Constellation-Based Deep Ear Recognition

Dejan Štepec*, Žiga Emeršič*, Peter Peer, Vitomir Štruc

Abstract This chapter introduces COM-Ear, a deep constellation model for ear recognition. Different from competing solutions, COM-Ear encodes global as well as local characteristics of ear images and generates descriptive ear representations that ensure competitive recognition performance. The model is designed as dual-path convolutional neural network (CNN), where one path processes the input in a holistic manner, and the second captures local images characteristics from image patches sampled from the input image. A novel pooling operation, called patch-relevant-information pooling, is also proposed and integrated into the COM-Ear model. The pooling operation helps to select features from the input patches that are locally important and to focus the attention of the network to image regions that are descriptive and important for representation purposes. The model is trained in an end-to-end manner using a combined cross-entropy and center loss. Extensive experiments on the recently introduced Extended Annotated Web Ears (AWEx)

Dejan Štepec
XLAB d.o.o.,
Pot za Brdom 100, SI-1000 Ljubljana, EU
e-mail: dejan.stepec@xlab.si

Žiga Emeršič (Corresponding author) ✉ · Peter Peer
Computer Vision Laboratory,
Faculty of Computer and Information Science,
University of Ljubljana,
Večna pot 113, SI-1000 Ljubljana, EU
e-mail: {ziga.emersic,peter.peer}@fri.uni-lj.si

Vitomir Štruc
Laboratory of Artificial Perception, Systems and Cybernetics,
Faculty of Electrical Engineering,
University of Ljubljana,
Tržaška cesta 25, SI-1000 Ljubljana, EU
e-mail: vitomir.struc@fe.uni-lj.si

* These authors contributed equally to this work.

dataset demonstrate the competitiveness of COM-Ear compared to existing ear recognition models.

Key words: Ear biometrics, Ear recognition, Part-based models, Constellation model, Convolutional neural networks.

1 Introduction

Ear recognition refers to the task of recognizing people from ear images using computer vision techniques. Ears offer appealing characteristics when used in automated recognition systems, such as the ability to distinguish identical twins [48], the potential to supplement other biometric modalities (e.g., faces) [65, 72] or the ability to capture images from a distance and without explicit cooperation of the subjects one is trying to recognize.

Person recognition based on ear images has seen steady rise of popularity over recent years. Nevertheless, despite significant advancements in this area and the shift towards deep-learning-based models, nuisance factors such as ear occlusions and the presence of ear accessories still adversely affect performance of existing recognition models. Moreover, while research on ear recognition has long been focused on recognition problems in controlled imaging conditions, the recent switch to unconstrained image acquisition conditions brought about new challenges related to extreme appearance variability caused by blur, illumination, and view-direction changes, which were thus far not considered problematic for ear recognition. These extreme conditions pose considerable challenges to existing ear recognition models and have so far not been addressed properly in the literature.

Existing approaches to address the challenges encountered in unconstrained imaging conditions focused mostly on fine-tuning existing deep learning models. In a recent competition, organized around the problem of unconstrained ear recognition [28], for example, most participants used established models, such as VGG-16 or Inception-ResNets pretrained on ImageNet data as a baseline and then fine-tuned the models on the training data of the competition. Other recent deep learning solutions [5, 23, 34, 52, 77] in this area also followed a similar approach and used pre-existing models or employed transfer learning and domain adaptation techniques to adapt the models for ear recognition. A common aspect of these works is the fact, the model were not designed specifically for ear recognition and processed ear images holistically ignoring the particularities and existing problems of ear recognition technology.

In this chapter, we take a step further and present a novel (deep) constellation model for ear recognition (COM-Ear) that addresses some of the problems seen with competing ear recognition approaches. As we show in the experimental section the model ensures state-of-the-art recognition performance for unconstrained ear recognition and exhibits a significant increase in robustness to the presence of partial ear occlusions (typically caused by ear accessories) compared to other techniques in

this area. The proposed COM-Ear model is designed around a Siamese architecture that takes an ear image and local image patches (sampled from a fixed grid) as input and generates an image representation that encodes both global and local ear characteristics at the output. To generate the output image representation features from different patches are first combined using a newly proposed pooling operation, called patch-relevant-information pooling (or PRI-pooling), and then concatenated with the global images features. Our constellation model is not limited to specific model topologies and can be built around any recent deep-learning model. We, hence, evaluate and analyze different backbone models (i.e., ResNet18, ResNet50 and ResNet152) for the implementation of COM-Ear. We train the proposed constellation model end-to-end using a combination of cross-entropy and center losses as our learning objectives. To the best of our knowledge, this is the first attempt at designing and training constellation deep model in the field of ear recognition.

