good trade-off between processing time and performance.
- We show the indispensable role of ear alignment in improving the performance of ear identification.
- We conduct comprehensive experiments on three public datasets, namely: AWE, AMI and IIT Delhi II. The obtained results demonstrate the efficiency of the pro-

posed method, which outscored relevant networks as well as several state of the art methods.

- We assess the performance of MDFNet under different occlusion levels. MDFNet has shown high tolerance towards occlusion.

The remainder of this paper is organized as follows. Section 2 presents an overview on the related works for ear recognition. Section 3 is devoted to present MDFNet. In Sect. 4, we provide details about the proposed method. Experimental results and discussions are given in Sect. 5. Finally, Sect. 6 draws some conclusions and perspectives.

## 2 Related work

### 2.1 Pre-processing-based methods

In unconstrained scenarios, ear pre-processing plays a vital role in the recognition pipeline. Main pre-processing acts including ear alignment, inpainting, colorization and region of interest segmentation. In Ribič et al. (2016), authors has addressed the first point and studied the impact of ear alignment on ear recognition for the case of AWE dataset. As for the case of ear images with slight roll and yaw, experiments have revealed that ear alignment improves the recognition outcomes. However, aligning ears with strong roll and yaw has negatively affected the recognition results. Authors suggested that further sophisticated ear alignment procedures are required.

More recently, deep learning techniques were utilized for ear image pre-processing. In Emeršič et al. (2018), authors proposed using inpainting techniques to replace the ear accessory by a surrogate region using CNN-based auto-encoders. Authors assumed that both shape and location of ear accessory is already know, and adopted three different approaches to replace the removed region. The first approach replaces the removed region with naturally looking regions, while the second approach simply fill in the blank region by a black color. As for the last approach it considers replacing the removed region by the average color of the image. A landmark detection-based geometrical ear normalization method is proposed in Hansley et al. (2018). In particular, a customized CNN is trained on augmented data to detect ear landmarks, scale and pose of ear images are then normalized using Principal Components Analysis (PCA). Different image features were considered, where an improvement of about 5% was achieved over results obtained using unaligned ear images. Ear alignment based on deformable models has consistently improved performance of the system proposed in Zhou and Zaferiou (2017). In addition, authors have compared different discriminant and generative approaches for ear landmark detection. Color information

losing once performing training using color images and testing using gray-scale images was the concern of the study in Khaldi and Benzaoui (2020a, b). To handle this problem, authors proposed colorizing test images using conditional deep convolutional generative adversarial networks. In Khaldi and Benzaoui (2020a, b), an image-to-image translation technique is used to segment region of interest within ear image, and to synthesize missing parts of ear such as occluded regions. In particular, a pix2pix GAN which is trained on the AWE dataset is used to synthesize the missing parts of the ear. Then, different handcrafted features, such as local binary patterns and local phase quantization, are extracted from different color channels of the ear image to describe and classify ear images. As can be noticed, the previous works differ from each other in terms of the work target i.e., colorization, inpainting and alignment. Meanwhile, they commonly used deep learning architectures due to their achievements on different vision tasks.

## 2.2 Feature extraction-based methods

Extract features that can effectively describe rich ear structure was the subject matter of several studies during the last few years. Local features, such as Local Phase Quantization (LPQ), have widely been used for different recognition tasks, including ear recognition (Al Rahhal et al. 2018; Korichi et al. 2018; Sarangi et al. 2021). As instance, in Al Rahhal et al. (2018), images are divided into horizontal stripes, where LPQ features are extracted from each stripe. The stripe-wise features are then concatenated to form the final descriptor. This local approach has greatly improved the recognition performance of LPQ with about 20%. To deal with 3D ear images, a descriptor has been formed from key-points in 2D ear images detected using curvilinear structure in Ganapathi et al. (2018). Authors suggested detecting the feature key-points from 2D ear images by using curvilinear structure and then mapping those features to the 3D ear images. In Sajadi and Fathi (2020), authors have considered extracting local and global spectral features such as LPQ and Gabor Zernike operators, which are combined using a genetic algorithm. To classify test images, k-nearest neighbor classifier (KNN) with Canberra distance is used. Due to the remarkable performance they reached, local features are extracted based on tunable filter bank for ear verification in Chowdhury et al. (2018). Authors in Hassaballah et al. (2020) presented a Local Binary Pattern (LBP) variant termed as robust local oriented patterns (RLOP), which is designed to be robust against noise and rotation. To make RLOP invariant to rotation, neighboring pixels are binarized according to the mean value of each patch instead of the central pixel value. More recently, local features, involving LPQ and Local Directional Pattern (LDP), were used for multimodal recognition using ear and face images (Sarangi et al.

2021). The kernel discriminative common vector (KDCV) technique is utilized to derive discriminative and non-linear features from the extracted features.

Authors in Alshazly et al. (2019a, b) compared the performance of handcrafted against deep learning-based features. Experiments on different public datasets revealed that deep learning-based features considerably outscored the conventional handcrafted features including LBP, Histogram of Gradient (HoG) and Binarized Statistical Image Features (BSIF). Noting that comparison of deep with handcrafted features has largely been debated in several literature works (Khaldi et al. 2019; Korichi et al. 2020). In another study (Priyadarshini et al. 2021), a customized six layer deep CNN which is composed of stacked convolution, subsampling, batch and output layers is proposed. The performance of this network is tested on AMI and IIT Delhi II datasets, where effect of varying the network parameters (e.g., learning rate and activation function) was studied. Similarly, Hamdany et al. (2021) proposed a CNN architecture for ear recognition. Different back-propagation techniques, including Adaptive moment estimation (ADAM) and stochastic gradient descent with momentum (SGDM), were tested. In general, due to its generalization capabilities, deep learning has been widely applied to solve different vision tasks (Kamilaris and Prenafeta-Boldú 2018; Połap 2019; Połap et al. 2021). Therefore, compared to the other kinds of features, features learned from deep networks have reached best performance in the task of ear recognition.

## 2.3 Learning-based methods

Designing the learning model is a challenging task, as it has a decisive role in determining the quality of recognition outcomes. The method proposed in Omara et al. (2021) is based on learning Mahalanobis distance metric using the LogDet divergence metric learning. This metric is learned on deep features that are extracted using pre-trained networks, where discriminant correlation analysis is used to fuse features from the last two layers of the selected networks. Authors in Alshazly et al. (2019a, b) suggested building an ensemble of deep models by combining decisions from multiple models to boost recognition yields. Different models have been investigated including pre-trained models, fine-tuned models and models trained with random weights. Fine-tuned models have shown superior performance over the others, thus, they have been considered for constructing ensembles based on voting committees. Another ensemble learning scheme based on CNN was presented in Dodge et al. (2018) to prevent individual networks from over-fitting. A set of CNNs are firstly fine-tuned, then, an ensemble of CNN is formed by averaging the soft-max outputs of the considered networks. In Mawloud et al. (2016), a sparse coding joint decision rule has been introduced for ear recognition, where

multi-scale co-occurrence of adjacent LBP was used as a feature extractor. The proposed method has been tested on IIT Delhi II (225 subjects) and IIT Delhi I (125 subjects) datasets, and shown a promising tolerance to occlusion.

