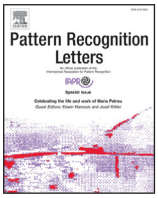
Pattern Recognition Letters 166 (2023) 143–150

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Zero-shot ear cross-dataset transfer for person recognition on mobile devices

David Freire-Obregón[a,*], Maria De Marsico[b], Paola Barra[c], Javier Lorenzo-Navarro[a], Modesto Castrillón-Santana[a]

[a] *Universidad de Las Palmas de Gran Canaria, Spain*
[b] *Sapienza Universitá di Roma, Italy*
[c] *Universitá di Napoli Parthenope, Italy*

ARTICLE INFO

Article history:
Received 10 October 2021
Revised 2 December 2022
Accepted 3 January 2023
Available online 20 January 2023

Edited by

MSC:
41A05
41A10
65D05
65D17

Keywords:
Ear recognition
Zero-shot cross-dataset transfer
Quadruplet loss

ABSTRACT

Smartphones contain personal and private data to be protected, such as everyday communications or bank accounts. Several biometric techniques have been developed to unlock smartphones, among which ear biometrics represents a natural and promising opportunity even though the ear can be used in other biometric and multi-biometric applications. A problem in generalizing research results to real-world applications is that the available ear datasets present different characteristics and some bias. This paper stems from a study about the effect of mixing multiple datasets during the training of an ear recognition system. The main contribution is the evaluation of a robust pipeline that learns to combine data from different sources and highlights the importance of pre-training encoders on auxiliary tasks. The reported experiments exploit eight diverse training datasets to demonstrate the generalization capabilities of the proposed approach. Performance evaluation includes testing with collections not seen during training and assessing zero-shot cross-dataset transfer. The results confirm that mixing different sources provides an insightful perspective on the datasets and competitive results with some existing benchmarks.

1. Introduction

Biometric recognition compares incoming templates extracted from physical or behavioral traits to stored ones to authenticate or identify an individual. Physical traits are mostly related to appearance (e.g., fingerprint, iris, face); behavioral ones reflect the user's behavior (e.g., keystroke dynamics, gait).

The research community has proposed different biometric-based techniques during the past two decades, with a relevant number focusing on mobile applications. The smartphone sales growth and the increasing amount of sensitive information stored on these devices have boosted security research by calling for reliable authentication techniques to unlock them. Traditional passwords have been used first. To this respect, a recent article by [37] states that, on average, Americans check their phones 262 times per day. This means that using a four-digit password scheme, each American will type 1048 characters per day to unlock the smartphone. However, repeated typing or, even worse, reuse [5] can expose the password, which can also be guessed or cracked. Moreover, the stronger the password, the hardest to remember.

More recently, biometric-based technologies, including iris, fingerprint, and face recognition, have gradually replaced or complemented the password-based methods [34]. Fingerprint and face recognition are widely used in smartphone unlocking, but they still can be attacked. Komkov and Petiushko [27] have proposed a method to attack the Face ID system based on a sticker placed on the forehead. On the other hand, fingerprints may not work due to dirty or sweaty fingers, and they can be quite easily reproduced. Therefore, robust biometric-based methods continue deserving attention.

The human ear is a biometric trait with several features that could be exploited in authentication. It is especially suited for smartphone protection due to its natural capture operation resembling a regular call [15]. categorized ear features into three levels: (1) the global ear appearance (overall shape, color), (2) the ear geometric structure (edges, folds, ridges, relative distances), and (3) unstructured micro features (piercings, birthmarks). Only the second and third levels provide highly discriminative biometric features, while the first is soft. Recently, deep learning ap-

[*] Corresponding author.
E-mail address: david.freire@ulpgc.es (D. Freire-Obregón).

https://doi.org/10.1016/j.patrec.2023.01.012
0167-8655/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

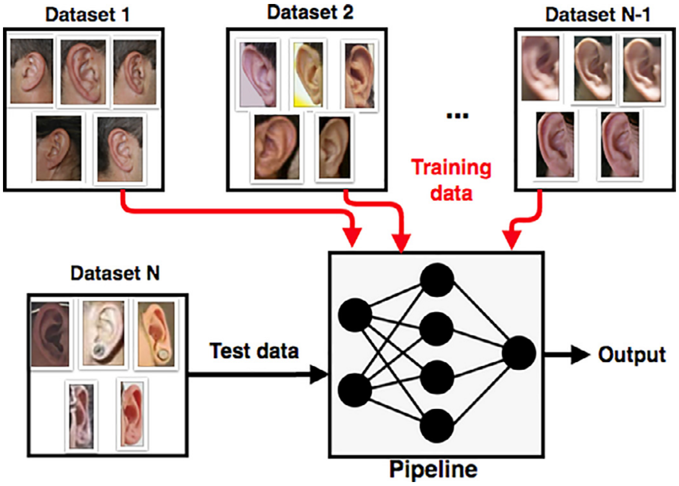


Fig. 1. Zero-shot ear cross-dataset transfer for person authentication. The proposed experiments analyze the person authentication performance with a model tested on datasets not seen during training. Top: a generic set of collections used for training. Bottom left: the unseen test dataset. Bottom right: the proposed pipeline to efficiently recognize a subject in a single forward pass.

proaches [2,3,20,30] have achieved significant progress compared with handcrafted features [1,9,28].