We evaluate the model in rigorous experiments on the challenging Extended Annotated Web Ears (AWEx) dataset [28, 29]. To demonstrate the robustness of the COM-Ear model to ear accessories and occlusions, we perform additional experiments using an artificially generated dataset [24] of images where accessories are added to the ear images. The results of our experiments show that the proposed COM-Ear model is a viable solution for the problem of ear recognition that ensures state-of-the-art performance and exhibits a considerable level of robustness to various factors adversely affecting the recognition accuracy of competing approaches.

To summarize, the main contributions of this chapter are the following:

- We present and describe COM-Ear, the first ever deep constellation model for the problem of ear recognition that ensures state-of-the-art performance on the most challenging dataset of ear image available.
- We introduce a novel pooling operation, called patch-relevant-information pooling (or PRI-pooling) that is able to select features from image patches that are locally important and integrate it into the COM-Ear model.
- We make all code, models, and trained weights publicly available via `http://ears.fri.uni-lj.si` and provide a strong baseline for future research in the field of unconstrained ear recognition.

The rest of the paper is structured as follows. In Section 2 we present the background and related work. Here, we discuss techniques for ear recognition in constrained setting as well as methods focusing on ear recognition in the wild. In Section 3 we describe the proposed constellation model in detail and elaborate on the idea behind the model, its architecture, and training procedure. In Section 4 we present a rigorous experimental evaluation of COM-Ear and discuss results. We also present a qualitative analysis to highlight the characteristics of our model. We conclude the chapter in Section 5 with some final remarks and direction for future work.

2 Related Work

The literature on automated ear recognition is extensive, starting with the early geometry-based recognition techniques to more recent deep learning model. The field has also seen a shift recently away from ear datasets from constrained environments towards unconstrained settings, which more closely reflect real-world imaging conditions. In this section we briefly survey the most important work in the field to provide the necessary context for our contributions. For a more complete coverage of ear recognition technology, the reader is referred to some recent surveys [29, 55]

2.1 Ear Recognition in Constrained Conditions

Until recently, most of the research on ear recognition was focused on controlled imaging conditions where the appearance variability of ear images was carefully controlled and typically limited to small head rotations and minute changes in the external lighting conditions [29]. Techniques for ear recognition (from 2D color/intensity images) proposed in the literature during this period can conveniently be grouped into the following categories: [29, 55]: *i)* geometric approaches, *ii)* global (holistic) approaches, *iii)* local (descriptor-based) approaches, and *iv)* hybrid approaches.

Geometric approaches dominated the early days of ear recognition [2, 29] and were often aided by manual intervention. Techniques from this group rely on certain geometric properties of ears and exploit relationships between predefined parts of ears. The first fully automated ear recognition procedure exploiting geometric characteristics was presented by Moreno et al. [**?**] and made use of ear geometric description and a compression network. Some other examples of geometric ear recognition include [11, 15, 16, 46, 59]. One of the more recent publications using a geometric approach is the work of Chowdhury et al. [18]. Here, the authors present a complete recognition pipeline including an Adaboost-based ear detection technique. The Canny edge detector is employed to extract edge features from images and ear comparisons are done using similarity measurements.

The second group of ear recognition approaches, global (also referred to as holistic) techniques, exploit the global appearance of the ear and encode the ear structure in a holistic manner. Even though techniques from this group seems to represent an obvious way to tackle ear recognition and improve upon the geometric methods, they are relatively sensitive to variations in illumination, pose or presence of occlusions. Some of the earliest examples of global approaches include the work from Hurley et al. [37] that relied on the Force Field Transform to encode ear images, methods using Principal Component Analysis [14,67] and others [1,3,8,47,71,74,76].

The third group of approaches, local techniques, extracts information from local image areas and use the extracted information for identity inference. As emphasized in the survey by Emersic et al. [29], two groups of techniques can in general be considered local-based: techniques that first detect interest points in the image and then compute descriptors for the detected interest points, and techniques that compute

descriptors densely over the entire images based on a sliding window approach. Examples of techniques from the first group include [7,12,58]. A common characteristic of these techniques is the description of the interest points independently one from another, which makes it possible to design matching techniques with robustness to partial occlusions of the ear area. Examples of techniques from the second group include [6,9,13,17,39,42,69]. These techniques also capture the global properties of the ear in addition to the local characteristics, which commonly result in higher recognition performance, but the dense descriptor-computation procedure comes at the expense of the robustness to partial occlusions. Nonetheless, trends in ear recognition favored dense descriptor-based techniques primarily due to their computational simplicity and high recognition performance.

The last group of ear recognition approaches, hybrid methods, typically describe the input images using both, global as well as local information. Techniques from this group first represent ear images using some local (hand-crafted) image descriptor (e.g., SIFT, BSIF or HOG) that captures local image properties and then encode the extracted descriptor using a global (holistic) subspace projection technique, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) or other related techniques [55]. Hybrid approaches and (powerful) local-descriptor-based methods represented the state-of-the-art in ear recognition for a considerable period of time and were only recently outperformed by deep learning based models [25,28]. The model introduced in this chapter builds on these techniques and similar to hybrid methods also tries to capture global ear characteristics as well as local ear details. In this sense it is related to hybrid techniques from the literature, but offers significant performance improvements as evidenced by the results presented in Section 4.