In the last few years, a remarkable advance has been made on the task of ear identification. However, existing methods suffer from some drawbacks. The first remark we can make is that most methods achieved a relatively good performance yet requiring a considerable amount of processing time, especially deep-based approaches. It would be beneficial if one can attain a good recognition accuracy in a reasonable processing time. In addition, a well-suited feature that consider ear's key characteristics, e.g., curvature degree of helix, could be a good alternative to generic handcrafted features.

### 3 MDFNet: an unsupervised lightweight network for ear description

This section presents the steps of our proposed lightweight network namely MDFNet, which is inspired by PCANet (Chan et al. 2015). The general flowchart of MDFNet is depicted by Fig. 2.

#### 3.1 PCA filter bank generation

The first step in MDFNet is generating a filter bank from training images using PCA. In fact, compared to predefined filters, data-driven PCA filters have a better generalization capability (Chan et al. 2015). Assuming we are given a training set of  $N$  images of size  $m \times n$ , denoted by  $\{T^i\}$ , such that

$i = 1, \dots, N$ . Suppose also that around every pixel we consider a  $p_1 \times p_2$  patch. All patches are gathered from the entire set  $T^i$  and then vectorized, such that patches of the  $i$ th image are referred to as  $x_{i,1}, x_{i,2}, \dots, x_{i,\tilde{m}\tilde{n}}$ , where  $\tilde{m} = m - \lfloor \frac{p_1}{2} \rfloor$  and  $\tilde{n} = n - \lfloor \frac{p_2}{2} \rfloor$ , and  $\lfloor b \rfloor$  stands for the smallest integer  $\geq b$ . Patch mean is then subtracted from each patch, thus, we obtain  $\bar{X}_i = [\bar{x}_{i,1}, \bar{x}_{i,2}, \dots, \bar{x}_{i,\tilde{m}\tilde{n}}]$ . The same matrix is built for all the images and combined together as follows

$$X = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_N] \quad (1)$$

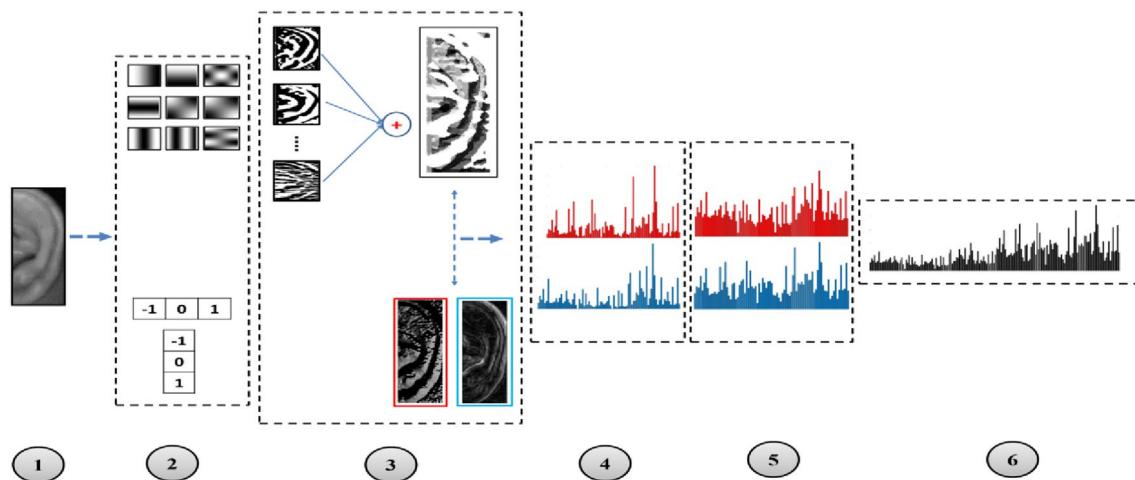
PCA targets finding an orthogonal matrix that can minimize the reconstruction error, such that number of filters is set to  $M$ , this can be done according to the next equation

$$\min_{V \in \mathbb{R}^{p_1 \times p_2 \times M}} \|X - VV^T X\|_F^2 \quad \text{s.t. } VV^T = I_M \quad (2)$$

where  $I_M$  and  $V$  stand for the unit matrix of size  $M \times M$  and the standard orthogonal matrix, respectively. PCA is performed on the set  $X$  to extract the principal eigenvectors, which are then sorted in descending manner according to their eigenvalues. The top  $M$  principal eigenvectors are reconstructed as PCA filters, according to Eq. (3)

$$W_k = \text{mat}_{p_1, p_2}(q_k(XX^T)) \in \mathbb{R}^{p_1 \times p_2}, k = 1, 2, \dots, M \quad (3)$$

where  $W_k$  is the  $k$ th PCA filter,  $XX^T$  represents the covariance matrix of  $X$ ,  $q_k()$  extracts the principal eigenvectors from the covariance matrix, and  $\text{mat}_{p_1, p_2}()$  is a function that maps a



**Fig. 2** General flowchart of MDFNet: (1) input image (2) filter bank generation using PCA, and preparing pre-fixed horizontal and vertical filters (3) a twofold layer in which input image is convolved using learned filters and filtered using pre-fixed filters to compute gradient magnitude (image with red border) and direction (image with blue

border), also a binary hashing process is done to fuse all feature maps in a single feature map (4) a dual histogram is generated based on the single feature map and both gradient images (5) dual histogram normalization (6) concatenating histograms (color figure online)



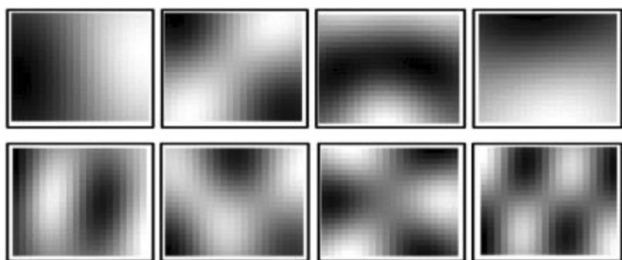


Fig. 3 Typical filters that are generated using PCA from AWE dataset

vector  $v \in \mathbb{R}^{p_1 p_2}$  to a matrix  $W \in \mathbb{R}^{p_1 \times p_2}$ . Figure 3 presents some typical filters learned using PCA.

### 3.2 Twofold single layer for image convolution and filtering

MDFNet is composed of a twofold single layer along with other successive steps for features learning. Hereafter, we present the two components of this layer namely convolution and filtering.

#### 3.2.1 Convolution

In this layer, the input image  $Tr^i$  is convolved using different PCA filters. At first, the boundary of image is zero-padded to have an output image, which is denoted by  $C^i$ , with same size as  $Tr^i$ . Convolution can be expressed as

$$C^i = Tr^i * W_k, i = 1, \dots, N \text{ and } k = 1, \dots, M \tag{4}$$

#### 3.2.2 Filtering

In this sub-step, to extract gradient image, the input image  $Tr^i$  is filtered in both vertical and horizontal directions using two pre-defined masks namely  $s_v$  and  $s_h$ , which are defined as  $s_v = [-101]^T$  and  $s_h = [-101]$ . Images are filtered horizontally and vertically using Eqs. (5) and (6), respectively

$$G_v = Tr^i \Delta s_v \tag{5}$$

$$G_h = Tr^i \Delta s_h \tag{6}$$

where the operator  $\Delta$  stands for filtering operation. Then, magnitude and direction, which are referred to as  $O_1^i$  and  $O_2^i$ , respectively, are computed from  $G_v$  and  $G_h$  as follows

$$O_1^i = \sqrt{G_v^2 + G_h^2} \tag{7}$$



Fig. 4 Binary hashing process: images with red border are feature maps produced by the convolution layer, the rightmost image with the blue border is the result of fusing these feature maps (color figure online)

$$O_2^i = \arctan \frac{G_v}{G_h} \tag{8}$$

Indeed, gradient magnitude and direction are adopted to depict the curvatures within different ear regions involving helix, tragus, lobe and the antihelix.

