CLEI electronic journal, Volume 25, Number 2, Paper 8, May 2022

# Domain Adaptation for Unconstrained Ear Recognition with Convolutional Neural Networks

**Solange Ramos-Cooper**
Universidad Catolica San Pablo, Department of Computer Science,
Arequipa, Peru,
*solange.ramos@ucsp.edu.pe*

and

**Guillermo Camara-Chavez**
Federal University of Ouro Preto, Computer Science Department,
Ouro Preto, Brazil
*guillermo@ufop.edu.br*

**Abstract**

Automatic recognition using ear images is an active area of interest within the biometrics community. Human ears are a stable and reliable source of information since they are not affected by facial expressions, do not suffer extreme change over time, are less prone to injuries, and are fully visible in mask-wearing scenarios. In addition, ear images can be passively captured from a distance, making it convenient when implementing surveillance and security applications. At the same time, deep learning-based methods have proven to be powerful techniques for unconstrained recognition. However, to truly benefit from deep learning techniques, it is necessary to count on a large-size variable set of samples to train and test networks. In this work, we built a new dataset using the VGGFace dataset, fine-tuned pre-train deep models, analyzed their sensitivity to different covariates in data, and explored the score-level fusion technique to improve overall recognition performance. Open-set and close-set experiments were performed using the proposed dataset and the challenging UERC dataset. Results show a significant improvement of around 9% when using a pre-trained face model over a general image recognition model; in addition, we achieve 4% better performance when fusing scores from both models.

**Keywords:** Ear recognition, CNN, Mask-RCNN, VGG16, VGGFace, transfer learning, domain adaptation, score-level fusion.

## 1 Introduction

Human identification through biometrics is an ever-growing need in our society due to its main application in security, forensic science, and surveillance. Biometric traits are used because of their invariance over time, and the uniqueness of each individual [1]. Common physical traits are fingerprints, palmprints, iris, face, ear, hand geometry, and voice. Among all these traits, using ears as a biometric structure to identify people has some appealing advantages over the rest. In the current COVID-19 pandemic scenario, fingerprints and palmprints rely on contact over a surface that could be contaminated. Face recognition fails due to the mask-wearing situation. Iris recognition involves close eye contact, besides special sensors are needed. Nonetheless, the ear structure remains stable over a person's lifetime, is not affected by aging or facial expressions, and is less prone to injuries than hands or fingers, for instance. Thus, ear recognition can be performed contactless, ear images can be captured from a distance, and the explicit cooperation of the individual is not needed.

On the other hand, there are some challenges that arise when it comes to ear recognition in real-world settings such as low resolution, blur, illumination variations, and mostly occlusions caused by hair, head accessories, and earrings. There are some datasets that have been released to tackle the ear recognition problem but not all of them gather images taken under uncontrolled conditions, most of them were taken under constrained settings or a laboratory-like environment. Moreover, not only the environment where

images were taken matters but the number of classes, and the number of samples per class are also important. Variation within images from different subjects is called inter-class variation, while distinction among images of the same individual is called intra-class variation. Building a robust feature description algorithm should be fed with great inter-class and intra-class variability. Thus, collecting data under uncontrolled conditions and big training sets play an important role when using Convolutional Neural Networks (CNN) for unconstrained recognition (a.k.a recognition in the wild) [2] [3].

In recent years, CNN-based approaches have been commonly used for object detection, segmentation, and classification. This can be reviewed in the submissions to one of the most very well-known computer vision competitions, the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), where algorithms based on DL techniques achieved state-of-the-art performance on such computer vision tasks [4]. Additionally, specific tasks like face and ear recognition have been also approached with DL-based techniques, even competitions for these two specific tasks have been launched [5] [6] [7] [8]. In this sort of competition, face recognition has reached a significant level of maturity against ear recognition. This huge progress is benefited from the large-scale datasets and deep CNN architectures that have been proposed in the literature to solve this specific task. Take for example, the Google FaceNet [9] that made use of 200 million images to train. The DeepFace is another architecture proposed in [10] that was trained using 4.4 million labeled images. Also, the VGG-Face, a VGG16-based model proposed in [11], was trained using 2.6 million unconstrained face images. These architectures have led to some of the most interesting applications in biometrics and security. Compared with existing unconstrained face databases, ear databases are much smaller regarding the number of classes and samples per class. Table 1 shows a summary of some face databases that have been employed to train and test state-of-the architectures for the face recognition task, while Table 2 presents the same information of some ear datasets available in the literature.

Face recognition in the wild stays many steps ahead of ear recognition, a fact that we can take advantage of. In this work, we intend to build a large-scale ear dataset as an extension of one face database that collects images from the web of thousands of subjects. Furthermore, we use pre-trained models on face images to fine-tune and adapt them to ear domain recognition. According to our experiments, fine-tuning face-based models performs better than general-imaging-based pre-trained models.

The major contribution of the proposed work is defined as follows:

- We prepare an ear dataset as an extension of the VGGFace dataset. Samples of the proposed dataset present high variability in illumination, head position, usage of accessories, and occlusion.
- We propose ear domain adaptation from a pre-trained model on face recognition using the proposed dataset, which enhances accuracy by around 9%.
- We fine-tune the VGG16 and VGGFace models to analyze the significance of pre-trained data, image input size, and final feature vector size that describes ears.
- We explore score-level fusion by combining matching scores of the two fine-tuning models, which enhances accuracy by around 4%.
- We analyze the effect of different annotation categories on data like gender, race/ethnicity, image aspect ratio, and size.

In a previous version of this paper [12], we showed ear domain adaptation by fine-tuning the pre-trained VGGFace model which leads to a better performance in contrast to a model trained on general image recognition. The model was trained using a dataset built as an extension of the VGGFace dataset. In this version, we extend the discussion by analyzing the effect of different annotation categories on the demographic data such as gender and race/ethnicity, and on image properties like size, and aspect ratio. The extended dataset is described based on these variables, and recognition performance is evaluated according to these annotation categories.

The rest of this document is organized as follows. Section II shows an overview of the publications and contributions in the context of this work. Section III gives a detailed description of the presented unconstrained ear database. The employed methods and models in this work are explained in section IV. The following section shows experimental results and analysis. Finally, conclusions and future works are detailed in section VI.

## 2 Related Work

Approaches in the ear recognition field have been through from early techniques like structural methods which rely on physiological and geometrical features such as shape, wrinkles, and ear points [20, 21, 22]. Later, some other approaches that use subspace learning methods like Principal Components Analysis (PCA)

Table 1: Face Datasets

| Dataset | Classes | Samples per class | Total images |
|---|---|---|---|
| FERET [13] | 994 | 5 - 91 | 11 338 |
| LFW [14] | 5 749 | 1 - 530 | 13 233 |
| CelebA [15] | 10 177 | - | 202 599 |
| Youtube Faces [16] | 1 595 | 48 - 2 157 | 621 126 |
| VGGFace2 [11] | 9 131 | 87 - 843 | 3 311 286 |

Table 2: Ear Datasets

| Dataset | Classes | Samples per class | Total images |
|---|---|---|---|
| IITD [17] | 125 | 3 - 5 | 493 |
| AMI [18] | 100 | 7 | 700 |
| WPUT [19] | 501 | 4 - 8 | 2 070 |
| UERC [6] | 346 | 10 - 93 | 4 104 |
| Proposed | 610 | 25 - 350 | 61 915 |

[23], Linear Discriminant Analysis (LDA) [24], Local Principal Independent Components (LPICs) [25], and Independent Component Analysis (ICA) [26] were also used to improve ear recognition performance. Lately, some popular techniques based on handcrafted feature extraction such as Gabor filters [27], Local Binary Patterns (LBP) [28], Scale-Invariant Feature Transform (SIFT) [29, 30], contourlet transform [31], wavelet transform [32], and gradient-based features [33] were widely used as well. Finally, the latest approaches that have been explored within the ear recognition field are the CNN-based methods. A clear outstanding advantage when using these kinds of methods is that they perform better recognition in unconstrained conditions due to their capability to provide in-depth representations, different from earlier methods which usually perform poorly to image variations [6, 34].