To the best of our knowledge, even the present proposals that use multiple datasets do so without mixing them, i.e., each method is evaluated separately on each dataset. However, generalizing those research results to real-world applications is almost impossible since the available ear datasets present different characteristics. When passing from cross-fold to cross-dataset evaluation, which is closer to the actual application's use, performance drops dramatically. This work advances toward cross-dataset assessment by proposing a novel ear recognition approach whose end-to-end model relies on a set of auxiliary pre-trained encoders. Its novelty relies on something other than designing a brand-new architecture to achieve higher SOTA results. Instead, it leverages a popular pre-trained backbone to set up an effective pipeline to tackle the cross-dataset evaluation challenge. The experiments test whether a mixed training strategy can provide better generalizable results using already available models. The main contributions are:

- the experimental zero-shot cross-dataset transfer protocol, training a model on a set of ear datasets and testing its performance on different unseen ones (Fig. 1); the goal is a more reliable estimate of "real-world" performance than training and testing on subsets of a single, and often biased, data collection [32,36];

- competitive cross-dataset results, namely more than a 70% *rank-5* on some of the collections considered ;

- a pipeline built on top of a leveraged pre-trained backbone, achieving highly competitive ear recognition and improving some state-of-the-art (SOTA) baselines (see Fig. 2 in Section 4), despite the testing on unseen collections.

2. Related work

The work by Iannarelli [25] provides a pioneer study about the viability of ear biometrics. Since then, several proposals have dealt with it, even in unconstrained settings [6]. Early works mainly used 2D ear images. A survey by Emeršič et al. [19] classifies the visual recognition techniques into geometric, holistic, local, and hybrid. Geometric techniques exploit the ear geometrical characteristics as in [7]. Holistic approaches consider the ear as a whole to extract features representing global properties, such as the force fields introduced in [24]. Local approaches extract features from local areas of an image with recognition purposes, as in the proposal by Bustard and Nixon [9]. Finally, hybrid approaches combine elements from the previously described categories, as in [28].

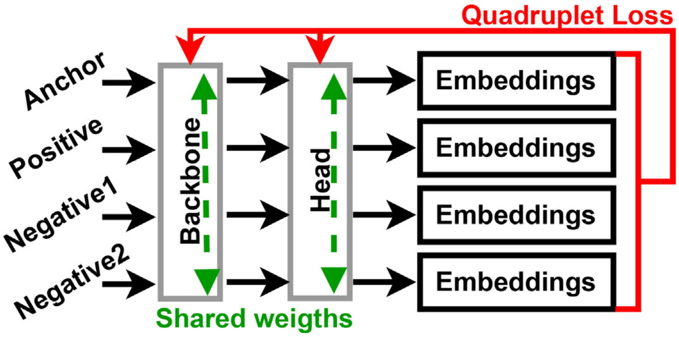


Fig. 2. The considered training scheme. The devised training process comprises three main modules: a backbone to extract features followed by a head, while the final embeddings are used to compute the quadruplet loss function.

2D images present substantial limitations due to the ear 3D structure, suffering from occlusions, illumination conditions, and camera points of view. Compared with the 2D data, the 3D data contains richer information about the ear shape and is more robust to these factors. The works by Chen and Bhanu [11] and [39] are among the first related proposals. Computational complexity and processing times are the main issues for the real-time application of 3D techniques. It is important to mention that most early works tested their approach on a single collection, such as the face-profile subset of the XM2VTS dataset exploited in [4].

Only a few works apply deep learning for ear recognition. Galdámez et al. [20] use a 3-layers convolutional neural network (CNN) and report accurate rates on their collection. Images are resized to 64 × 64 to fit into the shallow CNN. More recently, Priyadharshini et al. [30] exploits a 6-layers CNN for ear recognition. In this case, experiments rely on two publicly available datasets: the IITD-II dataset and the AMI dataset [21,29]. More recently, Alshazly et al. [2] has achieved high performance on the AMI dataset and the CVLE dataset [16] by combining image augmentation and fine-tuning of a pre-trained neural network. Precisely, the approach gathers the probabilities of distinct pre-trained networks to achieve noticeable accuracy rates on the EarVN1.0 dataset [23].

Regarding evaluation, the experiments in the first works on ear recognition mainly rely on in-house collections, which are not all publicly available [10]. proposes one of the earliest ear recognition datasets (CP). However, it has not been used much in literature due to its limitations in pose variations and subject identities (17 subjects only). Some approaches use image collections not explicitly aimed at ear recognition, e.g., XM2VTS, a generic multi-modal face dataset. Significant progress has been made in the past few years to remove these limitations. The Ear Recognition Laboratory at the University of Science & Technology of Beijing (USTB) introduced four distinct ear datasets [14]. The first two only contain grayscale cropped ear images, while the other two include whole head profile shots. Contrary to the CP dataset, the USTB ear datasets provide illumination variations, occlusions, and pitch angles in the range $[-30^\circ, +30^\circ]$ making it more challenging. The following section describes different datasets that have been used in this work.

It is reasonable to argue that highly accurate deep-learning models for ear recognition can operate on a relatively wide and unconstrained range of ear samples. However, the need for large-scale data in various conditions can limit performance generalizability. Commonly used collections feature homogeneous ear layouts in pose, illumination, and occlusions. Therefore models trained and tested on a single dataset might be biased and prone to fail in a different context. To our knowledge, the evaluation in

Table 1
The ear datasets considered in this work in chronological order. The last column shows the resolution variation.

Dataset	Authors	Source	# Subjects	# Images	Color	Ethnic	Pose	Res. Var.
AMI	[21]	Manually	100	700	RGB	Caucasian	Profile	None
IITD-I	[29]	Manually	125	493	GRAY	Indian	Profile	None
IITD-II	[29]	Manually	221	793	GRAY	Indian	Profile	None
AWE	[19]	Internet	100	1000	RGB	Varied	Varied	Strong
CVLE	[16]	Internet	16	804	RGB	Caucasian	Varied	Strong
AWEx	[18]	Internet	220	2200	RGB	Varied	Varied	Strong
BIPLab	[1]	Manually	100	300	RGB	Caucasian	Profile	Low
EarVN1.0	[23]	Internet	164	28412	RGB	Asian	Varied	Strong

the mentioned works did not mix different collections and relied on each dataset separately.