### 3.3 Binary hashing

Since that using a unique layer and several feature maps may cause our model to over-fit, a binary hashing process is conducted on feature maps to quantify filter responses and relieve over-fitting. Since filters responses are real-valued, the first step is binarizing feature maps using zero as a threshold i.e., assign pixels having positive responses with a value equals to one, and zero otherwise. Then, the binarized feature maps are fused into a single image (denoted by  $O_3^i$ ), according to the next equation

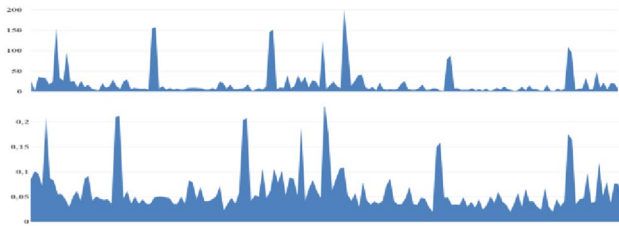
$$O_3^i = \sum_{r=1}^M 2^{r-1} \times \text{bin}(C^r) \tag{9}$$

such that  $\text{bin}$  stands for the binarization function. As a result, every pixel in  $O_3^i$  falls within the range  $[02^M - 1]$ . Figure 4 depicts the binary hashing process.

### 3.4 Block-wise filter response-based direction/magnitude histograms generation

The bottleneck of MDFNet is to simultaneously characterize three types of crucial information namely: gradient direction and magnitude as well as filter responses. We separately extract two histograms: the first one binds gradient direction and filter responses, while the second one binds gradient magnitude and filter responses. Note that the histogram dimension is equal to  $2^M$ . Filter response-based magnitude histogram (denoted by  $HISTM$ ) is generated using Eq. (10)

$$HISTM(w) = \sum_{a=1}^u O_1^i(Ind_w) \tag{10}$$



**Fig. 5** Histogram before normalization (at top) and histogram after normalization (at bottom). This latter seems to be more evenly distributed

where  $u = |Ind_w|$ ,  $w$  is an index for the histogram bins, and  $Ind_w$  indicates the set of pixels' spatial coordinates for which  $O_3^i$  equals to  $w$ ,  $Ind_w$  is given by

$$Ind_w \equiv (O_3^i = w) \quad (11)$$

such that  $\equiv$  stands for the assignment operator, which assigns  $Ind_w$  with spatial coordinates of pixels for which  $O_3^i = w$

Similarly, filter response-based direction histogram (denoted by *HISTD*) is generated according to the following equation

$$HISTD(w) = \sum_{a=1}^u O_2^i(Ind_w) \quad (12)$$

where  $u = |Ind_w|$  and  $Ind_w \equiv (O_2^i = w)$

Such that  $\equiv$  assigns  $Ind_w$  with spatial coordinates of pixels for which  $O_2^i = w$ . We point out that  $O_2^i$  is quantized, before generating the direction histogram, to consider significant angles and improve the method robustness. To incorporate spatial relationship, we extract both histograms in a block-wise manner. Therefore,  $O_1^i$ ,  $O_2^i$  and  $O_3^i$  are divided into blocks, and for each block a dual histogram is extracted.

### 3.5 Histograms normalization and concatenation

After having block-wise dual histogram (*HISTD* and *HISTM*) extracted, we normalize them for two reasons, the first one is to gain some robustness against illumination changes, while, the second is to cope with disparity of visual features. Moreover, normalization helps making the histogram evenly distributed, and alleviates disparity of visual features, which can actually, distort the recognition yields. We use power- $L_2$  norm, for a histogram  $H = (h_1, \dots, h_f)$ , this normalization is done as follows

$$h_i = \frac{|h_i|^\beta \times \text{sign}(h_i)}{H} \quad (13)$$

$\|H\|$  represents the  $L_2$  norm of  $H$  and  $\beta$  is a constant, such that  $0 \leq \beta \leq 1$ . In this work, we set  $\beta$  to 0.5. Figure 5 depicts the effect of normalization. To concatenate the two

histograms, we evaluate different score-level fusion schemes including sum, min and max alongside with the baseline feature-level scheme.

### 3.6 Recapitulation

The architecture of the proposed network is designed to faithfully describe the ear image. The twofold layer acts as a feature detector, where input ear image is convolved using learned filters along with pre-fixed filters to extract image gradient magnitude and direction. By carefully inspecting the ear image, we can notice that gradient magnitude and direction can provide interesting description of the ear structure. In addition, convolving the ear image using learned filters will highlight the distinguishing regions in the ear, which can help discriminating different classes. Binary hashing is considered to prevent the network from over-fitting and to quantify the filter responses by fusing different feature map into a single map. The bottleneck of the network is to establish a link between gradient images and the fused feature map. Such a representation can be reached by histogramming the gradient images based on the fused map. In addition, to strengthen our representation, we include the spatial relationship information by extracting the histograms from local image blocks. These local histograms are normalized to improve the network robustness against illumination changes.

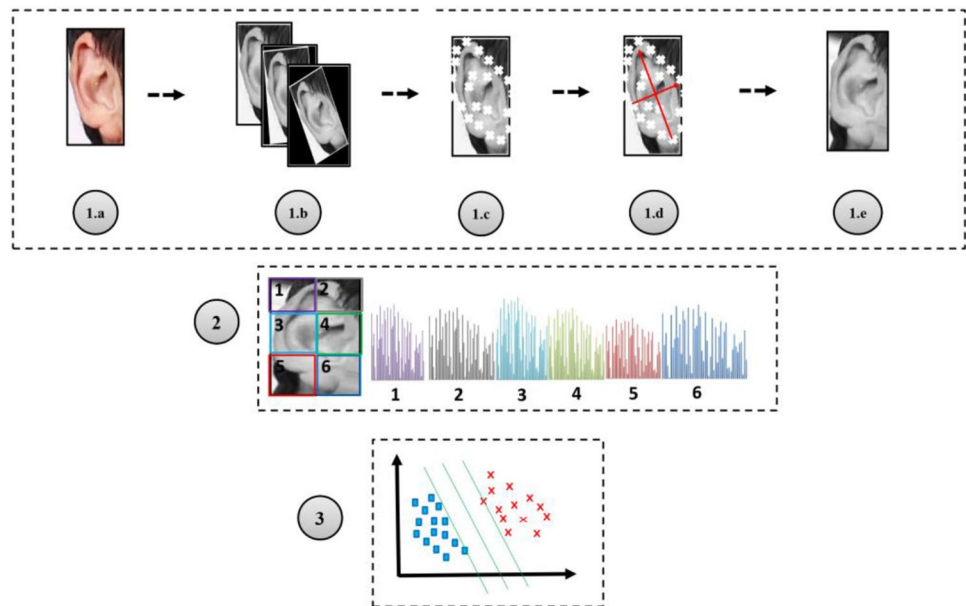
Indeed, one can apply transfer learning using MDFNet. However, it is worth noting that applying this strategy depends on the nature of the images on which the network is trained (original images) and the images to which the transfer learning will be applied (targeted images). If the original and targeted images are close to each other, thus, it will be possible to use the learned filters for the targeted images. In addition, it should also be noted that it might be better to tune the other network parameters such as block size to get a better performance.