Besides approaches and methods proposed in the literature, another aspect that also motivates ear recognition research is the availability of databases. Some databases that were created specifically for research purposes but present some variability and aimed to simulate real-world-conditions are: The West Pomeranian University of Technology Dataset (WPUT) [19], The University of Notre Dame Database (UND) [35], which gathers several different collections of 2D images as well as range images, The Indian Institute of Technology Delhi Ear Database (IITD) [17], The University of Beira Interior Ear Dataset (UBEAR) [36], and The Mathematical Analysis of Images (AMI) Ear Database [18] which presents little variability within their samples.

Some other datasets gather images from uncontrolled settings without paying particular attention to head position, illumination, occlusion, and of different sizes and quality. First, the Annotated Web Ears (AWE) database [37] which contains 1,000 images collected from the web of 100 subjects. This database has an extended version called the Annotated Web Ears Extended (AWEx), which is described in [38]. A later new extended version was presented in the Unconstrained Ear Recognition Challenge (UERC) [6] gathering 2,304 images from 166 subjects for the train set, and 1,800 images from 180 subjects (10 images per subject) for the test set. There is also a third set of images from 3,360 subjects with between 1 to 3 images per subject, which is also considered a test set. At the time of writing, this is the only available unconstrained ear dataset that can be used to train and test models for recognition. Second, the In-the-wild Ear Dataset [39] gathers around 2,600 labeled images with 55 landmark points used for detection, segmentation, and alignment. However, this dataset is not for recognition task training since annotations do not include subject information. Finally, a third dataset, the WebEars, was presented in [40] and contains 1,000 images. Later, the authors extend this database with the help of the University of Science & Technology of Beijing (USTB) and The Helloear Co. Ltd. It is called USTB-Helloear Database, and it contains a great amount of 2D images taken under

uncontrolled conditions. However, it is not publicly available. Fig. 1 shows sample images from some of the databases mentioned above. It can be noticed the variability among samples from the same class. Similarly, Table 2 lists ear datasets that we found available to explore.

(a) IITD

(b) WPUT

(c) AMI

(d) UERC

Figure 1: Different ear dataset samples

Regarding approaches that employed CNN-based methods, one of the first proposals that make use of this method was presented in [41]. The authors designed a convolutional neural network with three convolutional layers, a fully-connected layer, and a softmax classifier to deal with the partial occlusion problem. Their approach was tested using the USTB dataset, and the authors concluded that having limited training data restricts the performance of CNNs. [42] propose a method based on Geometric Morphometrics and Deep Learning (DL) for automatic ear detection and description. Using a large-scale landmark annotated dataset, they trained a model to detect 45 landmarks in 2D gray-scale images automatically and then used these landmarks as a feature vector for ear description.

Since one the main problem when using CNN-based techniques is the lack of enough training data. Authors in [2] proposed a transfer learning in a two-stage fine-tuned process for domain adaptation. In the first stage, a model was fine-tuned with constrained images, while in the second stage, unconstrained images were used. Experiments were performed using three models: VGG16, AlexNet, and GoogLeNet. Results show the best performance of VGG16 over the other two models. Similarly, [43] applied an aggressive data augmentation process to show the impact in full and selective learning models. Full learning refers to learning from scratch, while selective learning refers to initializing a model with parameters already learned from a different dataset. The authors used three models in their experiments VGG16, AlexNet, and SqueezeNet. [44] used the pre-trained AlexNet architecture, replaced the last three layers, and fine-tuned the model. The authors collect their own dataset of 300 images in unconstrained conditions to train and test the fine-tuned model. Dodge *et al.* [45] performed unconstrained ear recognition using transfer learning as well. However, this work compares existing DNNs as feature extractors, fine-tuned models, and a deep learning-based averaging ensemble architecture. Performance results are provided on unconstrained ear recognition datasets, the AWE, CVLE datasets, and a combined AWE+CVLE dataset. Their ensemble architecture resulted in the best recognition performance on these unconstrained datasets.

Likewise, Zhang *et al.* [3] showed that by assembling different architectures trained on the same dataset, the recognition rate improves than using single architectures. Authors replaced pooling layers for spatial pyramid pooling layers to fit arbitrary data sizes on models and get multi-level features; they also implemented center loss to obtain more discriminative features. Experiments were performed over a dataset that the authors built for training and testing their approach and the AWE dataset. Alshazly *et al.* [**?**] also built ensembles of deep CNN-based models and tested them using ear images acquired under controlled and uncontrolled conditions. They concluded that it is crucial to utilize multiple representations for ear images acquired under uncontrolled settings to improve recognition performance. [46], and [47] ensembles ResNet-like models and show a comparative analysis between two scenarios, using CNNs as feature extractors only and then training top layers for classification. A critical fact that led to a slightly better upgrade in one proposal was preserving the aspect ratio of images when training. Nonetheless, both proposals obtained better performance by averaging ensembles of fine-tuned networks with custom input sizes.

Some authors relied on combining different techniques to improve recognition. [48] trained a landmark detector on ears to later use this information for alignment and normalization. Then a CNN model was trained from scratch using unconstrained ear images. Besides, feature vectors were obtained using hand-

crafted techniques. After feature extraction, authors use PCA to reduce vectors to the same size. Finally, matching scores from both feature vector types were summed up with "sum fusion" after applying a min-max normalization process. The best result was given by fusing the CNN and HOG feature vectors. [49] presented a model called ScoreNet, which includes three basic steps. First, create a pool of modalities (image resizing, pre-processing, image representation, dimensionality reduction, and distance measurement). Second, select randomly a pipeline that includes different modalities. Third, the pipeline is applied to the training and validation set to generate a similarity score matrix. The algorithm selects the best groups of modalities and calculates the necessary fusion weights in a cascade network structure. These two last approaches were presented in Unconstrained Ear Recognition Challenge 2019 [6], being the top two techniques that used hybrid approaches (learned and handcrafted) and got the best results.