3. Considered datasets

The available ear datasets generally differ in the source (captured images/image web crawler), capture pose (front/profile ear), number of participants, dataset size, ethnicity, and camera settings. Each collection has its unique characteristics and/or limitations, possibly causing experimental bias [36]. Literature testifies that training and testing on partitions of the same ear dataset can lead to a robust performance [2,3,23]. However, it may lack generalization capabilities to unseen data with different characteristics (camera configuration, ear pose, environment). On the other hand, acquiring a new large-scale ear dataset from scratch is hard, too, whereas a large one collected from the Internet may raise legal (e.g., copyright) or social problems (e.g., privacy). This work instead proposes to train and test on a mix of unrelated datasets. This approach has been proposed before in different contexts. For instance, Ranftl et al. [32] applies it for monocular depth estimation. The authors train and test on a different collection of datasets to analyze the performance of each combination and obtain a robust model able to generalize. Table 1 details the collections used in this work.

AMI [21] contains seven noiseless images per subject (six right ear and one left ear images) collected under fixed illumination conditions using both a 135 mm and 200 mm focal length. Poses barely vary in yaw but severely vary in pitch (around 40–45°).

The IITD-I and IITD-II are two distinct gray-scale collections created by Kumar and Wu [29]. They were both captured under the same indoor illumination and with a fixed camera position (approximately the same profile angle). The main difference between them relies on the pre-processing of the IITD-II images: all ears are cropped, centered, and aligned. The number of images per user varies from 3 to 6 samples.

The AWE dataset introduces the notion of *ear images captured in the wild* [19]. collected all the celebrity images from the Internet with a wide range of image conditions. Each subject has ten images, and the image size severely varies, from 15 × 29 pixels for the smallest sample to 473 × 1022 pixels for the largest sample. Later on [16], presented the CVLE dataset. This dataset is smaller in terms of subjects (just 16) but has a higher variance in the number of samples (from 18 to 93 samples per subject). The same authors created the AWE_Ext dataset by mixing the AWE dataset, the CVLE dataset, and 2200 new images of 220 subjects. The three subdatasets are disjoint. In order to compare the present work with previous ones, we have split AWE_Ext into its three previously described components (see Table 1), so that AWE_Ext only refers to the new 2200 samples. Also, AWE_Ext new images were collected automatically from the Internet and manually screened to ensure that ears were indeed present in all of them. The AWE_Ext subset is highly diverse, with ten samples per subject and extreme variations in the image size. The UERC dataset was also presented by Emeršič et al. [17]. The experiments here will not consider this collection because it highly correlates to the AWE, CVLE, and AWE_Ext datasets.

Recently, [1] proposed the new dataset BIPLab. It includes 300 images of 100 distinct participants. Contrary to other manual collections, images were taken under uncontrolled illumination and with a non-fixed camera position. The authors tried to simulate the ear portion captured during a call to cover approximately 90% of the image. Samples can be blurred, and ear poses barely vary in yaw and pitch.

The last considered dataset is the EarVN1.0 [23]. This collection was gathered from the Internet and provided images of both ears per person under unconstrained conditions. Therefore, it exhibits significant variations of pose, scale, and illumination. This collection presents a high variance in the number of samples, between 107 and 300.

4. Description of the proposal

4.1. The considered training scheme

Fig. 2 depicts the considered training scheme corresponding to a quadruplet network [12]. This network is made up of four branches with shared parameters that are fed with four samples (quadruplet) named anchor ($\mathbf{x}$), positive ($\mathbf{x}^+$), negative 1 ($\mathbf{x}_1^-$) and negative 2 ($\mathbf{x}_2^-$). The network aims to find an embedding function $f(.)$, that reduces the intra-class dispersion and increases the inter-class margin. A pre-trained convolutional neural network realizes the embedding function in our proposal, with the last set of layers modified to fit our problem at hand. Section 4.2 extensively studies and compares different CNN architectures, leading to choosing VGG16. In this work, VGG16 is modified, replacing the last maxpool layer with the head.

The five-layers head encoder in the second step transforms the previously computed features into a more specific and smaller set of features. The first layer applies a global average pooling operation to the pre-trained encoder output. Then, the data passes through two dense layers (512 units each) separated by a batch normalization layer. A final dense layer embeds the information into a space of 20 elements. This allows the next module to compute the distance between embeddings.

The proposed network is fine-tuned using the quadruplet loss function [12], QL hereafter, defined as follows:

$$QL(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \mathbf{x}_2^-) = \max(g(\mathbf{x}, \mathbf{x}^+) - g(\mathbf{x}, \mathbf{x}_1^-) + m_1, 0) + \max(g(\mathbf{x}, \mathbf{x}^+) - g(\mathbf{x}_1^-, \mathbf{x}_2^-) + m_2, 0) \quad (1)$$

where the L2-distance is used to compute the loss function, $g(a, b) = ||f(\mathbf{a}) - f(\mathbf{b})||_2^2$, and $f(.)$ is the embedding function. The margin parameters are denoted as m_1 and m_2. Accordingly, similar objects are closer while dissimilar objects are pushed away from each other. The extra negative sample distance ($g(\mathbf{x}_1^-, \mathbf{x}_2^-)$) added to the loss in Eq. (1) helps the network to learn a better-generalized rule in terms of similarity. Similarly, Proença et al. [31] have used the QL for an identification problem. Their work uses a 128-dimensional space obtained by combining several datasets with more than 700K identities. They show that results stabilize for dimensions larger than 128. However, this research deals with much fewer identities. A grid search using some of the considered datasets has shown that the results stabilize for a 20-dimensional space.

Once the model is adequately trained, at the recognition time it will generate an embedding vector for each input sample and compare it with the embedding vectors stored in the gallery during the enrollment.