In addition to the ear recognition approaches described above, some works focus on solving specific issues regarding ear recognition, such as occlusions or alignment [54,60,61,68,75,78,79]. Furthermore, existing research related to ear recognition also studies multi-modal biometrics systems that incorporate ear images into the recognition procedure [4,31,53], different data-acquisition techniques, such as light-field cameras [62] or other ways of ear-based recognition that do not rely on visual information [45].

2.2 Ear Recognition in Unconstrained Conditions

More recent work on ear recognition is increasingly looking at unconstrained image acquisition conditions, where the appearance variability of ear images is considerably greater from what is seen in constrained settings. Several ear datasets for research in ear recognition have been proposed towards such unconstrained scenarios, starting with the Annotated Web Ears (AWE) [29], The Unconstrained Ear Recognition Challenge (UERC) dataset [28] and others. Research on these datasets is dominated by deep learning models based primarily on convolutional neural networks (CNNs), while techniques using local, hand-crafted descriptors are far and few between.

Numerous deep learning models have been presented for ear recognition over the course of the last two years [23, 28, 30, 34, 52, 64, 66, 77], all significantly outperforming local-descriptor-based and hybrid methods in the most challenging scenarios [26–28, 66]. One of the earliest approaches to ear recognition using deep neural used aggressive data augmentation to suffice the training needs [27]. Another example of one of the earliest uses of deep learning was presented by Galdámez et al. [32]. Here, the authors used a Haar-based detection procedure, but used a CNN for recognition. Another work using deep neural networks to facilitate ear recognition was presented by Tian and Mu [66]. However, the authors focused on datasets captured in the constrained environments and did not evaluate the performance of their model on more recent datasets captured in unconstrained conditions. The work of Eyiokur et al. [30] introduced an new (constrained) dataset of ear images, which was used to train AlexNet, VGG and LeNet for ear recognition. The models were later fine-tuned on the UERC data and tested in unconstrained conditions. Another work of using AlexNet for ear recognition includes [5]. Dodge et al. [23] developed a deep neural networks for ear recognition and tested their model on the unconstrained AWE and CVLE datasets. In the work of Ying et al. [73] a shallow CNN architecture was presented, but the ear dataset used for testing was not described. In the work of Zhang et al. [77] the authors presented a new dataset of ear video sequences captured with a mobile phone. For recognition the authors used the fully convolutional SPPNet (Spatial Pyramid Pooling Network) capable of accepting variable sized input images.

In this chapter we build on the outlined body of work and introduce a novel deep-learning model for ear recognition in unconstrained conditions. Similar to existing work our model relies on a CNN to learn an descriptive representation of ear images, but unlike competing solutions does not capture only global information, but also local image cues that may be important for recognition purposes. Moreover, due to the design of the model, the amount of local information that is added to the computed ear representation is adaptively added to the learned descriptor depending on the given content of the input images.

3 COM-Ear: A Deep Constellation Model for Ear Recognition

In this section we present the general structure and the idea behind our (deep) constellation model for ear recognition (COM-Ear). COM-Ear represents, to the best of our knowledge, the first part-based deep learning model for ear recognition. Compared to competing deep learning models, which process images in a holistic manner, COM-Ear also takes into account local image information and combines it with holistic features, which (as we show in the experimental section) improves robustness to occlusions and results in state-of-the-art results on established benchmarks.

3.1 Motivation

Deep learning models, and particularly convolutional neural networks (CNNs), which learn discriminative image representations from the whole input image have recently achieved great results in all areas of biometrics, including ear recognition [27]. However, global methods that process input images as a whole are in general sensitive to illumination changes, occlusions, pose variations and other factors typically present in unconstrained real-world environments. Local methods that extract discriminative information from local image regions, on the other hand, are by typically more robust to occlusions and related nuisance factors and represent a viable alternative to global approaches.

CNNs are by design global in nature, but with their hierarchical design and characteristics, such as convolutional kernels with local connectivity, high dimensionality and non-linearity, are also capable of encoding local discriminative information exceptionally well. However, because of the nature of operations in CNNs, this local information is aggregated and propagated along the model layers and the amount of local information that is preserved is limited to the most discriminative parts of the input. In unconstrained setting different parts of the input might be occluded, differently illuminated and especially in the case of ear recognition, different accessories might be present on different parts of the ear which greatly affects the performance of such methods. With COM-Ear we try to address these issues and present a deep model with the following characteristics:

- **Aggregation of global and local information:** We design our model to follow the approach of hybrid techniques, which were popular before the era of deep learning and utilized both global and local information. Compared to traditional hybrid approaches, we design COM-Ear in a fully convolutional way such that both, global and local information is captured by a single CNN model, resulting in a highly descriptive and discriminative image representation that can be used for identity inference.
- **Selective attention to image parts:** Local features are combined in a novel way which gives the model the capability to focus it's attention on locally important discriminative parts. The proposed model also offers a straight-forward way of exploring the importance of each image part which contributes towards high explainability of the proposed model.
- **Robustness to occlusions:** Our method is designed to specifically address issues known to be problematic for ear recognition. Specifically, it offers a natural way of decreasing sensitivity to partial occlusions of the ear that are typically caused by the presence of ear accessories.