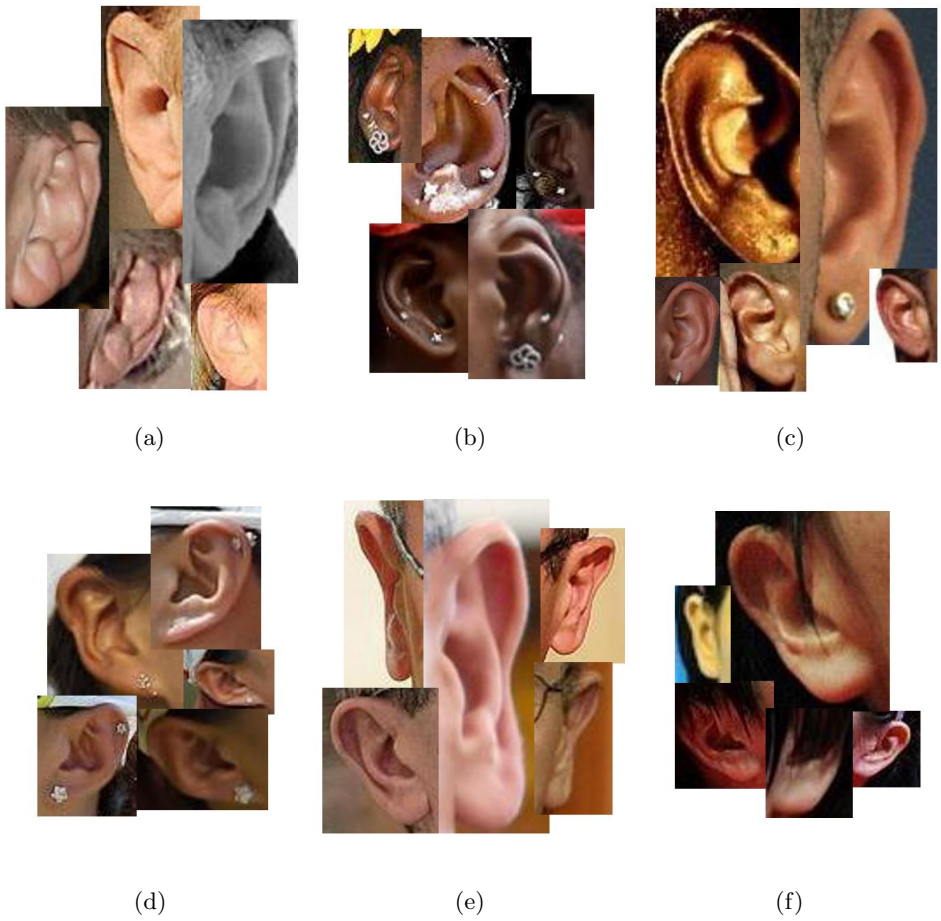
Emersic *et al.* [38] presented an experimental evaluation of several descriptors- and deep learning-based ear recognition models, studying the characteristics of the recognition techniques and the impact of various covariates such as gender, ethnicity, accessories, and head movements. Experiments were performed over the Annotated Web Ear (AWE) dataset, and results indicated that the presence of accessories and head movements significantly affect the identification performance, whereas other covariates of gender and ethnicity only affected the performance to a limited extent. In the same vein, [50] presented a comparison among extracted features using handcrafted descriptors and learned features using CNNs. The authors used seven top handcrafted descriptors and four AlexNet-based CNNs models to evaluate performance extensively. The obtained results demonstrated that CNN features were superior in recognition accuracy, outperforming all handcrafted features. The performance gain in recognition rates was above 22% over the best performing descriptor on the AMI and AMIC datasets, where the number of images per subject was relatively small. However, the performance gain was within 3% for the CVLE dataset, which had fewer subjects and more images per subject, but higher intraclass variability. Even though authors show that CNNs can be trained using small datasets, they point out that it is possible to improve performance when more training data is available.

In [51], authors presented a complete deep ear recognition pipeline. They segmented the ear using RefineNet, then features are extracted using a ResNet model, and finally K-NN classifier was trained to perform matching. The method is evaluated on the UERC database, and the deep learning-based approach has shown superior results over using handcrafted feature extractors. [52] evaluated a complete pipeline for ear recognition using a Faster Region-based CNN as object detector, a CNN as feature extractor, PCA for feature dimension reduction, a genetic algorithm for feature selection, and a fully connected artificial neural network for feature matching. Experimental results showed the time needed for the complete pipeline execution, 76ms for matching (database of 33 ear images features), 15ms for feature extraction, and ear detection and localization requiring 100ms. Experiments were performed on a combined AMI dataset and images taken from the author's environment. [53] proposed a simple 6-layer CNN architecture for ear recognition. They studied CNNs by varying the parameters such as kernel size, learning rate, epochs, and activation functions. Since the model is simple, it requires very little memory, making it feasible to port into any embedded/handheld device. Finally, the approach presented in [54] proposed a deep learning pipeline using transformer neural networks: Vision Transformer (ViT) and Data-efficient image Transformers (DeiTs). Similar to the concept of transfer learning on pre-trained CNN architectures, this study replaced the final layer of ViT and DeiT to enable the transformer network to learn the features from the extracted training ear images of the two unconstrained ear datasets.

## 3 VGGFace-Ear Database

Employing deep learning techniques usually demands considerable amounts of training data. This statement was also pointed out by [3]. However, they build an ear database from a nonpublic face database; thus, the final ear database can not be shared with the community research. In this case, we take advantage of one of the biggest face datasets available to detect ears and build a large-scale ear dataset. The VGGFace2 dataset gathers around 3.3M images collected from the web. We used this dataset to build an extended version, which gathers around 61.9K images from 610 different subjects. Samples per class vary from 50 to 350. Some sample images from this generated dataset can be seen in Fig. 2.

First of all, for the ear detection task, we used the Mask Region-based Convolutional Neural Network (Mask R-CNN). This architecture was introduced in [55]. Unlike other region-based architectures, this model detects objects at a pixel level, which means the output is a binary mask that indicates whether or not a given pixel is part of an object. Even though we are using bounding box regions instead of masks in the ear recognition stage, this architecture helps us avoid unwanted information close to the ear but not part of the ear, like accessories. It also helps us avoid extreme occlusion since the bounding box only encloses information from ears, if this bounding box has little information from ears, we discard it. We used an



(a) (b) (c)

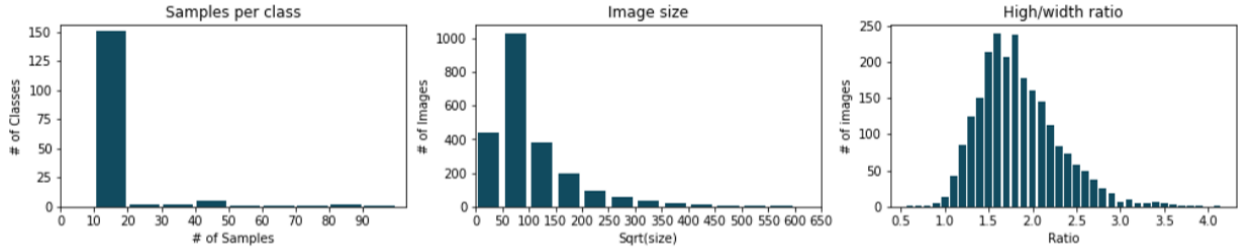
(d) (e) (f)

Figure 2: Sample images from the VGGFace-Ear dataset (all images were scaled the same to point out real size variation and resolution)

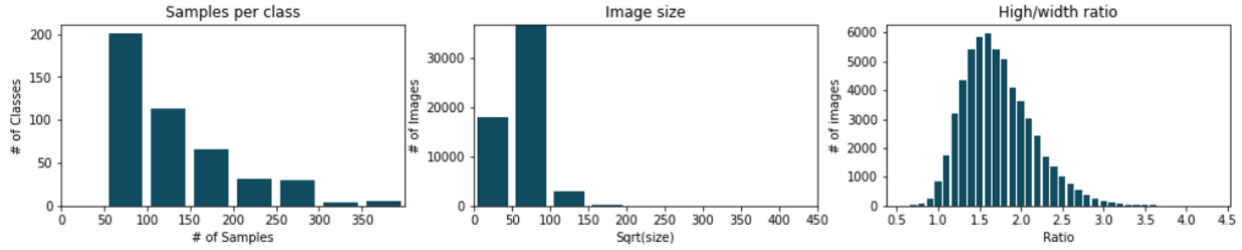
available online implementation of the Mask R-CNN architecture [56] within the Ear Segmentation Dataset available at [57] to fine-tuned the architecture to detect ears. Over the 1,000 available images, we used 750 images for training and 250 for testing the model. Considering an Intersection over Union (IoU) threshold of 0.5, the Average Precision (AP) value is 97%, and considering a range of thresholds from 0.5 to 0.95 IoU, the AP is 72.9%. Thus, the model detects ears with high accuracy. However, it also detects some other objects as ears in some cases. We detected ears on 1,000 classes of the VGGFace dataset, approximately one-eighth of the total number of classes. From this group, only classes with the most samples were included in the VGGFace-Ear dataset. Finally, we obtained 450 classes with several samples between 50 and 350 that will be used as the train set and 160 classes with 25 samples per class used as the test set. It is worth mentioning that all data were manually checked to ensure that it gathers only ear images.