4.2. The adopted experimental protocol

Sample mining strategy. When using QL the input of the proposed training pipeline is a quadruplet. A careful selection of quadruplet samples able to exemplify a wide range of variations is a challenging step [12]. introduces a margin-based online *hard* negative mining to select *hard* samples to train the model. Generally, the *hardness* of a sample depends on how much loss it will generate. *Easy* samples barely generate a loss value, if any, because the positive sample is very close to the anchor, whereas the negative sample/s are far from each other. In contrast, *hard* samples generate a high loss value because the positive sample is far from the anchor and the negative sample/s are close to each other. Contrary to easy samples, hard samples help the model to reduce the loss aggressively. The proposed protocol entails loading batches of 128 sample quadruplets, as shown in Fig. 1, where half of the samples are random items (hard or not) and the other half are hard ones. It has been experimentally verified that this strategy allows the model to converge on a regular basis. In addition, samples in the same quadruplet come from the same dataset, but quadruplets can be drawn from different datasets.

Data augmentation. Most considered collections provide a too low number of samples per subject. For instance, the BIPLab dataset collects three images per subject, while during training at least four images (two from the same subject) are necessary to create a quadruplet sample. In these cases, a data augmentation procedure ensured 100 samples per subject, except for EarVN1.0 that already contains from 107 to 300 images per subject. The applied data augmentation transformations include random brightness, random contrast, motion blur, horizontal flip, shift, scale, and rotate [8]. Augmented subsets are only used for training.

Backbone comparison and final choice. The comparison involved several pre-trained encoders: VGGNet [33], InceptionV3 [35], ResNet [22] and ResNeXt [38]. All those backbones were trained on the ImageNet dataset [13] considering 1000 different classes. The pipeline in Fig. 2 combined each pre-trained encoder with a trainable encoder, using Adam optimizer [26] with a learning rate of 10^{-5} and the decay rate by a factor of 0.4. Fig. 3 shows the results on five representative datasets out of those in Section 2. It shows as baseline the worst performance that InceptionV3 achieved on each collection, and the performance of the other backbones as a relative improvement with respect to it. Even though the performance always depended on the dataset, VGG16 outperformed any other pre-trained model counterpart if considering the overall results. For instance, ResNext101 equals VGG16 performance on IITD-II. However, VGG16 outperforms ResNext101 when any other collection is considered. The performance difference is significant when considering a more challenging dataset, as a significant boost is achieved with in-the-wild datasets AWE and CVLE (20% on average). It is reasonable to assume that, since these results hold for the single datasets, it is appropriate to choose VGG16 in our approach.

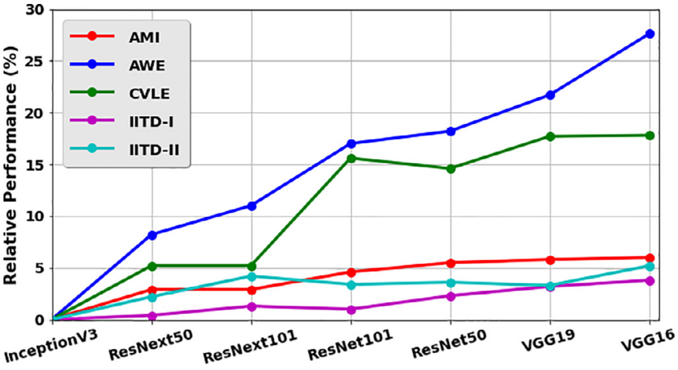


Fig. 3. Relative *rank-1* performance of backbones across different datasets with respect to InceptionV3.

Table 2
Recognition Error when fine-tuning on (single) different training collections. Lower is better. The best cross-dataset performance for each testing collection (columns) is bold, second best is underlined.

Train\Test	AWE	CVLE	IITD-II	BIPLab
AMI	95.0	75.0	74.7	60.0
IITD-I	97.0	87.5	**66.1**	61.0
AWE_Ext	**92.0**	**62.5**	71.5	**57.0**
EarVN1.0	96.0	**62.5**	77.8	82.0
In-dataset LBP_Base	88.0	81.3	14.5	28.0
In-dataset VGG_Base	94.0	87.5	43.9	47.0

Table 3
The different datasets combinations used for training.

	AWE_Ext	AMI	IITD-I	EarVN1.0
MIX 1	✓	✓		
MIX 2	✓	✓	✓	
MIX 3	✓	✓	✓	✓

5. Experimental evaluation

Similarly to [32], the collections were partitioned into training and test datasets to test the performance on the most cited datasets in literature, collected in different conditions. After an extensive study of the selected backbone VGG16 on each collection in Table 1 (see Fig. 3), we identified similar distributions of conditions for training and testing: captured in the wild (AWE_Ext and EarVN1.0 for training, AWE and CVLE for testing), in a controlled environment (AMI and IITD-I for training, IITD-II for testing), and a mixed environment (BIPLab testing only). The experiments assessed different mixtures of training conditions reflected by the combination of different corresponding datasets (see Table 3).

Data are imbalanced in some collections, so the reported performance considers the average of ten random splits to generate the gallery and the probe, with a sample per user in both sets. L2-distance (see Section 4) measures the distance between probe/gallery embeddings pairs. Distances are used to order the list of gallery embeddings for each probe (closed set identification). Two baseline algorithms are further tested. The implementations were included in the participants' starter kit for the Unconstrained Ear Recognition Challenge - UERC 2019: (i) an LBP-based approach (*In-dataset LBP_Base* hereafter), and (ii) a CNN-based model [17] built around the VGG-16 architecture (*In-dataset VGG_Base* hereafter). Regarding the first one, each test sample's feature vectors are computed without any training according to the hand-crafted LBP features. Since histograms represent them, they are compared using the Bhattacharyya distance. Regarding the second baseline, the embeddings for test samples are computed using the pre-trained VGG model on ImageNet. Then, Euclidean distance is used to compare the VGG embeddings. This process may lead to a non-generalizable set of embeddings the dataset's intrinsic features may bias [36]. The performance is reported in terms of the *RecognitionError* (*RE*), intended as 1 - *Rank-1*, where Rank-1 is the percentage of probes for which the correct identity was returned in the first position of the ordered list.