3.2 Overview of COM-Ear

The architecture of the COM-Ear model is presented in Fig. 1. The model is designed as a dual-path architecture, where the first path (marked 1) in Fig. 1) encodes global information about the ear appearance and the second path (marked 2) in in Fig. 1) captures local image cues to supplement the information extracted from the global processing path. For both paths a CNN-based feature extractor is used as the backbone model. Below we describe all parts of the model in detail.

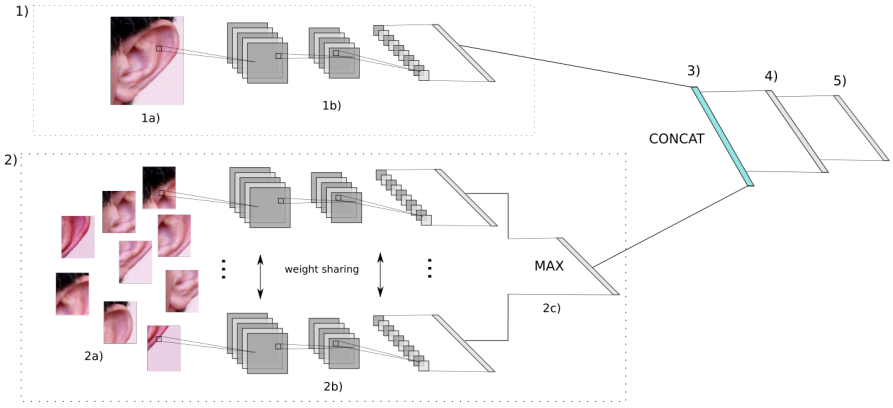


Fig. 1: Overview of the proposed Deep Constellation Model for Ear Recognition (COM-Ear). The model is designed as a two-path architecture. First path, denoted by 1) represents the global processing path that encodes the input image at a holistic level using a backbone CNN-based feature extractor denoted by 1b). The second path, denoted by 2) represents the local patch-based processing path which extracts features from local image patches via the backbone CNN, denoted by 2b). Local features are then combined with the PRI-Pooling operator, denoted by 2c). Global and local features are concatenated in 3) and used in the fully connected layers 4) and 5) to predict outputs and to compute losses during training.

3.2.1 The Global Processing Path

The input to the global processing path (1) is the whole ear image (1a) as shown in Fig. 1. From this input, a feature representation is computed using a backbone CNN (1b). While different models could be used for this task, we select different ResNet incarnations (ResNet-18, ResNet-50, ResNet-152) [35] because of their state-of-the-art performance in many vision tasks and the fact that open source implementations are readily available. To make full use of the ResNet models, d-dimensional features are taken from the last pooling layer of the models and the pooling operation is

replaced by adaptive pooling to make the model applicable to differently sized input images.

3.2.2 The Local Processing Path

For the second processing path (2), the input image is first split into N smaller patches (2a). The patches are then fed to the COM-Ear model for processing. The local processing path is designed as a parallel Siamese architecture (2b) with shared model parameters and the same backbone feature extractor as used in the global processing path (1b). Feature representations are extracted from each of the patches and aggregated using max-pooling with kernel size of 1 (i.e. max operation along patch dimension) (2c). We decided for a Siamese architecture to reduce the number of parameters that need to be learned during training and to decrease the possibility of over-fitting. With our scalable design we are able to exploit information from a variable number of input patches with no influence on the number of parameters and without changes to the network topology.

To aggregate information from the local features we use a novel pooling procedure we refer to *patch-relevant-information* pooling or PRI-pooling for short. The idea behind PRI-pooling is to use only the most relevant information from the local image patches in the final image representation. For the proposed pooling procedure, every patch is first passed through the Siamese CNN ensemble to get a set of N corresponding d-dimensional feature representations - similar to TI-Pooling [43]. A max pooling operation is then applied on the N feature vectors along the patch dimension to generate the final aggregated d-dimensional feature representation for the local processing path.