Comparing the proposed dataset with the UERC dataset, we present an overall analysis of the number of classes per range of samples, the number of images per range of sizes (the squared root of size), and the number of images per ratio of high over width. Fig. 3a shows that from a total of 166 classes, around 150 classes contain between 10 and 20 samples per class. The squared root of the area of most samples is less than 100 pixels. However, some samples reach an area root squared size of 500 to 600 pixels. The aspect ratio of all samples ranges from 1 to 3. Fig. 3b shows same aspects for the VGGFace-Ear dataset. It outnumbers samples per class, where more than half of classes gather from 50 to 150 samples. None of the samples' sizes are over 200 pixels (the squared root of size), but most samples' aspect ratio is in the same range as the UERC dataset. Moving onto test sets of both databases, Fig. 4 shows a similar analysis.

In addition, Fig. 5 and 6 show information about percentage distribution regarding gender and race/ethnicity from the UERC and VGGFace-Ear datasets, respectively. When it comes to gender, it is inevitable having more male classes than females, however, when building the VGGFace-Ear test set collection, we intended to find a similar proportion among all classes. On the other hand, race/ethnicity proportions outnumber samples from White people over Asian, black, and Hispanic/Indian in most sets. Nonetheless, in the case of the VGGFace-Ear test set, a major number of classes from Asian, Black, and Hispanic/Indian people were tried to include more than White people. All these intentions were done to fully analyze the impact of these

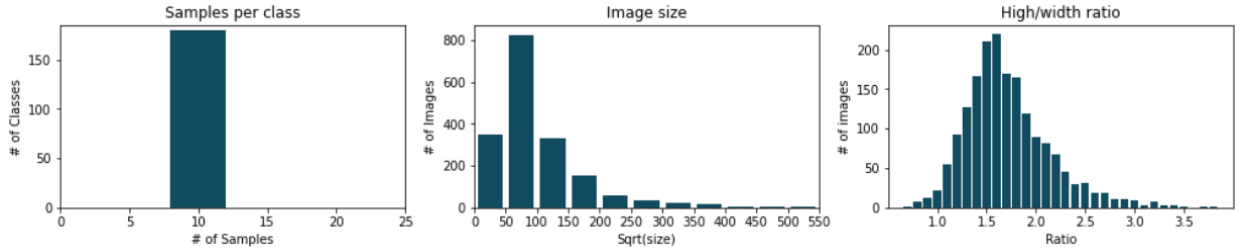



(a) UERC train set. 166 classes

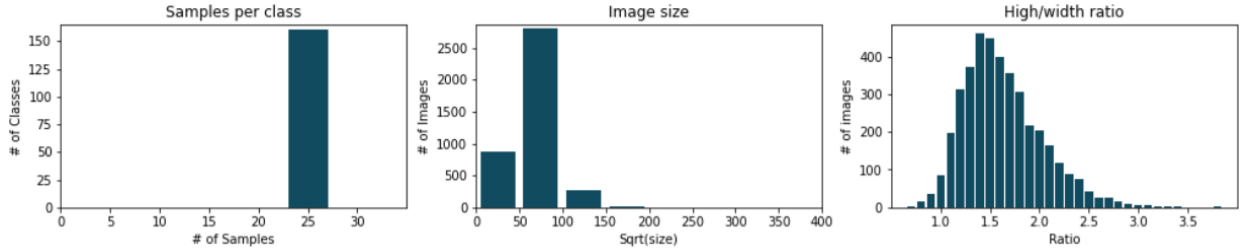



(b) VGGFace-Ear train set. 450 classes

Figure 3: Distribution and comparison of train sets from the UERC and the proposed VGGFace-Ear datasets considering samples per class, and samples image size and aspect ratio




(a) UERC test set. 180 classes



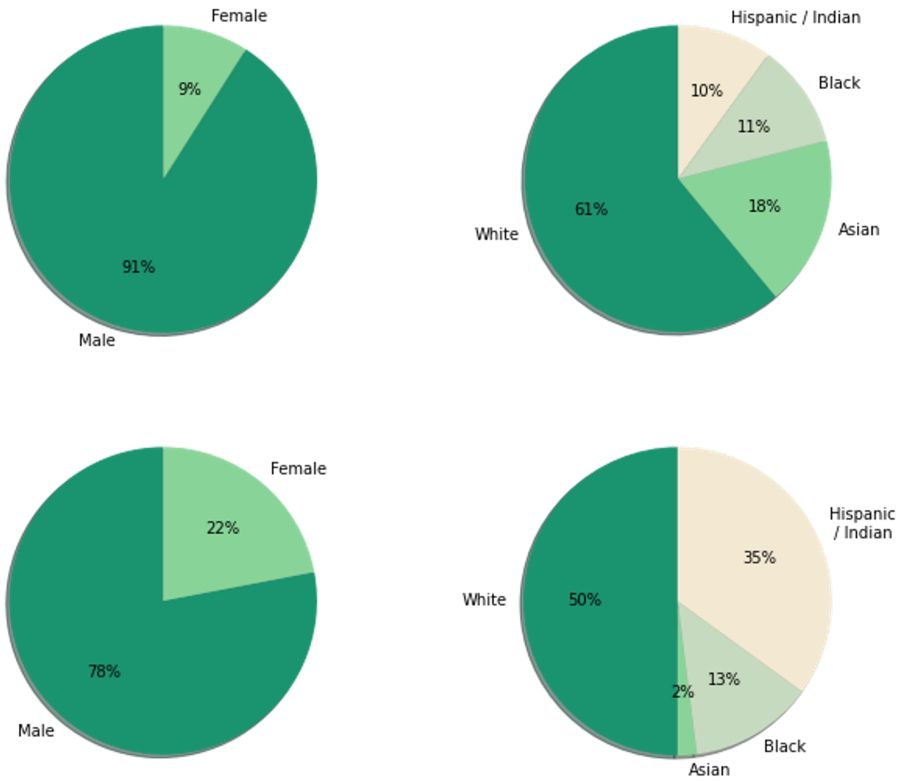

(b) VGGFace-Ear test set. 160 classes

Figure 4: Distribution and comparison of test sets from the UERC and the proposed VGGFace-Ear datasets considering samples per class, and samples image size and aspect ratio

distributions when training and testing models. This analysis is presented in Section V.

## 4 Ear Recognition using CNNs

This section presents a comprehensive experimental evaluation of the proposed approach for the ear recognition task under unconstrained conditions. Using the proposed dataset, we employed two pre-trained CNNs to adjust them to the ear domain. Best performance is achieved using a model pre-trained on face images rather than pre-trained for general image recognition.




(a) Gender

(b) Race/Ethnicity

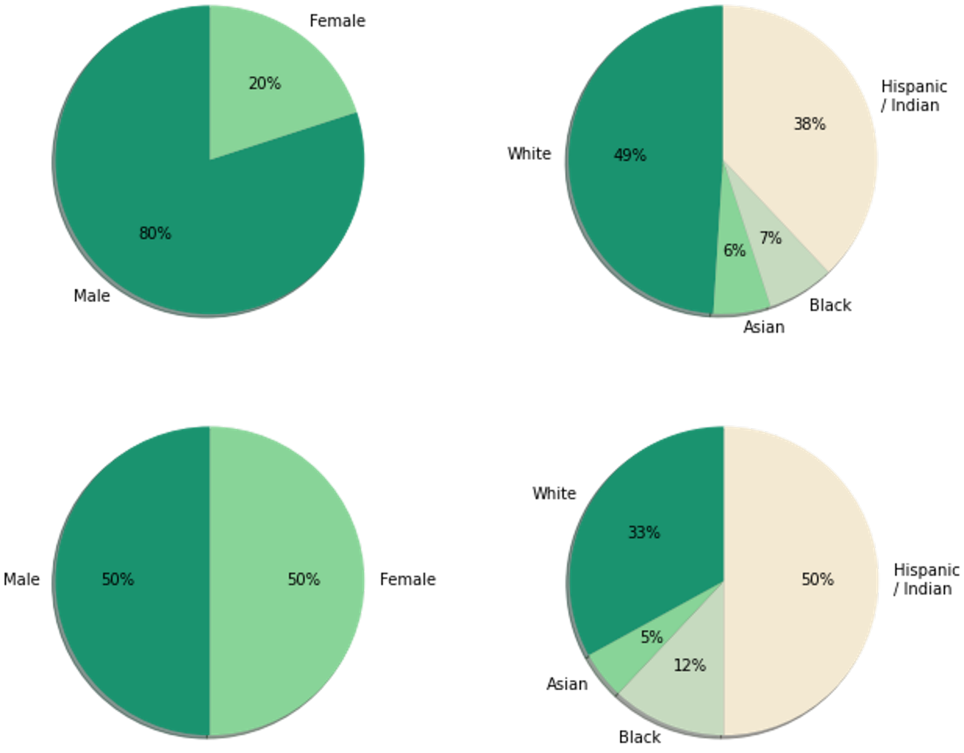
Figure 5: Overview of the UERC dataset regarding gender and race/ethnicity sampling. The first row shows information from the train set while the second row, from the test set