Table 2 reports the results of experiments with a single training dataset (rows) and a single (unseen) testing dataset (columns), i.e., cross-dataset evaluation when the collection used for training is different (with possibly very different characteristics) from the one used for testing. Worse results are expected under these conditions than splitting train and test sets from a single collection. These tables show the generalizability of training on the single datasets, which dataset is better for training and in which conditions, which dataset is worse, and the performance differences. When training with AWE_Ext, the best performance is achieved with AWE, CVLE, and BIPLab due to the large variability of samples captured under different angles that it includes. In most experiments, AMI provides good results. Due to similar characteristics of IITD-I and IITD-II, the former outperforms any other training dataset when testing on the latter. Likewise, IITD-I performed poorly when any other dataset but IITD-II was tested due to the specificity of their samples. Finally, AMI provides positive results in most experiments despite not being the best in any experiment. Regarding the considered baselines, the cross-dataset experiments also provide compelling results when testing on a wild dataset. However, baselines are pretty robust on indoor datasets, especially the In-dataset LBP_Base. This effect can be seen in Fig. 4, which better details the above considerations by showing a set of Cumulative Match Characteristic (CMC) curves for each test dataset. The top pair of plots corresponds to less controlled datasets (AWE and CVLE), while the bottom pair to more controlled conditions (IITD-II and BIPlab). Different colors correspond to different single (cross-)training datasets, and black colors are baselines (solid black is *In-dataset LBP_base*, dashed black is *In-dataset VGG16_base*). In the cross-dataset experiment, performance barely improves baselines on BIPLab and does not improve the LBP-based baseline on IITD-II. This may be due to the robustness of these baselines due to the intrinsic features of the indoor datasets. On the other hand, baselines on wild collections, i.e., AWE and CVLE, do not exhibit the same behavior due to data variability.

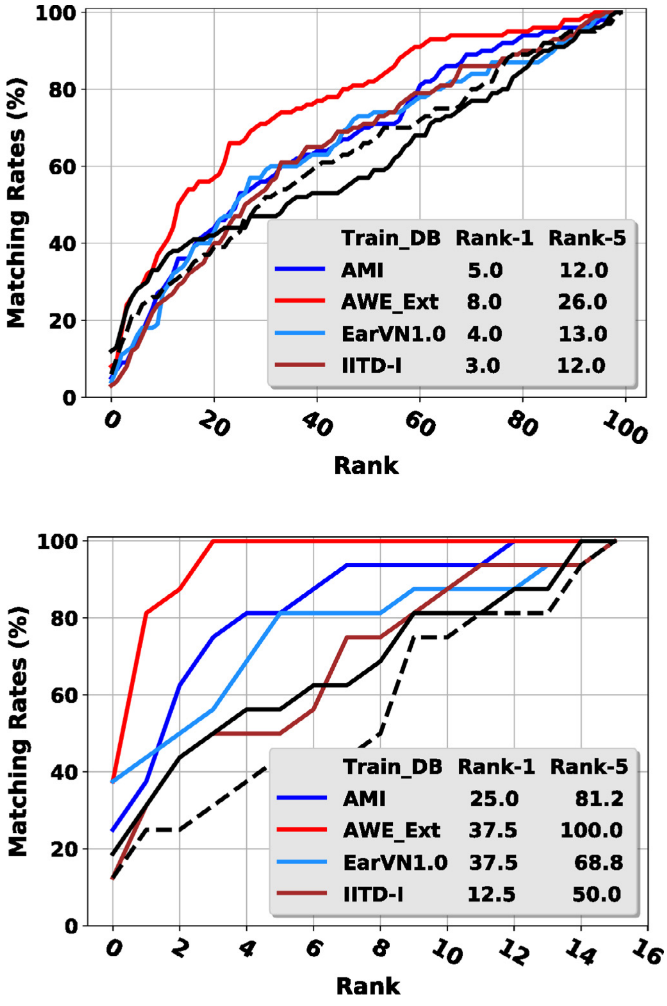


Fig. 4. CMC curves when fine-tuning on different (single) training collections. Black solid line is for *In-dataset LBP_Base*, whereas black dashed line is for *In-dataset VGG_Base*.