## 4 Proposed method

The proposed method consists in three main stages (Fig. 6) namely: pre-processing (alignment), features extraction and matching, as follows.

- *Ear image alignment*: as shown by the relevant studies, ear pre-processing helps improving recognition results by mitigating intra-class variation, particularly in unconstrained scenarios. Therefore, we consider, pre-process unconstrained ear images based on the method in (Hansley et al. 2018). As a first step, landmarks provided with the ITWE dataset are used to train a customized CNN (Table 1) on landmarks locations. To prevent over-fitting

**Fig. 6** General flowchart of the proposed method: (1.a) input image (1.b) data augmentation (1.c) landmark detection using CNN (1.d) geometrical ear normalization using PCA (1.e) output image (2) block-wise features learning using MDFNet (3) classification using SVM



**Table 1** CNN architecture for ear landmarks detection

No.	Type	Input	Filter	Stride	Drop	Output
1	Conv/Relu	$96 \times 96 \times 1$	$3 \times 3 \times 1 \times 32$	1	10%	$96 \times 96 \times 32$
2	MaxPool	$96 \times 96 \times 32$	$2 \times 2$	2		$48 \times 48 \times 32$
3	Conv/Relu	$48 \times 48 \times 32$	$2 \times 2 \times 32 \times 64$	1	20%	$48 \times 48 \times 64$
4	MaxPool	$48 \times 48 \times 64$	$2 \times 2$	2		$24 \times 24 \times 64$
5	Conv/Relu	$24 \times 24 \times 64$	$2 \times 2 \times 64 \times 128$	1	30%	$24 \times 24 \times 128$
6	MaxPool	$24 \times 24 \times 128$	$2 \times 2$	2		$12 \times 12 \times 128$
	Flattening	$12 \times 12 \times 128$	–	–	–	18,432
7	Fc/Relu	18,432	–	–	50%	1000
8	Fc/Relu	1000	–	–	–	1000
9	Fc	1000	–	–	–	110

and improve the CNN generalization power, PCA is used to perform data augmentation on the available labeled ear images. Specifically, for a given training image PCA is used to get the upright orientation of the ear from 2D landmark coordinates. Afterwards, different images were generated by rotating the upright from  $-45^\circ$  to  $+45^\circ$  with a step of  $3^\circ$ . Images are also transformed into a random scale and translated of up 20% of the ear size in both axes. After generating these additional images, a total of 15,500 images are considered for training with a size of  $96 \times 96$  pixels. We used a batch size of 36 images, and carried out the training for 2000 epochs, which gives about 860,000 iterations. Scales and poses of ear images are then normalized using PCA. Figure 7 shows examples for normalized ears and top of Fig. 6 presents the ear alignment procedure.

- *Features extraction*: in this step, features are extracted from training images using our MDFNet. To take advantage of spatial relationship, we extract features from local



**Fig. 7** Geometrical ear normalization, first row presents original ear images, and the second row presents their respective normalized ears

image blocks. One advantage of the local approach is that it prevents pooling all features in a compacted vector where discriminative features may be dominated by the common global features.

- *Matching*: a linear one-versus-all support vector machine classifier is trained, and used to classify probe images.

## 5 Evaluation and discussion

### 5.1 Experimental datasets

#### 5.1.1 AWE

The Annotated Web Ear (AWE) database (Emeršič et al. 2017a, b) involves images that are collected from the Internet. AWE constituted of 1000 ear images for 100 subjects, where each subject has 10 images. Image sizes vary from  $15 \times 29$  to  $473 \times 1022$  pixels. It is considered as one of the most challenging ear databases, as it has been taken under uncontrolled conditions with a significant difference in terms of image size, illumination and angles. It also contains images that are occluded by ear accessories, earrings and hair.

#### 5.1.2 AMI

The Mathematical Analysis of Images (AMI) ear dataset contains 700 ear images from 100 subjects. Images have been taken in an indoor environment, with a resolution of  $492 \times 702$  pixels. Each subject has seven different images taken at different poses namely right, left, forward, up and down. In particular, five images represent the right ear, such that each image is taken with a different head pose (i.e., right, left, forward, up and down). As for the remaining two images, the first one is also for the right ear, but it is taken with a different focal length, while the second one is for the left ear.

#### 5.1.3 IIT Delhi II

This dataset (Kumar and Wu 2012) is acquired from student and staff at IIT Delhi University (India). It is made up of 793 images of 221 subjects, where each subject includes at least 3 images. All images in the dataset are remotely taken in an indoor environment and under different lighting and pose conditions. Figure 8 presents typical samples from the three datasets.

### 5.2 Experimental settings and evaluation metrics

To perform recognition experiments, we split each dataset into two disjoint groups, one for training and the other for testing. Specifically, for the AWE and AMI datasets, 60% of images are used for training, and the remaining (40%) are intended for testing. Regarding the IIT Delhi II dataset, we considered two images from the first three images for training, while the remaining images are considered as probes. As each subject has at least three images, we performed three permutations and computed the average performance.



**Fig. 8** Representative samples from each dataset, first row corresponds to AWE, second row to AMI and third row to IIT Delhi II

It is worth noting that we have considered aligning images from the AWE dataset only, as IITD II and AMI are inherently well aligned.

To measure the performance of the proposed method, we consider two metrics which are rank-1 recognition rate and Cumulative Match Characteristics (CMC) curve. Given a test image denoted by  $T_i$ , MDFNet produces a similarity matrix which indicates the similarity scores between training and testing samples. By sorting the similarity scores, we obtain the rank of the ground truth subjects i.e.,  $rank(T_i)$ . The rank- $k$  recognition rate is given by Eq. (14)

$$Recognition\_Rate(k) = \frac{|\{T_i | rank(T_i) \leq k\}|}{N} \quad (14)$$

where  $|\cdot|$  stands for the set cardinality, and  $N$  represents the number of test images. As for the CMC curve, it is rank-based metric that measures the probability of determining the correct matching within the top  $k$  ranks. CMC curve plots the recognition rate in terms of  $k$  rank, as follows

$$CMC(k) = Recognition\_Rate(k) \quad (15)$$

### 5.3 Experimental results

#### 5.3.1 MDFNet parameters tuning

The aim of this experiment is to measure the performance of MDFNet by varying different parameters. Therefore, we carry out extensive experiments to determine the parameter subset that maximizes the recognition outcomes. Especially, three hyper-parameters namely number of filters, filter size and block size are required for computing MDFNet. Among several, we have experimented six subsets as reported by Table 2.