PRI-pooling is applied on the feature vector in order to obtain features that are locally important. We argue that with patches as an input to the PRI-pooling topology the features learned capture local information that is relevant and may supplement the global features produced by the global processing path of the COM-Ear model. With this approach the network can automatically infer, which parts of the input image are important and, vice versa, these regions can be identified from the composition of the final aggregated feature vector produced by the local processing path. Thus, the PRI-pooling operation gives our COM-Ear model as a level of explainability not available with purely holistic competitors. We provide a few qualitative examples of this explainability in the experimental section.

3.2.3 Combining Information

Global and local features are combined in the final part of the COM-Ear model by simple concatenation. Given that the feature representation from each model path is d-dimensional the combined feature representation comprises $2d$ elements. As we show in the experimental section, both types of features are important for the performance of the COM-Ear model, as with holistic features or local features alone

we are not able to match the performance of the combined representation. Especially, the performance of the local features is observed to be limited when no holistic information is used. We believe the reason for this setting is that global processing path of our model affects the learning of the local processing path due to end-to-end learning procedure designed for the proposed architecture. The combined holistic and global features are passed through another series of fully connected layers where the final layer is softmax layer upon which a loss is defined during training. The softmax classification layer can also be used during run-time for closed-set recognition experiments.

3.3 Model Training

To train the COM-Ear model, we design an end-to-end learning procedure and a combined training objective, defined as follows:

$$L_{total} = L_S + \lambda L_C, \tag{1}$$

where L_S denotes the cross-entropy loss defined on the softmax COM-Ear layer, L_C stands for the center loss defined on features that represent inputs to the first fully connected layer, and λ represents a hyper parameter that balances the two losses. The motivation behind the center loss is to enhance the discriminative power of the learned features. The cross-entropy loss forces the deep features of different classes to be separable, while the center loss efficiently pulls the deep features closer to their corresponding centers (learned on mini batches). In this way inter-class feature differences are enlarged and intra-class feature variations are reduced effectively making features more discriminative. Our preliminary experiments during the design phase suggested that the inclusion of the center loss is highly important as the results without center loss were significantly less convincing.

3.4 Implementation Details

We implement our model using PyTorch and use built-in implementations of ResNet (ResNet-18, ResNet-50, ResNet-152) models as the backbone CNN-based feature extractor. All backbone models are used with pretrained weight on the ImageNet dataset. We modify the models in order to accept arbitrary sized input images and replace all average pooling layers with adaptive average pooling operations. The outputs of the adaptive average pooling layers are used as features in both the global as well as local processing paths.

As described above, local features from the patches are aggregated with the proposed PRI-pooling operation, which is implemented as an element-wise maximum over patch dimension. Both, the aggregated local feature and the global holistic

features are of the same dimension (512 for ResNet-18 and 2048 for ResNet-50 and ResNet-152) and are combined using a simple concatenation operation. Concatenated inputs are transformed via the fully connected layer of the same output dimension as the input (e.g. 512 for ResNet-18) and this represents the input to the final fully connected softmax layer.

4 Experiments and Results

In this section we describe the experiments performed to highlight the main characteristics of the proposed COM-Ear model. Since our focus is the ability of the proposed model to perform well in unconstrained environments we used Extended Annotated Web Ears (AWEx) for our experiments. However, the dataset contains a limited number of images. Based on our previous experience and findings [27] we used severe data augmentation to stimulate training and to prevent overfitting. The AWEx is then used to train and evaluate different variations of the model – reduced patch size and omission of center loss. We compare our model directly to some of the state-of-the-art approaches. Furthermore, to evaluate the performance of our proposed COM-Ear model as well as possible, we also present a comparison with the deep-learning approaches submitted to the 2017 Unconstrained Ear Competition Challenge [28] - a recent group-benchmarking effort of ear recognition technology applied to data captured in unconstrained conditions. Additionally, we also evaluate the robustness of our model in regard to one of the most problematic aspects of ear recognition – occlusions. For this part of our analysis we generate a synthetic dataset with artificial ear accessories superimposed over ear images. Lastly we present a in depth qualitative analysis, where we first visualize the impact of the proposed patch-based processing by analyzing performance of separate parts and then visually compare ranking performance of the proposed model vs the performance of the deep-learning approaches from the 2017 Unconstrained Ear Recognition Challenge [28].

4.1 Experimental Datasets

We conduct experiments on the Extended Annotated Web Ears (AWEx) dataset, which represents one of the largest datasets of unconstrained ear images available. Images from the dataset were gathered from the web and therefore exhibit a significant level of variability due to differences in head rotations, illumination, age, gender, race, occlusion and other factors. A few example images intended to highlight the difficulty of the dataset are presented in Fig. 2. The ear images in the dataset are tightly cropped and are not normalized in size. A total of 336 subjects and 4004 images is available in the AWEx dataset and is used in our experiments.

We use the dataset in identification experiments and follow various experimental protocols to be able to compare our model with published results from the literature.

Fig. 2: Sample images from the AWEx dataset. As can be seen, the images exhibit a wide range of appearance variability due to different resolution, ethnicity, varying levels of occlusion, presence/absence of accessories, head rotations and other similar factors.