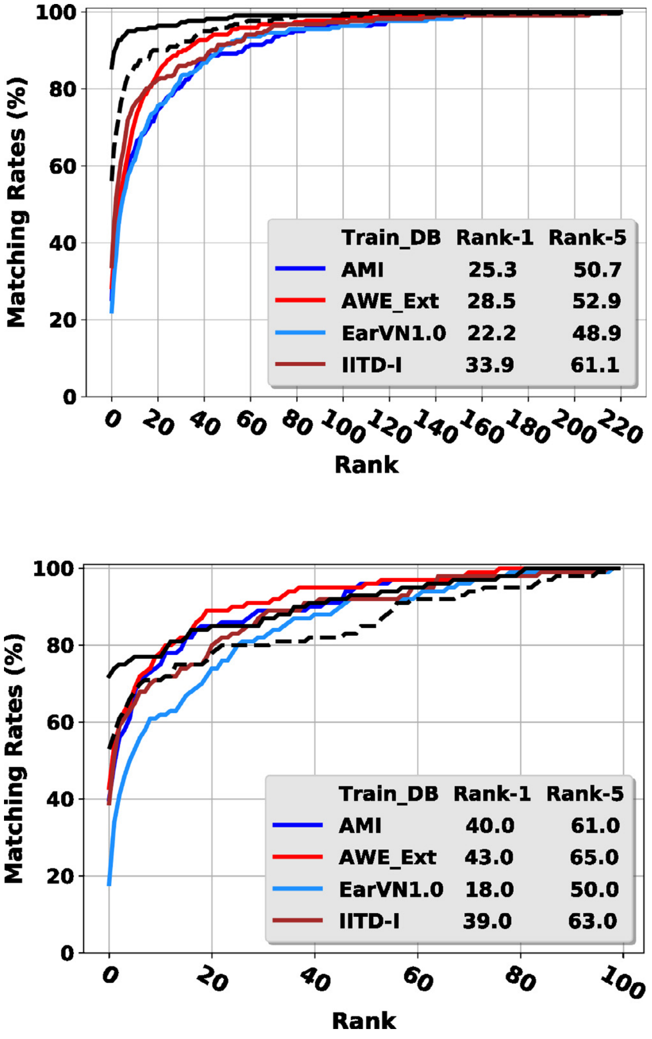


Fig. 4. Continued

Subsequent experiments adopted mixed training collections to analyze mixture performance on the (single) test datasets. Table 3 summarizes the training combinations. Test datasets were never used for fine-tuning. AWE_Ext is the compared baseline single training dataset, being the most promising in Table 2. For the same reason, it is included in all the mixed collections. In this regard, Table 3 shows that we have adopted an incremental approach when generating the mixes, similar to [32]. Therefore, MIX1 combines a wild dataset (AWE_Ext) and the most stable indoor dataset (AMI), MIX2 includes all the considered datasets in MIX1 and the

Table 4
Absolute performance (Recognition Error) when mixing the training datasets against the AWE_Ext baseline. Lower is better. The best cross-dataset performance is bold, second best is underlined.

Train\Test	AWE	CVLE	IITD-II	BIPLab
AWE_Ext	92.0	62.5	71.5	57.0
MIX 1	93.0	68.8	58.4	51.0
MIX 2	**86.0**	**50.0**	**57.5**	48.0
MIX 3	90.0	75.2	59.7	**45.0**
In-dataset LBP_Base	88.0	81.3	14.5	28.0
In-dataset VGG_Base	94.0	87.5	43.9	47.0

best dataset on IITD-II (IITD-I), and MIX3 combines all the considered training datasets. The goal is twofold. First, this is done to take advantage of the most effective training samples. Secondly, it can further highlight the improvement that can be achieved over even the best training collection when used alone.

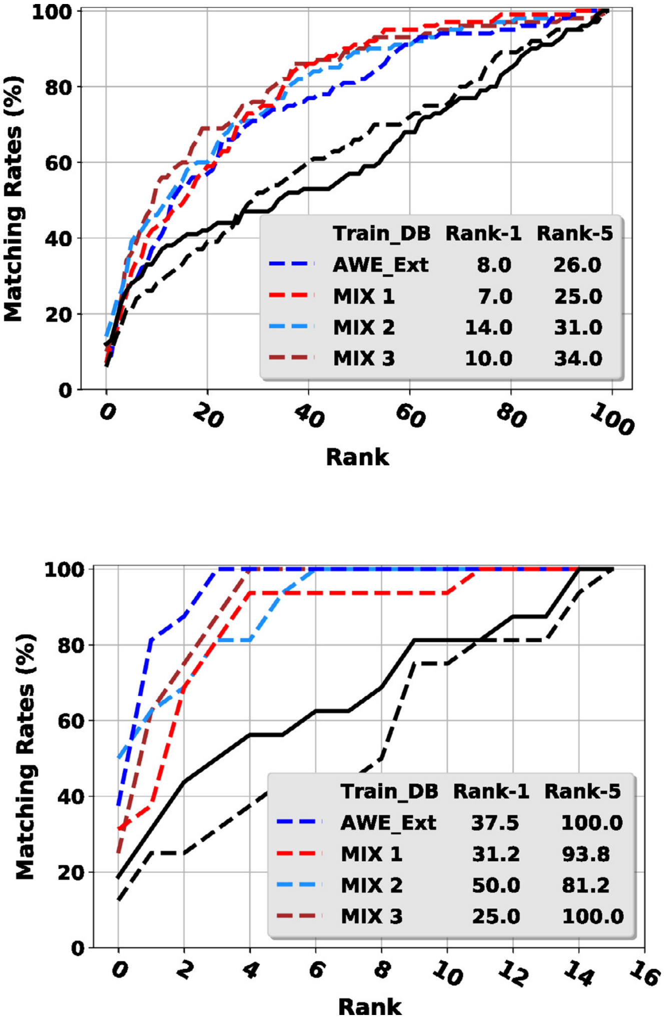
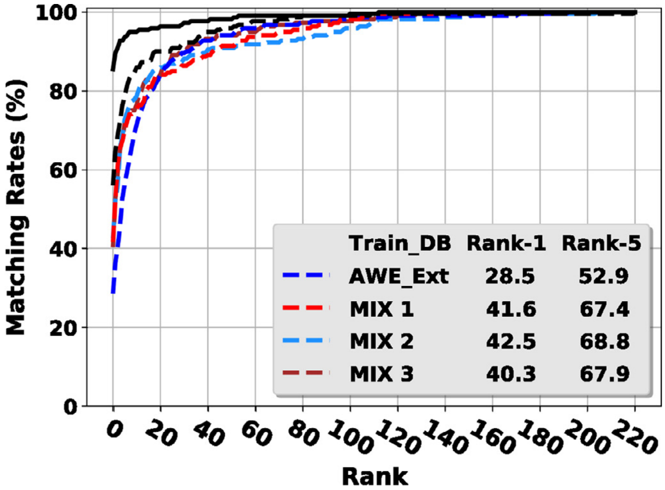


Fig. 5. CMC curves with mixed training datasets. Black solid line is for *In-dataset LBP_Base*, whereas black dashed line is for *In-dataset VGG_Base*.