**Table 2** The subsets of parameters that are considered

Subset of parameters	Number of filters for PCA	Filter size	Block size
1	5	5 × 5	16 × 16
2	5	7 × 7	32 × 32
3	9	9 × 9	16 × 16
4	9	9 × 9	100 × 100
5	9	9 × 9	25 × 25
6	7	7 × 7	100 × 100

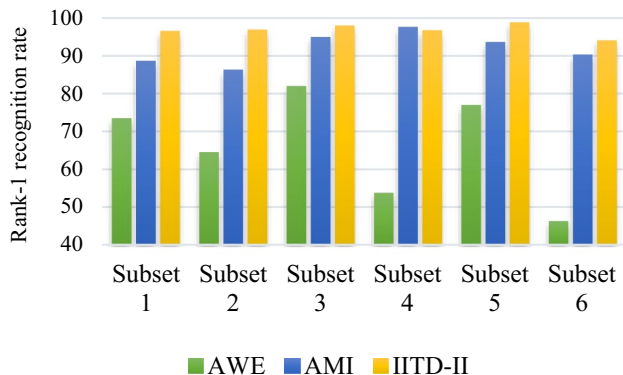
**Fig. 9** Rank-1 recognition rates yielded by different subsets of parameters for the three datasets

Figure 9 presents rank-1 recognition rate for the three datasets. From this figure, we can make the following observations

- As for the number of filters, we note that, in most cases, using 9 filters scored the best results compared to 5 and 7 filters. Note that by increasing number of filters, the dimension of feature vector increases exponentially, because dimension of block-wise histograms is power of two. Thus, it is recommended to choose the best trade-off between the number of filters and feature vector dimension.
- Regarding filter size, 9 × 9 filters have obtained the best results. Filter size in this case is big enough to capture interesting characteristics of ear.
- Regarding block size, we notice that each dataset requires a different size. This can be justified by the significant difference between images of each dataset as shown by Fig. 8.
- From the six subsets, we notice that the third (82%), fifth (97.67%) and fourth (98.86%) subsets have yielded the best score for AWE, AMI and IIT Delhi II, respectively.
- We can notice the better performance of MDFNet on IIT Delhi II compared to AWE and AMI. This can be attributed to the nature of images in these three data-

sets. Several images in the AWE dataset are occluded by ear accessories, earrings and hair. Moreover, images of this dataset are taken from different angles. This will maximize the intra-class variance and increases the rate of potential confusion between different classes. Similarly, images of AMI dataset are taken at different poses, where ear is occluded by hair and accessories. In the contrary, compared to the AWE and AMI datasets, images in the IITD-II are well-aligned, taken in an indoor environment and manually cropped by the dataset creators, which makes possible for the proposed method to distinguish different ear classes successfully. Nevertheless, it is worth noting that the main challenge of IITD-II is the difference in illumination level of different images. To cope with this issue, we have considered normalizing the local histograms extracted from different ear regions.

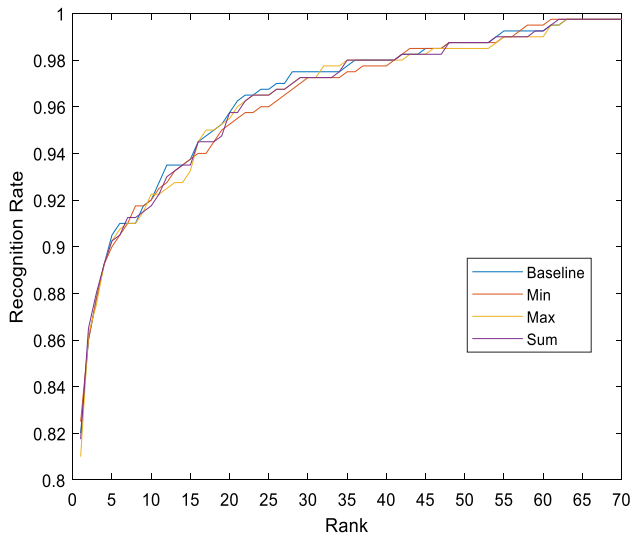
- We notice also that subset 6 records a significantly low recognition rate on the AWE dataset. In fact, the main reason behind this low rate (46%) is the block size parameter (100 × 100). In the case of AWE dataset, after performing ear alignment using CNN, the image size becomes 128 × 128. This means that features are extracted from the entire image, without considering the block-wise strategy because the block size is roughly equal to the image size (100 × 100 and 128 × 128, respectively). In such a case, the spatial relationship information is lost and key features will be pooled in a compacted feature vector, which can negatively affect the network performance. Even by increasing the number of filters and keeping the same block size (subset 4), the recognition rate still low (53%) compared to the other subsets. In addition, if we look at AMI dataset, we can notice that using the same value (i.e., 100 × 100) as block size hasn't much influenced the network performance because size of AMI images is 492 × 702, which allows extracting the features in a block-wise manner.

### 5.3.2 Evaluation of multimodal recognition system

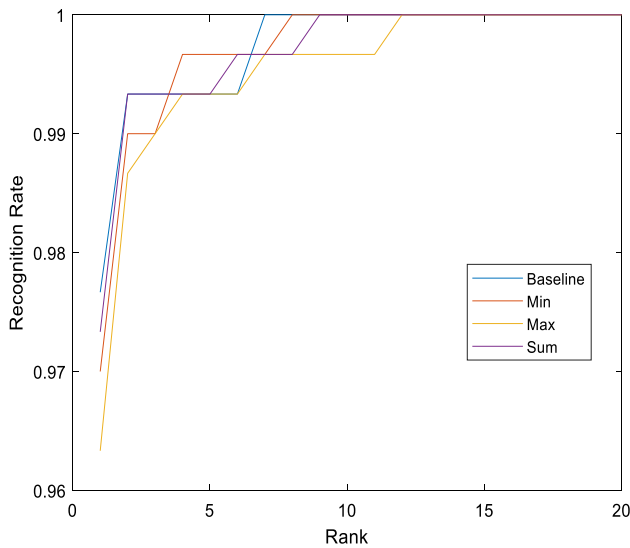
A multimodal biometric system may consist in fusing information provided by different modalities. Another form of multimodal systems considers fusing different feature extraction techniques. In this work, we fuse different features at score level to enhance the performance of unimodal systems. More precisely, we fuse filter responses-based magnitude/direction histograms at score-level using three rules namely sum, min and max. In addition, we report performance of baseline fusion in which the two histograms are combined at feature level. Table 3 reports performance of multimodal system for the three datasets (bold scores are the best). Figures 10, 11 and 12 present the CMC curves of both baseline and multimodal systems for the three datasets.