### 4.1 Pre-trained models

We employed VGG-based models to train ear image representation. First, the VGG16 [58] was presented as a submission of the ILSVRC 2014 competition. The model achieves 92.7% top-5 test accuracy over 14 million images belonging to 1,000 classes. This model and others are implemented in most DL frameworks and the weights of the trained models using the ImageNet Database. This architecture contains 16 layers divided into 5 convolutional groups and 3 final dense layers. We reduced the input size for our experiments, removed one convolutional group, and added one dense layer before the softmax activation layer. Second, the VGGFace [11], this model was trained using the VGGFace2 dataset, and the architecture model and trained weights were shared by the authors. In this case, we used the same input size and did not remove any layer.

### 4.2 Fine tuning

Parameters from both architectures, VGG16 and VGGFace, were fine-tuned to adapt these networks to the ear domain recognition. A fine-tuning process trains the network for some more iterations on a target dataset. Thus, filters trained on ImageNet and VGGFace2 datasets will be adapted to the new VGGFace-Ear dataset. The VGG16 architecture was modified as follows. First, the input size was reduced to $112 \times 112$ pixels. Second, the original VGG16 network includes 5 convolutional blocks before classification, in our model, only four convolutional blocks were included in the network. Third, all top dense layers were removed and replaced by 4 new dense layers, two of 4,096 units, one of 2,622, and the last layer with 450 units, which are the number of classes in the target dataset. So, these 4 last layers were trained from scratch while the rest, convolutional layers, were fine-tuned. In this case, we removed all dense layers because these units were trained to classify the 1,000 objects from the ImageNet, which is much different from the new target object, the ear. Nonetheless, for the VGGFace model, the top dense layers remained and only the last layer was replaced with a new layer with 450 units, the number of classes in the VGGFace-Ear dataset. In this case, we considered there is a close relation between face and ear objects, thus the top layers were also fine-tuned




(a) Gender (b) Race/Ethnicity

Figure 6: Overview of the VGGFace-Ear dataset regarding gender and race/ethnicity sampling. The first row shows information from the train set while the second row, from the test set

and the rest of the layers. This new model has an input size of $224 \times 224$, just like the original network. Both networks were trained using the Stochastic Gradient Descent (SGD) optimizer on a sparse categorical cross-entropy loss using momentum with a decay of 0.9, and the learning rate was set to 0.0001. All layers from both networks were set trainable and trained for 50 epochs.

### 4.3 Score-level fusion

This sort of fusion is also known as confidence level or measurement level fusion and it is widely used in the literature [59]. This technique combines matching scores produced by different match methods to generate one final score. Since different modalities could be used for feature extraction, matching scores can be in different ranges. Thus, the first step to fuse information is choosing an appropriate normalization technique to convert the matching scores into a similar common domain. Then, combining scores can be performed by using some common rules such as sum, mean, maximum, and minimum. Finally, the resulting score of multiple matches is considered the final score. In this work, we employed this technique to fuse scores of both fine-tuned models.

### 4.4 Data augmentation and pre-processing

Some common data augmentation techniques were applied to the train set. First, the train set consists of 450 classes with between 50 to 350 samples per class. Randomly 14 samples from every class were separated to be part of the validation set used in the training process. Remainder samples were applied data augmentation techniques to reach a number of 1,000 samples per class. Some of the techniques used are scale in both axes $(-20\%, 20\%)$, translation in the x-axis $(-15\%, 15\%)$ and y-axis $(-5\%, 5\%)$, rotation $(-15°, 15°)$, brightness and contrast increasing, and decreasing, histogram equalization, Gaussian noise addition, and vertical flipping. Besides data augmentation, pre-processing and re-scaling methods were also employed. Common methods of pre-processing images to fit different architectures' input sizes are resizing, cropping, and filling. In Fig. 7, it is shown graphical examples of these three methods. Resizing results in geometric

distortion. Cropping may discard important information. Black filling adds irrelevant information, but if the image is so thin, this unwanted information could overwhelm important information. Finally, the color filling method is our choice, with this technique we prevent distortion and losing information by adding a background extended from the border pixels of the image. Once images were pre-processed to be squared, a re-scaling process is performed to reduce or increase the image size to fit the input layer size of the networks.

(a) Original (b) Resize (c) Cropped (d) Black Filled (e) Nearest Filled

Figure 7: Ear region pre-processing to fit CNN models input size

## 5 Experimental Results and Discussion

### 5.1 Datasets

Experiments were performed on three different datasets and on the validation and test sets of the proposed dataset. One first dataset is AMI [18] which includes 700 images of 100 subjects, 7 images per person. All images are the same size. This constrained set collects images from students, teachers, and staff of the Computer Science department at Las Palmas de Gran Canaria University in Spain. One particular characteristic of this dataset is that for each individual, six images were taken from the right ear and only one from the left ear. Is a low variation set according to pose and illumination but occlusion and accessories are present in some cases. The second semi-constrained dataset is WPUT [19]. This set gathers semi-constrained images because some were taken in different outdoor places and at different periods. It collects 2,701 images from 501 subjects. All samples are the same size and in high resolution. The authors tried to reflect real-life conditions with occlusion, pose variation, and the use of accessories. The third is the UERC dataset, which gathers 2,304 samples from 166 subjects in the train set and 1,800 images from 180 subjects in the test set, 10 samples per subject. Different from the above, images in this dataset were collected from the web using crawlers and public figures, and famous people. Finally, the validation and test sets of the VGGFace-Ear. The first gathers 6,300 images of the 450 subjects the model was trained on, and the test set contains 4,000 samples from 160 subjects different from the ones used to train the model.

### 5.2 Experimental protocols

The proposed VGGFace-Ear dataset was divided into 2 disjoint sets. One first set gathers 450 classes and a test set of 160 classes. The train set was used to fine-tune VGG16 and VGGFace architectures. From this train set, one second split was done by taking 14 images of every class to form the validation set and the rest of the samples were augmented to 1,000 samples per class, both subsets were used in the training process. Training and experiments were conducted using the Tensorflow framework [60]. For performance evaluation, we used Rank-1 and Rank-5 percentage values. Rank-n means that $n$ observations from the top set of the closest samples are taken into account to calculate the percentage. So, the recognition rate at Rank-1 corresponds to the fraction of probe images, for which an image of the correct identity was retrieved from the gallery in the top match; while the recognition rate at Rank-5 corresponds to the fraction of correct images found within the five top matches. For the evaluation process, images were pre-processed to make them the same height size as width size and then resize to fit input size models. For the fine-tuned VGG16, images were resized to $122 \times 122$, and for the fine-tuned VGGFace, to $224 \times 224$. Both fine-tuned models were used as descriptors to obtain a feature vector of every image. Feature vectors were obtained from the last three layers of the models, 2,622 and 4,096, and used to compare performance. Once feature vectors from all images were obtained, the cosine distance from all-vs-all images was calculated. In this all-vs-all matrix, small distances indicate similarity between two images while large distance means opposition. For every image, all their matches were ordered according to similarity. Finally, the recognition rate was then calculated using the Rank metric. For the score-level fusion evaluation, distances were first normalized using min-max normalization and then ordered according to similarity as well. Then, matching score matrices from both models were summed up and normalized again to finally order matches according to these scores values and calculate the recognition rate.