These protocols include two evaluation protocols from UERC 2017 [28] which allows for a direct comparison with approaches from the challenge.

4.2 Performance Metrics

As already indicated above, we perform identification experiments to evaluate the COM-Ear model and compare it to existing approaches. Identification aims at predicting the identity of the given ear image, as opposed to verification experiments where the prediction is binary – whether the observed ear-image belongs to a given subject or not.

To measure performance in our experiments, we report the following performance metrics, wherever possible:

- The rank one recognition (rank-1): is the percentage of probe images, for which an image of the correct identity was retrieved from the gallery as the top match. If there are multiple images per class available in the gallery, the most similar image is selected and used for the rank calculation.
- The rank five recognition (rank-5): is the percentage of probe images, for which an image of the correct identity was among the top five matches retrieved from

the gallery. Same, as for rank-1, if there are multiple images per class in the gallery, the most similar sample is considered. This procedure applies for all the rank calculations.

- The Area under the CMC curve (AUC): is the normalized area under the Cumulative Match Score Curve (CMC), which is similar to the standard AUC measure typically computed for Receiver Operating Characteristic (ROC) curves. This metric measures the overall performance of the tested recognition model and is commonly used in identification experiments [28].

These identification metrics are widely used in literature and have, therefore also been selected for this work. For all described performance metrics a higher value means better performance. The rank values range from 0 to the number of classes present in the test set, whereas for the AUC score, values range between 0 and 1 and denote the fraction of the surface area under the CMC curve. For a more in-depth explanation of the metrics used in the experiments, we refer readers to [38].

4.3 Training Details

We train the COM-Ear model using images from the AWEx dataset. For the training procedure we use the training objective in (1) with a value of $\lambda = 0.003$, as used in original paper [70], to balance the impact of the cross-entropy and center losses. We set the learning rate to 0.01 for the cross-entropy loss and to 0.5 for center loss. We train the model for 100 epochs with stochastic gradient descent (SGD) and a step size of 50, decay rate of 0.1 and a batch size of 32 input images and their corresponding patches. We sample patches from the input images in a grid-like fashion with overlap[2]. A summary of the hyper-parameters used during training is given in Table 1.

Table 1: Hyper-parameters used during training.

number of epochs	100
weight decay	0
learning rate for loss function	0.01 (0.9 momentum)
learning rate for center loss	0.5
λ (as defined above)	0.003

To avoid overfitting, we perform data augmentation to increase the variability of the data. The importance of data augmentation in the ear recognition domain was first mentioned in [27]. However, compared to [27] we used online augmentations (i.e. augmenting the data on the fly) so that the network almost never sees the exact same image multiple times in order to improve generalization performance.We perform

[2] Note that we study the influence of the size of the patches in the experimental section

data augmentation with the Imgaug [3] Python library and use the following image transformations:

- horizontal flipping,
- blurring with Gaussian filters with σ in the range (0, 0.5),
- scaling by a factor in the range (0.9, 1.2), and
- rotation in the range $\pm 30^{\circ}$.

All listed operations are performed in random order and each operation is applied with 50% chance. With this setting we leave the chance that there could be no augmentations applied at all - albeit with a very low probability. The images are also normalized with per channel mean and standard deviation values from ImageNet [21] as is general practice.

Image patches are cropped after performing augmentations on the image so that both the image and patches are transformed in the same way. With this we ensure that holistic and local models are looking at the same input.

4.4 Ablation Study

In our first series of experiments we investigate the impact of some of our design choices when developing the COM-Ear model. For this ablation study we, therefore, focus on separate parts of the proposed COM-Ear model and observe how specific design choices affect the performance of our model. For this experiment we follow the experimental protocol from [25] and split the available data from the AWEx dataset into two, subject disjoint sets, i.e.:

- *A training set* comprising 1804 images of 116 subjects. These images are used to learn the parameters of the COM-Ear model (and it's variants) and monitor training progress via a validation set during the learning procedure.
- *A testing set* comprising 2200 images of 220 subjects intended for final performance evaluation. Images from the set are used to compute performance metrics and report results.

To allow for open-set identification experiments, we perform network surgery on the COM-Ear model and use the $2d$-dimensional concatenated global and local features as the descriptor for the given input ear image. To measure similarity between ear descriptors we compute cosine similarities. For the experiments, we use an initial image size of 224×224 pixels and a patch size of 112×112 pixels. Patches are sampled with a 50% overlap resulting in a total of 9 patches for the local processing path of COM-Ear.

Using the above protocol, we first explore the performance of the backbone ResNet feature extractors and compare the performance of different ResNet variants, i.e., ResNet-18, ResNet-50 and ResNet-152. We train all models on our training data

[3] https://github.com/aleju/imgaug

using the same loss as for the COM-Ear model (see Eq. (1)) and use features from the penultimate model layer with the cosine similarity for recognition. The results in Table 2 show no significant difference in the performance of the backbone models. We, therefore, select ResNet-18 as the final backbone model for COM-Ear due to it's light-weight architecture compared to the other two ResNet variants.