**Table 3** Performance of different fusion schemes in terms of rank-1 recognition rate

Schemes	AWE	AMI	IITD II
Baseline	82	<b>97.67</b>	98.86
Sum	81.75	97.33	98.95
Min	<b>82.5</b>	97	<b>98.96</b>
Max	81	96.33	98.77

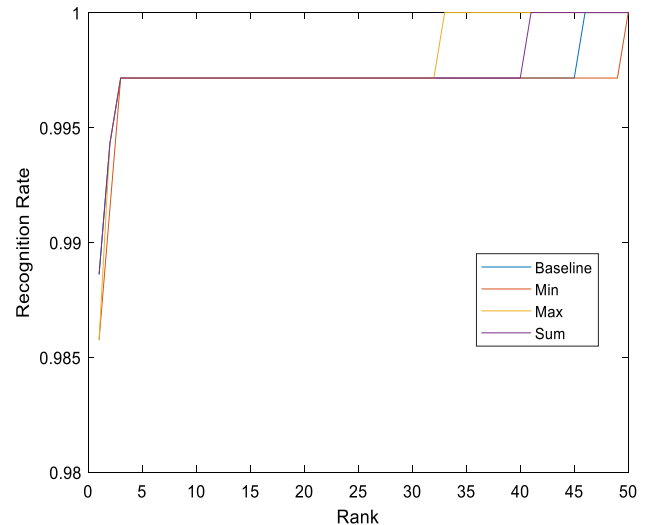


**Fig. 10** CMC curve for different fusion schemes in AWE dataset



**Fig. 11** CMC curve for different fusion schemes in AMI dataset

From Table 3, we note that by using min rule, recognition rates have been improved for the AWE and IIT Delhi



**Fig. 12** CMC curve for different fusion schemes in IITD II dataset

II. As instance, the baseline scheme has yielded 82% for the AWE, an improvement of 0.5% is obtained by considering the multimodal system based on min rule. Similarly, a slight improvement of 0.1% has been reached in the case of IIT Delhi II dataset. This confirms the strength of multimodal system compared to unimodal system. In contrast, in AMI dataset, identification rate scored by the baseline scheme was slightly higher than rates yielded by the remaining rules. From the CMC curves, we can notice that curves of different rules are, to some extent, close to each other, especially in the case of AWE dataset. However, the main difference appears in the rank value for which all rules attained 100% of performance. For instance, in the IIT Delhi II dataset, max rule has achieved 100% when the rank equals to 33, whereas, sum rule achieved the same performance when the rank equals to 41.

### 5.3.3 Comparing to relevant networks

In this experiment, we compare the performance of MDFNet with two relevant networks namely PCANet (Chan et al. 2015) and DCTNet (Ng and Teoh 2015). Although these two methods are somewhat similar, they differ in the manner used for generating filter banks. PCANet generates a data-driven filter bank as shown in Sect. 3.2, while, DCTNet, as its name indicates, considers a learning-free approach by using pre-fixed DCT filters. Another point differentiating DCTNet from PCANet is the Tied-Rank normalization, which is integrated on the top of DCTNet to make local histograms more evenly distributed. We conduct an experiment to assess the identification performance of the both networks. For a fair comparison, we consider the same parameters as in MDFNet. Table 4 shows the rank-1 recognition

**Table 4** Performance of the relevant networks

Networks	AWE	AMI	IITD II
PCANet	80	96.67	98.38
DCTNet	75	90	98.86
MDFNet	<b>82.5</b>	<b>97.67</b>	<b>98.96</b>

rate achieved by the three networks in AWE, AMI and IIT Delhi II (bold scores are the best).

The first observation we can make is that our MDFNet has outperformed the other networks in the three dataset. Specifically, the proposed method has outperformed DCTNet and PCANet by 7.5% and 2.5% in AWE, 7.67% and 1% in AMI, respectively. In fact, IIT Delhi II is relatively easy than the two other datasets, thus, slight improvements of 0.10% and 0.58% has been reached by MDFNet compared to DCTNet and PCANet, respectively. This outperformance confirms the utility of considering gradient information along with filters responses. This latter acts as features detectors, and gradient magnitude/direction provides interesting information on local ear structure. In brief, MDFNet has achieved a better performance than DCTNet and PCANet for two reasons. The first reason is that MDFNet considers additional and crucial features (i.e., gradient magnitude and direction) that can well describe the ear structure. The second reason is that local histograms are extracted by jointly considering the filter responses (which are fused in the binary hashing stage) and the gradient images (i.e., gradient magnitude and direction). This representation allows linking these two kinds of crucial information, and thus, strengthening the final descriptor.

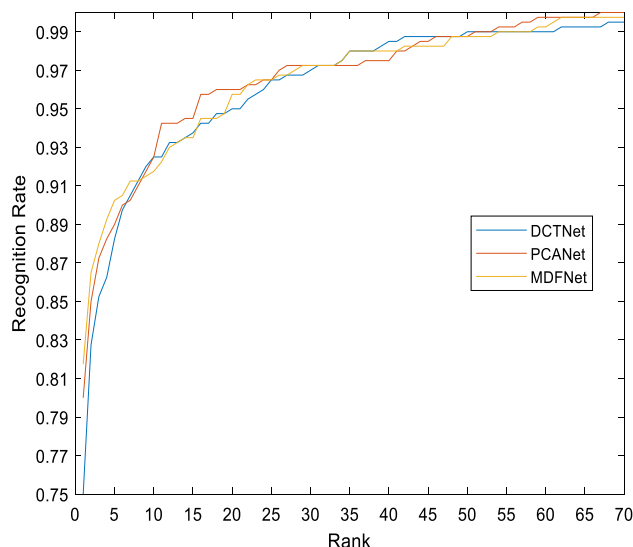
From Table 4, we remark that PCANet has significantly outperformed, by 6%, DCTNet in the two challenging datasets AWE and AMI. Although DCT and PCA-learned filters are somewhat equivalent, as argued by DCTNet authors, this experiment indicates that data-driven learned filters are more faithful than DCT filters, and can generalize better for unseen test images. Figures 13, 14 and 15 present the CMC curves for the three networks.

From these CMC curves, we can note that MDFNet still yielding better performance than the two others until, at least, rank-5. Then, they interchangeably achieved the best performance in the three datasets. In AWE, performance of the three networks is somewhat similar, but, it is different for the case of AMI and IIT Delhi II datasets.

### 5.3.4 Measuring the processing time

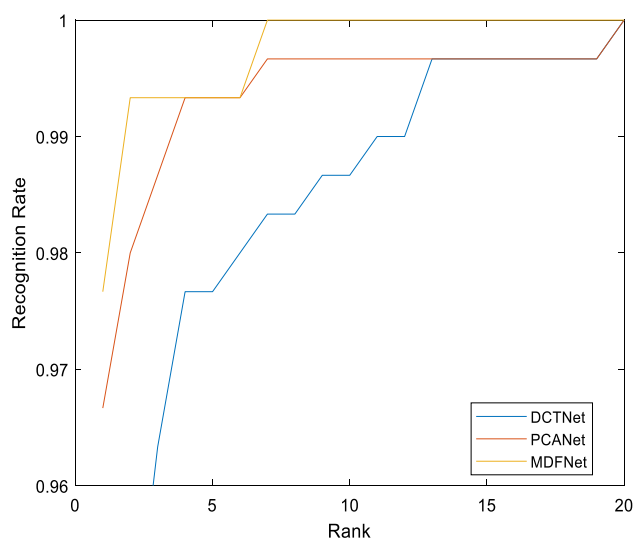
We measure the computation time required by the proposed method compared to several others that are used by the related works (Table 5).