### 5.3 Experimental results

A first evaluation on the fine-tuned models performance is shown in Table 3. AMI and WPUT datasets gather images in controlled and semi-controlled conditions respectively, most images are high resolution and the same size. For these cases, fine-tuned VGG16 results in a slightly better performance than fine-tuned VGGFace trained with uncontrolled face images. However, testing over validation and test set of the VGGFace-Ear datasets shows different behavior. In this case, a little better performance is achieved using the fine-tuned VGGFace model. Uncontrolled ear images description is more effective using a model trained for a similar task than a model trained for general image recognition. It is worthy to point out that the evaluation of the VGGFace-Ear dataset validation set falls into a close-set experimental protocol since the trained model knows all classes. While evaluation over the VGGFace-Ear test set is an open-set evaluation because classes are different from the ones the model was trained on. It seems that models trained for recognition in the wild perform poorly on controlled images, this reveals the bias set on controlled and semi-controlled datasets that drives them away from real-world situations. In this case, a 4096-size feature vector was used to describe images. Fig. 8 and Fig. 9 also shows how much better VGGFace performs over the VGG16 on every different category of data. Those categories are people's race/ethnicity, image aspect ratio, and image size. It can be noticed also that in every sub-category where the VGG16 has poor performance, the VGGFace has good performance. Even though this improvement reduces the VGGFace-Ear test set (see Fig. 9), the VGGFace model still performs better than the VGG16 model. Moreover, analyzing specifically every sub-category of data, it is observed that when it comes to race/ethnicity, performance decreases for White, Asian, and Hispanic/Indian people but not for Black people. In the case of aspect ratio, performance reduces drastically on images where the ratio of height/width goes over 2.5, this means images that are in frontal face position and with little ear information. Finally, image size also reduces performance when it comes to images smaller than 50px squared.

Table 3: VGG16 vs VGGFace fine-tuning performance

| **Dataset** | **VGG16** | | **VGGFace** | |
|---|---|---|---|---|
| | **Rank-1** [%] | **Rank-5** [%] | **Rank-1** [%] | **Rank-5** [%] |
| AMI | **83.57** | **94.71** | 83.14 | 93.14 |
| WPUT | **63.24** | **80.77** | 61.16 | 78.84 |
| VGGFace-Ear (validation set) | 74.06 | 87.40 | **83.63** | **91.73** |
| VGGFace-Ear (test set) | 68.43 | 82.62 | **71.10** | **84.87** |

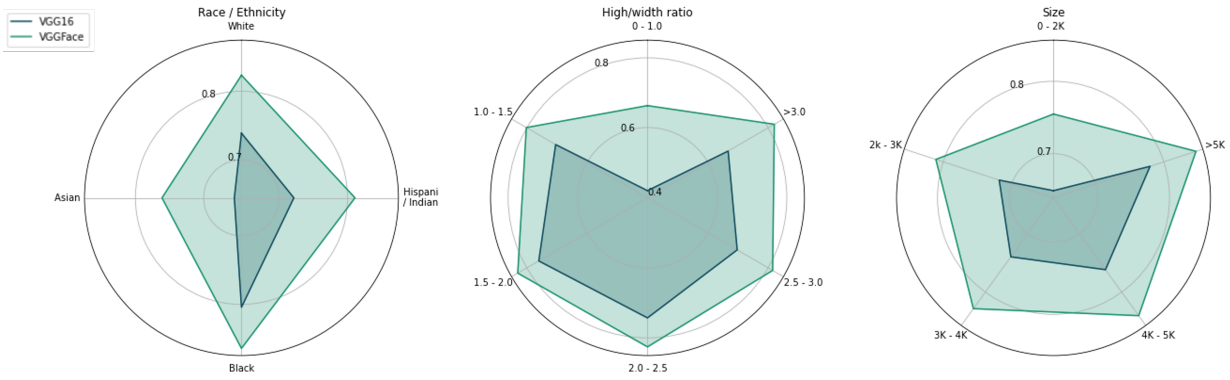



Figure 8: VGGFace-Ear validation set analysis on people's race/ethnicity, image aspect ratio, image size

One second analysis was done to evaluate fusion performance. This time we evaluate fusion on the UERC dataset test set. We calculate Rank-1 and Rank-5 values for both single models and then by combining the matching score of both models. We calculate rank values using two different feature descriptors of different sizes. Table 4 shows quantitative values while Fig. 10 shows radar charts for the three cases, two single

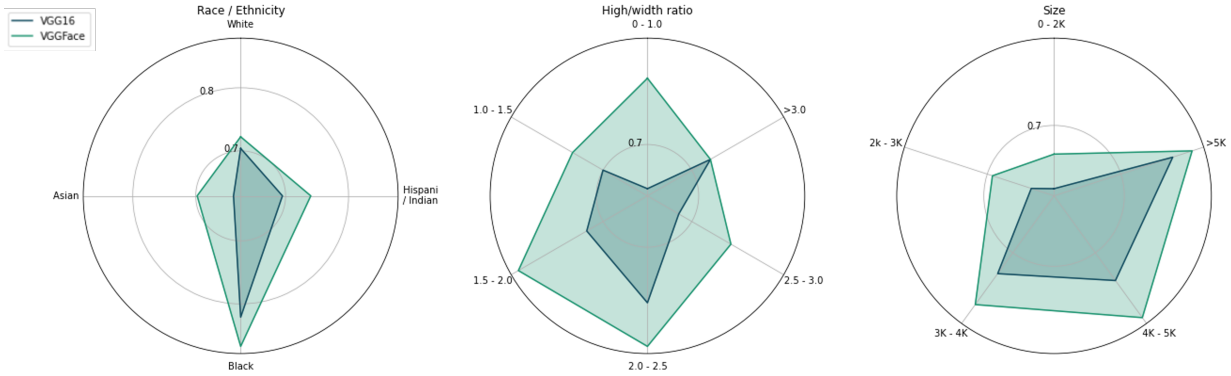



Figure 9: VGGFace-Ear test set analysis on people's race/ethnicity, image aspect ratio, image size

models, and fusion. It can be observed that recognition accuracy improves when combining both models not only in general performance but for every category and sub-category of data. Another category of data that was evaluated is the gender category, male and female classes. In this case, female recognition always drops compared to male classes. For the VGGFace-Ear test set, where recognition is around 71% using the VGGFace model, the same model reaches an 83.5% accuracy on male classes while on female classes go down to around 65%. Similar behavior is observed in the UERC dataset evaluation.

Table 4: Score-level fusion evaluation

| **Feature Vector** | **Fine-tuned VGG16** | | **Fine-tuned VGGFace** | | **Score-level Fusion** | |
|---|---|---|---|---|---|---|
| **Size** | **Rank-1** [%] | **Rank-5** [%] | **Rank-1** [%] | **Rank-5** [%] | **Rank-1** [%] | **Rank-5** [%] |
| 2622 | 57.67 | 78.72 | 69.44 | 86.50 | 73.56 | 88.67 |
| 4096 | 64.17 | 83.11 | 71.33 | 87.00 | **75.33** | **89.11** |

Finally, a third evaluation was performed on the UERC test set and compared with different approaches proposed in the literature (Table 5). This evaluation follows an open-set evaluation protocol, using the fine-tuned VGGFace as a feature descriptor and with a 4096-size feature vector. In this case, a second fine-tuning process was done using the UERC training set. This adjustment leads to an improvement of around 9% against other approaches. In addition, applying a score-level fusion of both models also improves the recognition rate by around 4%. Table 4 shows a comparative evaluation of both single models and the fusion approach, as well as, shows recognition rates reached using the two feature vector sizes. This evaluation is also performed using the UERC testing set. Evaluation is reported using Rank values and Area Under the Curve (AUC) of the Cumulative Matching Score (CMC).