Next, we report results for the COM-Ear model obtained with and without the use of center-loss. We observe that the performance of COM-Ear drops by a larger margin when no center-loss is used, which points to the importance of the combined loss during training. Additionally, when looking at the performance difference between the backbone ResNet-18 model and COM-Ear, we see that the addition of the local processing path significantly improves performance, as the rank-1 recognition improved from 26.1% (for ResNet-18) to 31.1% (for COM-Ear).

Finally, we report results for COM-Ear using smaller patches of size 56×56 pixels – patches are still sampled from the input image with a 50% overlap. In comparison with the initial patch size of 112×112 pixels sampled with a 50% overlap results are worse. These results suggest that patches need to be of a sufficient size in order to carry enough context to be informative. Smaller patches can also carry background information which is not beneficial for recognition purposes and may also introduce ambiguities among different subjects. Examples of 112×112 pixel input patches with 50% overlap can be viewed in Fig. 1.

Table 2: Ablation study for the COM-Ear Model.

Method	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
ResNet-18	26.1	52.2	92.7
ResNet-50	26.1	50.8	92.6
ResNet-152	26.1	49.9	92.4
COM-Ear	**31.1**	**54.6**	**93.2**
COM-Ear [no center loss]	27.1	52.5	92.6
COM-Ear [patch size / 2]	29.4	52.1	91.6

4.5 Performance Evaluation against the State-of-the-art

In our next series of experiments we benchmark the COM-Ear model against state-of-the-art models from the literature. We conduct two types of experiments to match the experimental protocols most often used by other researchers.

Table 3: Comparative evaluation of the COM-Ear Model.

Method	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
ResNet-18 [35]	24.5	48.5	91.4
ResNet-50 [35]	25.9	49.9	92.0
ResNet-152 [35]	26.1	52.8	92.6
MobileNet ($\frac{1}{4}$) [36]	17.1	36.1	88.0
MobileNet ($\frac{1}{2}$) [36]	16.0	38.5	88.5
MobileNet (1) [36]	26.9	50.0	91.8
LBP [49]	17.8	32.2	79.6
HOG [19]	23.1	41.6	87.9
DSIFT [44]	15.2	29.9	77.5
BSIF [40]	21.4	35.5	81.6
LPQ [50]	18.8	34.1	81.0
RILPQ [51]	17.9	31.4	79.8
POEM [69]	19.8	35.6	81.5
COM-Ear	**31.5**	**55.9**	**93.3**

4.5.1 Comparison with Competing Methods

In the first experiment of this series we use the same protocol as during the ablation study. This protocol is taken from [25] and we report results from this publication for comparison purposes. Specifically, we include results for dense-descriptor-based methods relying on Local Binary Patterns (LBPs [10, 29, 33, 56, 57]), (Rotation Invariant) Local Phase Quantization Features (RILPQ and LPQ [50, 51]), Binarized Statistical Image Features (BSIF [29, 40, 56]), Histograms of Oriented Gradients (HOG, [19, 20, 29, 56]), the Dense Scale Invariant Feature Transform (DSIFT, [22, 29, 42]) and Patterns of Oriented Edge Magnitudes (POEM, [29, 69]). For deep learning based models, we report results for ResNet-18, ResNet-50 and ResNet-152 (taken from [25]). Additionally, we provide results for the MobileNet model, which represent a light-weight CNN architecture, developed with mobile and embedded vision applications in mind. The architecture uses two main hyperparameters that efficiently trade off between latency and accuracy [36]. These hyper-parameters allow to tweak the size of the model in accordance with the problem domain and use-case scenarios. In this work we evaluate three such versions with different width multipliers. The lower the value the more lightweight the model, the higher the value (highest, being 1) the heavier the footprint. In Table 3 we report results for three levels of multipliers: $\frac{1}{4}$, $\frac{1}{2}$ and 1 [36].