As can be seen from Table 5, the proposed method requires 0.05 compared to 0.02 s required by both PCANet

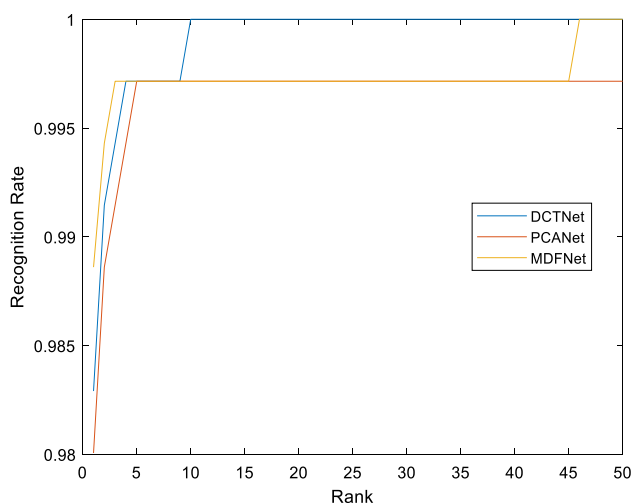


**Fig. 13** CMC curve for the relevant networks along with MDFNet in AWE dataset

and DCTNet. As a second observation, we note that Gabor filters and POEM are computationally demanding (0.22 and 0.11 s, respectively), whereas, LBP is the fastest feature due to its simplicity. We also notice that HOG, LPQ and BSIF take a reasonable time. Because of the relatively high number of stacked layers in CNN, extracting deep features has taken a significant deal of time. Knowing that MDFNet has outscored all the cited features, we can conclude that it is the best compromise between accuracy and processing time.



**Fig. 14** CMC curve for the relevant networks along with MDFNet in AMI dataset



**Fig. 15** CMC curve for the relevant networks along with MDFNet in IITD II dataset

### 5.3.5 Comparison with state of the art

To prove the effectiveness of the proposed method, we compare it with several recent and relevant state of the art methods (Table 6, where bold scores are the best). We follow the evaluation protocol that is explained in Sect. 5.2 for the three datasets namely AWE, AMI and IIT Delhi II.

From Table 6, it evidently appears that our proposed MDFNet outperforms the majority of the cited related works. As instance, a recognition rate of 82.5% has been reached compared to 78.13% and 75.6% achieved by Omara et al. (2021) and Hansley et al. (2018), respectively. Actually, the most substantial remark concerning the AWE dataset is the huge improvement that is made by MDFNet compared to some recent works. For instance, the proposed method surpassed the methods in Khaldi and Benzaoui (2020a, b) and Sajadi and Fathi (2020) by 32% and 29%, respectively. To take an idea about processing time required by the deep-based competing methods, we present a concrete example. The method in Dodge et al. (2018) has a recognition rate of 68.5% based on deep ensemble learning using different fine-tuned CNNs of varying depths. Nevertheless, such an approach requires a significant deal of processing time. For the sake of illustration, according to Table 5, ResNet101 and GoogleNet have taken 0.86 and 0.42 s (0.64 in average), respectively, compared to 0.05 taken by our method. Approximately, if three networks are used in the ensemble learning, the whole feature extraction procedure will take about 1.92 s, which is equal to the processing time of MDFNet multiplied by 38. This confirms once again that MDFNet provides a good trade-off between performance and speed. Now, if we look at AMI dataset, we notice that our proposed method has surpassed all the cited methods

**Table 5** Processing time required by each method in seconds for one image

Method	Processing time
BSIF	0.0372
HOG	0.0206
LPQ	0.0421
POEM	0.1106
LBP	0.0179
Gabor	0.2201
PCANet	0.0204
DCTNet	0.0239
GoogleNet	0.4270
ResNet101	0.8697
MDFNet	0.0573

**Table 6** Comparison with state of the art methods based on rank-1 recognition rate

Method	AWE	AMI	IITD II
(Benzaoui et al. 2017)	–	–	97.63
(Hansley et al. 2018)	75.6	–	–
(Dodge et al. 2018)	68.5	–	–
(Al Rahhal et al. 2018)	–	–	98.4
(Raghavendra et al. 2016)	–	86.36	–
(Chowdhury et al. 2018)	–	67.26	–
(Emeršič et al. 2017a; b)	49.60	–	–
(Omara et al. 2018)	–	95.5	–
(Hassaballah et al. 2020)	54.10	72.29	–
(Alshazly et al. 2019a; b)	–	94.50	–
(Alshazly et al. 2019a; b)	–	97.50	–
(Zhang et al. 2019)	50	–	–
(Priyadharshini et al. 2021)	–	96.99	–
(Sajadi and Fathi 2020)	53.5	–	97.13
(Hassaballah et al. 2019)	–	73.71	–
(Sarangi et al. 2018)	–	–	98.79
(Khaldi and Benzaoui 2020a, b)	50.53	96.00	–
(Khaldi and Benzaoui 2020a, b)	48.48	–	–
(Omara et al. 2021)	78.13	<b>97.84</b>	–
(Hamdany et al. 2021)	–	–	94
MDFNet (without alignment)	<b>60.25</b>	–	–
MDFNet	<b>82.5</b>	97.67	<b>98.96</b>

except the method in Omara et al. (2021) which has slightly outperformed our method. Similarly, by taking a look on the IIT Delhi II, we notice that the proposed method slightly outscored most of the compared works in terms of rank-1 recognition rate.

We report the performance of MDFNet on AWE without alignment to show the great importance of taking such an aspect into account. Obviously, it can be seen how much performance of MDFNet has been enhanced when aligning ear images. In particular, a huge leap of about 20% has been



**Table 7** Performance of the proposed method under occlusion

Images	AWE	AMI	IITD II
Original image	<b>82.5</b>	<b>97.67</b>	<b>98.96</b>
-1 block	79.25	97	98.58
-2 block	79	96.33	98.86
-3 block	79	95.67	98.58
-4 block	66.50	92	96.58
-5 block	66.50	90	93.16

gained. This confirms the actual need for considering this pre-processing step.

### 5.3.6 Occluded ear recognition

In unconstrained scenarios, ear may be occluded by earrings or by ear accessories, which makes very challenging to recognize it. One way to overcome such an issue is by eliminating pixels from the bottom of the ear image. The aim of this experiment is to assess the robustness of the proposed method when performing such an elimination. In particular, to perform a meaningful elimination, we repeatedly truncate one block in each time from the bottom of ear image i.e., one block in the first time, then, two blocks, three blocks and so on. Noting that, for each dataset, we have considered the block size yielding the best performance in experiment 1. We report the method performance in each iteration as shown by Table 7 (bold scores are the best).

From Table 7, we can notice the high robustness of the proposed method, which yields high recognition rates in spite of the successive cropping. We remark that identification rates slightly decreased whenever the crop size increased. What is actually surprising is the high recognition rates still attained despite significant truncation on ear images. As instance, in AWE, even though by cropping a total of 5 blocks, a high rate of 58.5% is scored. Similarly, in AMI and IITD II, although 5 blocks are entirely eliminated, recognition accuracy remains relatively high (90% and 93.16%, respectively). This suggests that top ear regions, including helix and crus of antihelix, are more distinguishing than the other ear parts. In fact, in the case of AWE dataset, ear alignment is a crucial factor in providing more robustness to the proposed method because AWE images are taken from different angles. Figure 16 illustrates different occlusion degrees.