Table 5: UERC Test set evaluation

| **Approach** | **HC/LR*** | **Rank-1** [%] | **Rank-5** [%] | **AUC** |
|---|---|---|---|---|
| CH-Fus [48] | Both | 47.6 | 69.3 | 0.946 |
| GooLeNet-MP [2] | LR | 60.9 | - | |
| VGG16-MP [2] | LR | 62.6 | - | |
| ScNet-5 [49] | Both | 62.8 | 82.1 | 0.966 |
| VGGFace-Ear | LR | **71.3** | **87.0** | **0.978** |
| VGGFace-Ear Fusion | LR | **75.3** | **89.1** | **0.981** |

*HC: Hand-crafted / LR: Learn / Both: Hybrid (HC + LR).

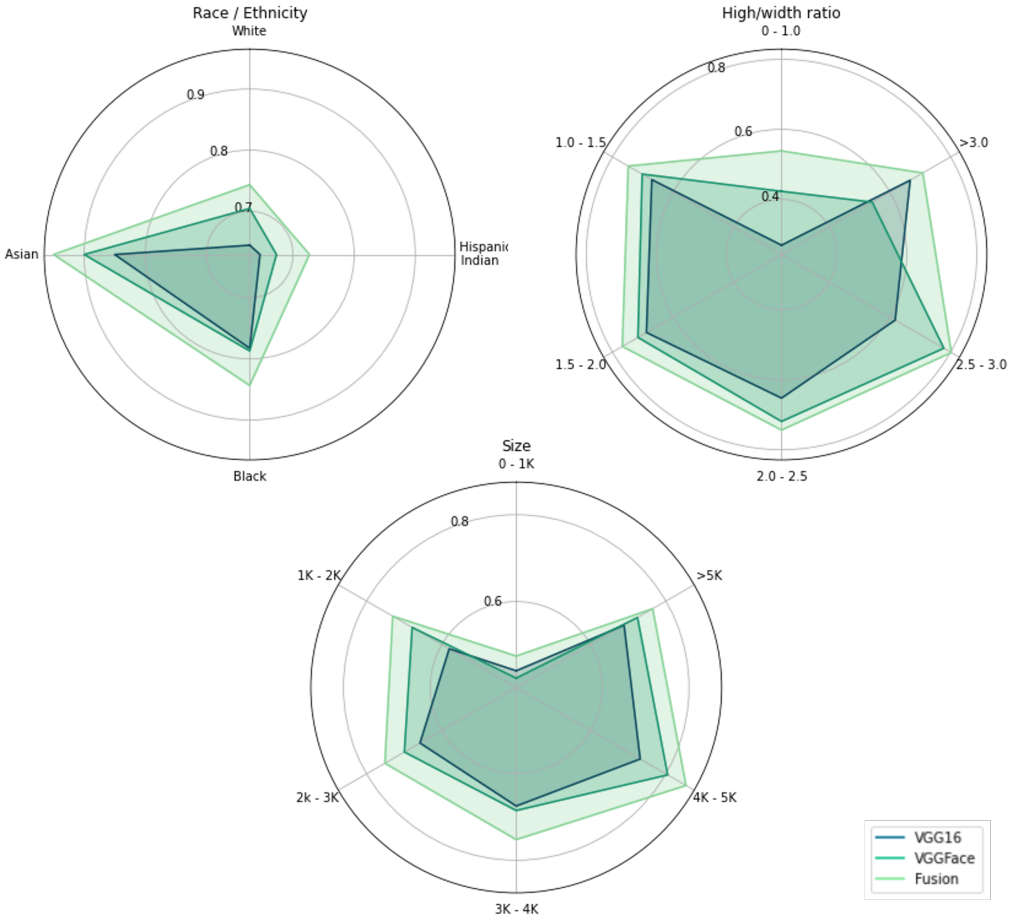



Figure 10: UERC test set analysis on people's race/ethnicity, image aspect ratio, and image size for single and fused methods

## 6 Conclusion and Future Work

This work introduces a new large-scale dataset for unconstrained ear recognition. Images gathered on this dataset were extracted from the VGGFace database using the Mask-RCNN for ear detection. Data shows high inter-class and intra-class variability intended to produce models with notable generalization ability when training CNN models. We propose ear domain adaptation from a pre-trained face model which leads to a better performance than using a pre-trained general image recognition model. Moreover, we explore a fusion technique at a score level of the two trained models, the VGGFace and VGG16, that result in even more positive recognition accuracy. Furthermore, we approach different aspects when training models like image input size, image pre-processing, and final feature vector size. Besides, using different annotation categories of data, we examine the generalization of models on these different categories and sub-categories. We conducted experiments on the UERC dataset, a challenging dataset of images taken under uncontrolled conditions. Even though we achieve an improvement against the state-of-the-art approaches, there is more room for further research. First, landmark detection and alignment or normalization may help the CNN focus more on the ear structure features than pose variation. Second, the ear recognition modality can be complemented by face recognition due to their proximity in space. Profile face positions of the individual show entire ear structure while frontal positions show little ear information but entire face information.

## Acknowledgment

This research was supported by the National Council for Science, Technology and Technological Innovation (CONCYTEC-Peru) and the World Bank within the Project "Incorporation of Researchers" through its executive unit ProCiencia [Grant 028-2019-FONDECYT-BM-INC.INV].

## References


[1] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, 1st ed. Springer Publishing Company, Incorporated, 2010.

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

[3] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, 2018.

[4] S. V. Lab, "Imagenet large scale visual recognition challenge," 2015, accessed on 08.01.2021. [Online]. Available: http://www.image-net.org/challenges/LSVRC/

[5] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdaii, Y. Zhou, S. Zafeiriou, D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "The unconstrained ear recognition challenge," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 715–724.

[6] Ž. Emeršič, A. K. S. V., B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, E. Hansley, M. P. Segundo, S. Sarkar, H. J. Park, G. P. Nam, I. . J. Kim, S. G. Sangodkar, U. Kacar, M. Kirci, L. Yuan, J. Yuan, H. Zhao, F. Lu, J. Mao, X. Zhang, D. Yaman, F. I. Eyiokur, K. B. Ãzler, H. K. Ekenel, D. Paul Chowdhury, S. Bakshi, P. K. Sa, B. Majhi, P. Peer, and V. Štruc, "The unconstrained ear recognition challenge 2019," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–15.

[7] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 3406–3415. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.363

[8] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2638–2646.

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2018, pp. 67–74. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/FG.2018.00020

[12] S. Ramos-Cooper and G. Camara-Chavez, "Ear recognition in the wild with convolutional neural networks," in *2021 XLVII Latin American Computing Conference (CLEI)*, 2021, pp. 1–10.

[13] NIST, "Face Recognition Technology (FERET)," 2017, accessed on 08.01.2021, National Institute of Standards and Technology. [Online]. Available: https://www.nist.gov/itl/products-and-services/color-feret-database

[14] G. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *Tech. rep.*, oct 2008.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2015, pp. 3730–3738. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.425

[16] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, 2011, pp. 529–534.