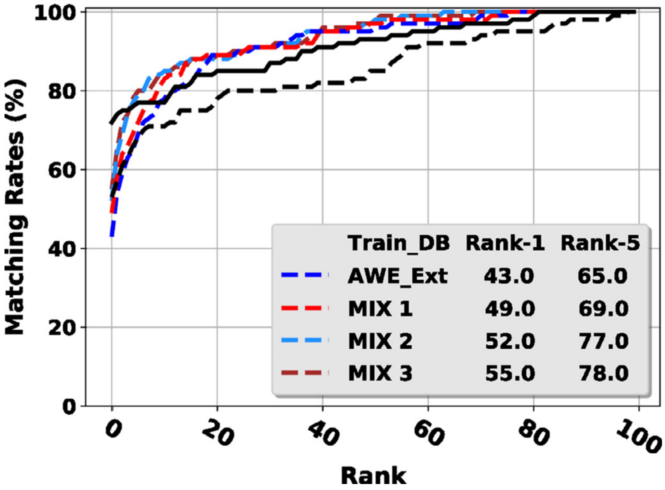


Fig. 5. Continued

Table 5
Relative performance (Recognition Error) when mixing the training datasets against the AWE_Ext baseline (top row). The best performance is bold, second best is underlined.

	AWE	CVLE	IITD-II	BIPLab	Mean
AWE_Ext	92.0	62.5	71.5	57.0	-
MIX 1	-1.1%	-10%	18.3%	10.5%	4.4%
MIX 2	**6.5%**	**20%**	**19.6%**	15.8%	15.5%
MIX 3	2.2%	-20%	16.5%	**21.1%**	5%

Tables 4 and 5 confirm the improvement of the mixed collections for each tested dataset. Considering the In-dataset baselines, Table 4 shows a notable improvement over the non-mixed experiment described previously. MIX2 outperforms both In-dataset baselines when wild datasets are tested and one baseline (In-dataset VGG_Base) when the BIPLab dataset is tested. Table 5 shows the relative performance over the AWE_Ext baseline. Several interesting insights can be inferred from this table. First, adding an indoor dataset (AMI) to the AWE_Ext dataset notably improves the performance over the tested indoor collections. Second, significant variations in performance must be seen in perspective due to the high variance number of subjects between test datasets. For instance, a single correctly classified sample on CVLE implies a 6.25% *rank-1* improvement. In contrast, a minor improvement under the same circumstance can be achieved for BIPLab (1%), AWE(1%), and IITD-II (0.45%) due to the higher number of subjects assessed on these collections. Finally, MIX2 provides better results because adding IITD-I boosted the performance when testing on IITD-II.

The comparison of Figs. 4 and 5 testifies the different and much more stable, therefore generalizable, performance measures obtained using a mixture of training collections. The gap between the experiment's CMC curves and the In-dataset baselines is higher for all collections, being particularly better on both wild collections (AWE and CVLE). Again, it is worth underlining that both result sets belong to cross-dataset evaluation. As we previously stated, testing on indoor datasets, i.e., IITD-II and BIPLab, has no improvements over baselines in terms of net *rank-1* CMC. However, it is worth underlining that the mixed training results stabilize at the best CMC curve on BIPLab. Another interesting intuition from Fig. 5 is that adding datasets does not improve rank-1 performance unconditionally during mixing but achieves a better overall curve. For instance, see MIX3 on the AWE dataset.

6. Conclusion

Ear recognition is a feasible biometric method and appears especially viable for smartphone authentication. However, a robust and generalizable assessment of the recognition methods is needed. Many deep learning approaches have been recently using ear datasets to outperform previous proposals. This work evaluated the robustness and generality of different models by applying the zero-shot cross-dataset transfer. Interestingly, our findings can be summarized as follows. (I) When performing cross-dataset experiments without mixing collections, in-the-wild datasets provide better generalization when used for training than those acquired in a controlled environment (see AWE_Ext on Table 2). This meets preliminary expectations since in-the-wild data include a wider variety of distortions and their more realistic combination, therefore feeding a more robust and generalizable model. (II) In the same conditions, datasets in controlled environments provide a more robust intra-dataset baseline than those created in the wild (see *In-dataset LBP_Base* and *In-dataset VGG_Base* on BIPLab and IITD-II in Table 2 and Fig. 4). Of course, this may be caused by similar, possibly equally biased conditions. (III) The zero-shot cross-dataset pipeline has a more limited impact on datasets with a solid intra-dataset baseline (see IITD-II on Figs. 4 and 5). Again, this may be caused by the fact that in intra-dataset baselines, the model computed during training is applied to test data with similar characteristics and possibly similar bias. IV) The intra-dataset baselines can be easily beaten on wilder datasets (see *In-dataset LBP_Base* and *In-dataset VGG_Base* on AWE and CVLE). This further demonstrates that the higher the variability of training data, the better the performance even concerning using a single, more realistic training collection. (V) Mixing datasets during training provides better results than just crossing datasets without mixing them (see Figs. 4 and 5). This is quite an intuitive and expected outcome since training with a single collection and testing on a different one causes to apply a model built on data with possibly very different characteristics. (VI) Finally, mixtures of datasets collected in a wilder environment better support generalization when applying zero-shot cross-dataset transfer than mixtures of controlled-environment collections. This may suggest low generalizability to new data or that some datasets may be more biased than others due to a lack of enough realistic variations. The consequence is that results achieved on these collections are seldom widely generalizable through mixing them.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No data was used for the research described in the article.

Acknowledgments

This work is partially funded by the ULPGC under projects ULPGC2018-08 and CEI-2018-4, by ACIISI-Gobierno de Canarias and European FEDER Funds Grant EIS 2021 04, the Spanish Ministry of Economy and Competitiveness (MINECO) under project RTI2018-093337-B-I00, and the Ministry of Science and Innovation under project PID2021-122402OB-C22, and by the ACIISI-Gobierno de Canarias and European FEDER funds under projects ProID2020010024, ProID2021010012 and ULPGC Facilities Net and Grant EIS 2021 04. We want to thank Professor Emeršič for providing us with the code of the baseline algorithms.