## 6 Conclusion

In this paper, we introduced an unsupervised lightweight single layer network, which is referred to as MDFNet, for human ear recognition. To extract faithful features, MDFNet

**Fig. 16** Illustration of different occlusion degrees

jointly considers gradient magnitude and direction beside outputs from PCA-based filters. To help our method extracting features that actually reflects the rich ear structure, we have aligned ear image using landmark-based approach, which incorporates both Convolution Neural Network (CNN) and Principal Component Analysis (PCA). To evaluate the proposed method, we carried out comprehensive experiments and analysis on three public datasets namely AWE, AMI and IITD II. Experimental results revealed some noteworthy points. First, MDFNet makes a good compromise between running time and recognition accuracy. Second, ear alignment has greatly improved the performance of ear recognition, and it seems to be indispensable in the case of unconstrained scenario. Third, MDFNet has outscored several state of the art methods and showed a high robustness to occlusion. As a future direction, one can investigate other strategies for filters learning and selection. Another track is to integrate other useful information within MDFNet such as filter responses in pixel's neighborhood.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Al Rahhal MM et al (2018) A dense phase descriptor for human ear recognition. *IEEE Access* 6:11883–11887
- Alshazly H et al (2019a) Ensembles of deep learning models and transfer learning for ear recognition. *Sensors* 19(19):4139
- Alshazly H et al (2019b) Handcrafted versus CNN features for ear recognition. *Symmetry* 11(12):1493
- Barra S et al (2018) Cloud-based biometrics (biometrics as a service) for smart cities, nations, and beyond. *IEEE Cloud Comput* 5(5):92–100
- Benzaoui A et al (2017) Experiments and improvements of ear recognition based on local texture descriptors. *Opt Eng* 56(4):043109
- Chan T-H et al (2015) PCANet: a simple deep learning baseline for image classification? *IEEE Trans Image Process* 24(12):5017–5032
- Chowdhury DP et al (2018) On applicability of tunable filter bank based feature for ear biometrics: a study from constrained to unconstrained. *J Med Syst* 42(1):1–20
- Dodge S et al (2018) Unconstrained ear recognition using deep neural networks. *IET Biometrics* 7(3):207–214

- Emeřić Ž et al (2017a) Training convolutional neural networks with limited training data for ear recognition in the wild. arXiv preprint [arXiv:1711.09952](https://arxiv.org/abs/1711.09952).
- Emeřić Ž et al (2017b) Ear recognition: more than a survey. *Neuro-computing* 255:26–39
- Emeřić Ž et al (2018) Towards accessories-aware ear recognition. In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB), IEEE
- Ganapathi II et al (2018) Ear recognition in 3D using 2D curvilinear features. *IET Biometrics* 7(6):519–529
- Hamdany AHS et al (2021) Earprint recognition using deep learning technique. *Telkonnika* 19(2):432–437
- Hamrouni L et al (2021) Plant leaves recognition based on a hierarchical one-class learning scheme with convolutional auto-encoder and siamese neural network. *Symmetry* 13(9):1705
- Hansley EE et al (2018) Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biometrics* 7(3):215–223
- Hassaballah M et al (2019) Ear recognition using local binary patterns: a comparative experimental study. *Expert Syst Appl* 118:182–200
- Hassaballah M et al (2020) Robust local oriented patterns for ear recognition. *Multimed Tools Appl* 79(41):31183–31204
- Kacar U, Kirci M (2019) ScoreNet: deep cascade score level fusion for unconstrained ear recognition. *IET Biometrics* 8(2):109–120
- Kamboj A et al (2021) A comprehensive survey and deep learning-based approach for human recognition using ear biometric. *Vis Comput* 1–34
- Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: a survey. *Comput Electron Agric* 147:70–90
- Khalidi Y, Benzaoui A (2020a) A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. *Evol Syst* 12:923–934
- Khalidi Y, Benzaoui A (2020b) Region of interest synthesis using image-to-image translation for ear recognition. In: 2020b international conference on advanced aspects of software engineering (ICAASE), IEEE
- Khalidi B et al (2019) Combining colour and grey-level co-occurrence matrix features: a comparative study. *IET Image Proc* 13(9):1401–1410
- Korichi A et al (2018) Off-line Arabic handwriting recognition system based on ML-LPQ and classifiers combination. In: 2018 international conference on signal, image, vision and their applications (SIVA), IEEE
- Korichi A et al (2020) Arabic handwriting recognition: between handcrafted methods and deep learning techniques. In: 2020 21st international Arab conference on information technology (ACIT), IEEE
- Kumar A, Wu C (2012) Automated human identification using ear imaging. *Pattern Recogn* 45(3):956–968
- Kumar A et al (2021) SARS-Net: COVID-19 detection from chest x-rays by combining graph convolutional network and convolutional neural network. *Pattern Recogn* 122:108255
- Mawloud G et al (2016) Sparse coding joint decision rule for ear print recognition. *Opt Eng* 55(9):093105
- Ng CJ, Teoh ABJ (2015) DCTNet: a simple learning-free approach for face recognition. In: 2015 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), IEEE
- Omara I et al (2018) Metric learning with dynamically generated pairwise constraints for ear recognition. *Information* 9(9):215
- Omara I et al (2021) A novel approach for ear recognition: learning Mahalanobis distance features from deep CNNs. *Mach vis Appl* 32(1):1–14
- Polap D (2019) Analysis of skin marks through the use of intelligent things. *IEEE Access* 7:149355–149363
- Polap D et al (2021) Automatic ship classification for a riverside monitoring system using a cascade of artificial intelligence techniques including penalties and rewards. *ISA transactions*
- Priyadharshini RA et al (2021) A deep learning approach for person identification using ear biometrics. *Appl Intell* 51(4):2161–2172
- Raghavendra R et al (2016) Ear recognition after ear lobe surgery: a preliminary study. In: 2016 IEEE international conference on identity, security and behavior analysis (ISBA), IEEE
- Ribič M et al (2016) Influence of alignment on ear recognition: case study on AWE dataset. In: International electrotechnical and computer science conference
- Sajadi S, Fathi A (2020) Genetic algorithm based local and global spectral features extraction for ear recognition. *Expert Syst Appl* 159:113639
- Samai D et al (2018) 2d and 3d palmprint recognition using deep learning method. In: 2018 3rd international conference on pattern analysis and intelligent systems (PAIS), IEEE
- Sarangi PP et al (2018) Fusion of PHOG and LDP local descriptors for kernel-based ear biometric recognition. *Multimedia Tools and Applications* 78:9595–9623
- Sarangi PP et al (2021) A feature-level fusion based improved multi-modal biometric recognition system using ear and profile face. *J Ambient Intell Humaniz Comput* 13:1867–1898
- Zhang J et al (2019) Few-shot learning for ear recognition. In: Proceedings of the 2019 international conference on image, video and signal processing
- Zhang Y-D et al (2021) MIDCAN: A multiple input deep convolutional attention network for Covid-19 diagnosis based on chest CT and chest X-ray. *Pattern Recogn Lett* 150:8–16
- Zhou Y, Zaferiou S (2017) Deformable models of ears in-the-wild for alignment and recognition. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), IEEE

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.