[17] A. Kumar and C. Wu, "Automated human identification using ear imaging," *Pattern Recognition*, vol. 45, no. 3, pp. 956–968, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320311002706

[18] E. Gonzalez, L. Alvarez, and L. Mazorra, "AMI Ear Database," 2019, accessed on 08.01.2021. [Online]. Available: https://ctim.ulpgc.es/research_works/ami_ear_database/

[19] D. Frejlichowski and N. Tyszkiewicz, "The west pomeranian university of technology ear database – a tool for testing biometric algorithms," in *Image Analysis and Recognition*, A. Campilho and M. Kamel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 227–234.

[20] Z. Mu, L. Yuan, Z. Xu, D. Xi, and S. Qi, "Shape and structural feature based ear recognition," in *Advances in Biometric Person Authentication*, S. Z. Li, J. Lai, T. Tan, G. Feng, and Y. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 663–670.

[21] M. Choras and R. S. Choras, "Geometrical algorithms of ear contour shape representation and feature extraction," in *Sixth International Conference on Intelligent Systems Design and Applications*, vol. 2, 2006, pp. 451–456.

[22] A. S. Anwar, K. K. A. Ghany, and H. Elmahdy, "Human ear recognition using geometrical features extraction," *Procedia Computer Science*, vol. 65, pp. 529–537, 2015, international Conference on Communications, management, and Information technology (ICCMIT'2015). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915029567

[23] K. Chang, K. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160–1165, 2003.

[24] Z. Xiaoxun and J. Yunde, "Symmetrical null space lda for face and ear recognition," *Neurocomputing*, vol. 70, no. 4, pp. 842–848, 2007, advanced Neurocomputing Theory and Methodology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231206002906

[25] Mamta and M. Hanmandlu, "Robust ear based authentication using local principal independent components," *Expert Syst. Appl.*, vol. 40, no. 16, p. 6478â6490, nov 2013. [Online]. Available: https://doi.org/10.1016/j.eswa.2013.05.020

[26] H.-J. Zhang, Z.-C. Mu, W. Qu, L.-M. Liu, and C.-Y. Zhang, "A novel approach for ear recognition based on ica and rbf network," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 7, 2005, pp. 4511–4515 Vol. 7.

[27] A. Kumar and C. Wu, "Automated human identification using ear imaging," *Pattern Recognition*, vol. 45, no. 3, pp. 956–968, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320311002706

[28] L. Jacob and G. Raju, "Ear recognition using texture features - a novel approach," in *Advances in Signal Processing and Intelligent Recognition Systems*, S. M. Thampi, A. Gelbukh, and J. Mukhopadhyay, Eds. Cham: Springer International Publishing, 2014, pp. 1–12.

[29] D. R. Kisku, H. Mehrotra, P. Gupta, and J. K. Sing, "Sift-based ear recognition by fusion of detected keypoints from color similarity slice regions," in *2009 International Conference on Advances in Computational Tools for Engineering Applications*, 2009, pp. 380–385.

[30] L. Chen, Z. Mu, B. Zhang, and Y. Zhang, "Ear recognition from one sample per person," *PLOS ONE*, vol. 10, no. 5, pp. 1–16, 05 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0129505

[31] C. Murukesh, A. Parivazhagan, and K. Thanushkodi, "A novel ear recognition process using appearance shape model, fisher linear discriminant analysis and contourlet transform," *Procedia Engineering*, vol. 38, pp. 771–778, 2012, iNTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705812020103

[32] Z. Hai-Long and M. Zhi-Chun, "Combining wavelet transform and orthogonal centroid algorithm for ear recognition," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 2009, pp. 228–231.

[33] H. A. Alshazly, M. Hassaballah, M. Ahmed, and A. A. Ali, "Ear biometric recognition using gradient-based feature descriptors," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, A. E. Hassanien, M. F. Tolba, K. Shaalan, and A. T. Azar, Eds. Cham: Springer International Publishing, 2019, pp. 435–445.

[34] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/19/4139

[35] C. V. R. Lab, "University of Notre Dame Datasets," 2018, accessed on 08.01.2021. [Online]. Available: https://cvrl.nd.edu/projects/data/

[36] R. Raposo, E. Hoyle, A. Peixinho, and H. ProenÃ§a, "Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions," in *2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, 2011, pp. 84–90.

[37] Å½iga EmerÅ¡iÄ, V. Å truc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017, bioinspired Intelligence for machine learning. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121730543X

[38] Z. Emersic, B. Meden, P. Peer, and V. Struc, "Evaluation and analysis of ear recognition models: performance, complexity and resource requirements," *Neural Computing and Applications*, 2018.

[39] Y. Zhou and S. Zaferiou, "Deformable Models of Ears in-the-Wild for Alignment and Recognition," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, may 2017, pp. 626–633.

[40] Y. Zhang and Z. Mu, "Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks," *Symmetry*, vol. 9, no. 4, 2017. [Online]. Available: http://www.mdpi.com/2073-8994/9/4/53

[41] L. Tian and Z. Mu, "Ear recognition based on deep convolutional network," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2016, pp. 437–441.

[42] C. Cintas, M. Quinto-Sánchez, V. Acuña, C. Paschetta, S. de Azevedo, C. Cesar Silva de Cerqueira, V. Ramallo, C. Gallo, G. Poletti, M. C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, A. Ruiz-Linares, R. Gonzalez-José, and C. Delrieux, "Automatic ear detection and feature extraction using Geometric Morphometrics and convolutional neural networks," *IET Biometrics*, vol. 6, no. 3, pp. 211–223, 2017.

[43] Z. Emersic, D. Stepec, V. Struc, and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, may 2017, pp. 987–994.

[44] A. A. Almisreb, N. Jamil, and N. M. Din, "Utilizing AlexNet Deep Transfer Learning for Ear Recognition," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2018, pp. 1–5.

[45] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, 01 2018.

[46] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," *IEEE Access*, vol. 8, pp. 170 295–170 310, 2020.

[47] H. Alshazly, C. Linse, E. Barth, S. A. Idris, and T. Martinetz, "Towards explainable ear recognition systems using deep residual networks," *IEEE Access*, vol. 9, pp. 122 254–122 273, 2021.

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

[49] U. Kacar and M. Kirci, "ScoreNet: deep cascade score level fusion for unconstrained ear recognition," *IET Biometrics*, vol. 8, no. 2, pp. 109–120, 2019.

[50] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus cnn features for ear recognition," *Symmetry*, vol. 11, no. 12, 2019. [Online]. Available: https://www.mdpi.com/2073-8994/11/12/1493

[51] Ž. Emeršič, J. Križaj, V. Štruc, and P. Peer, *Deep Ear Recognition Pipeline*. Cham: Springer International Publishing, 2019, pp. 333–362. [Online]. Available: https://doi.org/10.1007/978-3-030-03000-1_14

[52] A. Alkababji, "Real time ear recognition using deep learning," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, pp. 523–530, 04 2021.

[53] R. Ahila Priyadharshini, S. Arivazhagan, and M. Arun, "A deep learning approach for person identification using ear biometrics," *Applied Intelligence)*, vol. 51, pp. 2161–2172, 04 2021.

[54] M. B. Alejo, "Unconstrained ear recognition using transformers," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 07, no. 04, pp. 326–336, 2021.

[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, oct 2017, pp. 2980–2988.

[56] W. Abdulla, "Segmentation with Mask R-CNN and TensorFlow explained by building a color splash filter." 2018, accessed on 08.01.2021. [Online]. Available: https://engineering.matterport.com/

[57] U. of Ljubljana, "Ear Recognition Research," 2018, accessed on 08.01.2021. [Online]. Available: http://awe.fri.uni-lj.si/datasets.html

[58] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1, 2015.

[59] L. M. Dinca and G. P. Hancke, "The fall of one, the rise of many: A survey on multi-biometric fusion methods," *IEEE Access*, vol. 5, pp. 6247–6289, 2017.

[60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, accessed on 08.01.2021. [Online]. Available: https://www.tensorflow.org/