References

[1] A.F. Abate, M. Nappi, S. Ricciardi, I-am: implicitly authenticate me-person authentication on mobile devices through ear shape and arm gesture, IEEE Trans. Syst. Man Cybern. Syst. 49 (2019) 469–481.
[2] H. Alshazly, C. Linse, E. Barth, T. Martinetz, Handcrafted versus cnn features for ear recognition, Symmetry 11 (2019).
[3] H. Alshazly, C. Linse, E. Barth, T. Martinetz, Deep convolutional neural networks for unconstrained ear recognition, IEEE Access 8 (2020) 170295–170310.
[4] B. Arbab-Zavar, M.S. Nixon, D.J. Hurley, On model-based analysis of ear biometrics, in: Proceedings of the 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–5.
[5] D.V. Bailey, M. Dürmuth, C. Paar, Statistics on password re-use and adaptive strength for financial accounts, in: Proceedings of the International Conference on Security and Cryptography for Networks, Springer, 2014, pp. 218–235.
[6] S. Barra, M. De Marsico, M. Nappi, D. Riccio, Unconstrained ear processing: what is possible and what must be done, in: Signal and Image Processing for Biometrics, Springer, 2014, pp. 129–190.
[7] M. Burge, W. Burger, Ear biometrics, in: A.K. Jain, R. Bolle, S. Pankanti (Eds.), Biometrics: Personal Identification in Networked Society, Kluwer Academic Publishers, 1999, pp. 273–285.
[8] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: fast and flexible image augmentations, Information 11 (2020) 1–20.
[9] J.D. Bustard, M.S. Nixon, Toward unconstrained ear recognition from two-dimensional images, IEEE Trans. Syst. Man Cybern. A Syst. Hum. 40 (2010) 486–494.
[10] M. Carreira-Perpinan, Compression Neural Networks for Feature Extraction: Application to Human Recognition from ear Images, Universidad Politcnica de Madrid, 1995.
[11] H. Chen, B. Bhanu, Human ear recognition in 3d, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 718–737.
[12] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[13] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[14] E.R. Laboratory, Introduction to USTB Ear Image Databases, University of Science and Technology, Beijing, 2002.
[15] S. El-Naggar, A. Abaza, T. Bourlai, On a taxonomy of ear features, in: Proceedings of the IEEE IEEE Symposium on Technologies for Homeland Security, 2016, pp. 1–6.
[16] v. Emeršič, L.L. Gabriel, V. Štruc, P. Peer, Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation, IET Biom. 7 (2018) 175–184.
[17] v. Emeršič, S.V. A. Kumar, B.S. Harish, W. Gutfeter, A. Pacut, E. Hansley, M. Pamplona Segundo, S. Sarkar, H. Park, V. Štruc, The unconstrained ear recognition challenge 2019, in: Proceedings of the International Conference on Biometrics (ICB), 2019, pp. 1–5.
[18] v. Emeršič, B. Meden, P. Peer, V. Štruc, Evaluation and analysis of ear recognition models: performance, complexity and resource requirements, Neural Comput. Appl. 32 (2018) 1–16.
[19] v. Emeršič, V. Štruc, P. Peer, Ear recognition: more than a survey, Neurocomputing 255 (2017) 26–39. Bioinspired Intelligence for machine learning.
[20] P.L. Galdámez, W. Raveane, A. González Arrieta, A brief review of the ear recognition process using deep neural networks, J. Appl. Logic 24 (2017) 62–70. SI:SOCO14
[21] E. González-Sánchez, Biometria de la Oreja, Universidad de Las Palmas de Gran Canaria, 2008. PhD dissertation
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[23] V.T. Hoang, Earvn1.0: a new large-scale ear images dataset in the wild, Data Br. 27 (2019) 104630.
[24] D. Hurley, M. Nixon, J. Carter, Automatic ear recognition by force field transformations, in: IEE Colloquium on Visual Biometrics (Ref.No. 2000/018), 2000, pp. 7/1–7/5.
[25] A. Iannarelli, Ear Identification, Paramount Publishing Company, 1989.

[26] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015.

[27] S. Komkov, A. Petiushko, Advhat: Real-world adversarial attack on arcface face id system, in: Proceedings of the 25th International Conference on Pattern Recognition (ICPR), 2021.

[28] A. Kumar, T.S.T. Chan, Robust ear identification using sparse representation of local texture descriptors, Pattern Recognit. 46 (2013) 73–85.

[29] A. Kumar, C. Wu, Automated human identification using ear imaging, Pattern Recognit. 45 (2012) 956–968.

[30] R.A. Priyadharshini, S. Arivazhagan, M. Arun, A deep learning approach for person identification using ear biometrics, Appl. Intell. 51 (2021) 2161–2172.

[31] H. Proença, E. Yaghoubi, P. Alirezazadeh, A quadruplet loss for enforcing semantically coherent embeddings in multi-output classification problems, IEEE Trans. Inf. Forensics Secur. 16 (2021) 800–811.

[32] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 44 (3) (2020) 1623–1637.

[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015.

[34] R. Spolaor, Q. Li, M. Monaro, M. Conti, L. Gamberini, G. Sartori, Biometric authentication methods on smartphones: a survey, PsychNology J. 14 (2016) 87–98.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[36] A. Torralba, A.A. Efros, Unbiased look at dataset bias, in: Proceedings of the CVPR 2011, 2011, pp. 1521–1528.

[37] T. Wheelwright, Cell phone behavior in 2021: how obsessed are we?, 2021, https://www.reviews.org/mobile/cell-phone-addiction.

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5987–5995.

[39] P. Yan, K.W. Bowyer, Biometric recognition using 3d ear shape, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 1297–1308.