The results of this experiment are presented in Table 3 and Fig. 3. We observe that COM-Ear achieves the best overall results, improving upon the state-of-the-art by a large margin. With a rank one recognition rate of 31.05% it significantly outperforms all traditional feature extraction methods as well as all tested deep learning based models. The closest competitor is MobileNet (1) with a rank one recognition rate of 26.9%. Descriptor-based methods are less convincing with the best performing

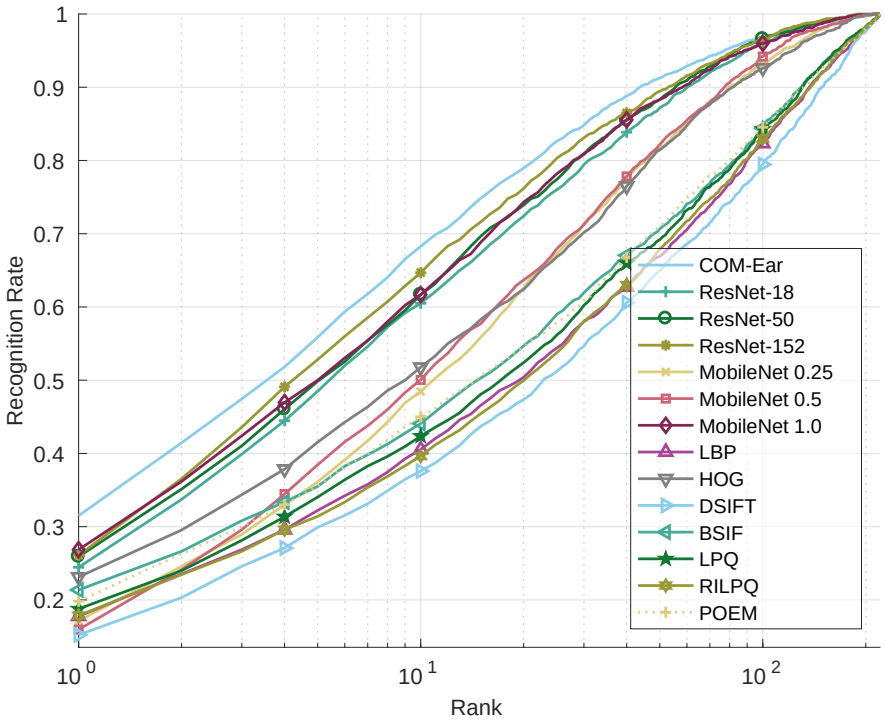


Fig. 3: CMC curves of the comparative evaluation of the COM-Ear Model. The results are presented in logarithmic scale to better visualize the performance differences at the lower ranks, which are more important from an application point of view. The figure is best viewed in color.

method from this group achieving a rank-1 recognition rate of 23.1%, 8% less (in aboslute terms) than the proposed COM-Ear model.

4.5.2 Comparison With Results From the 2017 Unconstrained Ear Recognition Challenge (UERC)

In the next experiments we compare COM-Ear on the data and experimental protocol used in the 2017 Unconstrained Ear Recognition Challenge (UERC 2017). UERC 2017 was organized as a group benchmarking effort in the scope of the 2017 International Joint Conference on Biometrics (IJCB 2017) and focused on accessing ear recognition technology in unconstrained settings. The challenge was conducted in part on the AWEx dataset using a slightly different protocol as used in the previous section. The reader is referred to [28] for details on the protocol. Several groups participated in the challenge and submitted results. Here, we include results for all

Table 4: Summary of deep learning based approaches from UERC 2017 included in the comparison. The table provides a short description of each approach, information on whether ear alignment and flipping was performed and the model size (if any). See [28] for details.

Approach	Description	Descriptor type	Alignment	Flipping
IAU	VGG network (trained on ImageNet) and transfer learning	Learned	No	No
ICL	Deformable model and Inception-ResNet	Learned	Yes	Yes
IITK	VGG network (trained on the VGG face dataset)	Learned	No	Yes
ITU I	VGG network (trained on ImageNet) and transfer learning	Learned	No	Yes
ITU II	Ensemble method (LBP + VGG-network)	Learned + Hand-crafted	No	Yes
LBP-baseline	Descriptor-based (uniform LBPs)	Hand-crafted	No	No
VGG-baseline	VGG network trained solely on the UERC training data	Learned	No	No

deep-learning-based methods from UERC 2017 - briefly summarized in Table 4 - and for the three ResNet variants also tested in the previous sections.

Table 5: Comparison with results from the Unconstrained Ear Recognition Challenge (UERC) [28]. The results were generated on the testing split of the AWEx dataset.

Method	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
ICL [28]	5.3	14.8	71.17
IAU [28]	38.5	63.2	94.0
IITK [28]	22.7	43.6	86.1
ITU-I [28]	24.0	46.0	89.0
ITU-II [28]	27.3	48.3	87.7
LBP-baseline [28]	14.3	28.6	75.9
VGG-baseline [28]	18.8	37.5	86.6
ResNet-18	28.0	57.1	93.0
ResNet-50	22.1	46.8	90.5
ResNet-152	15.9	40.7	88.8
COM-Ear	**41.8**	**67.7**	**94.7**

The results of the comparison are presented in Table 5 and Fig. 4. Similarly to the results from Table 3 we observe that among the ResNet models, ResNet-18 performs the best (rank-1 = 28%). The overall top performer is again COM-Ear, which achieves state-of-the-art results, improving upon the best results included in the comparison (IAU) by more than 3% in terms of rank-1 and more than 5% rank-5

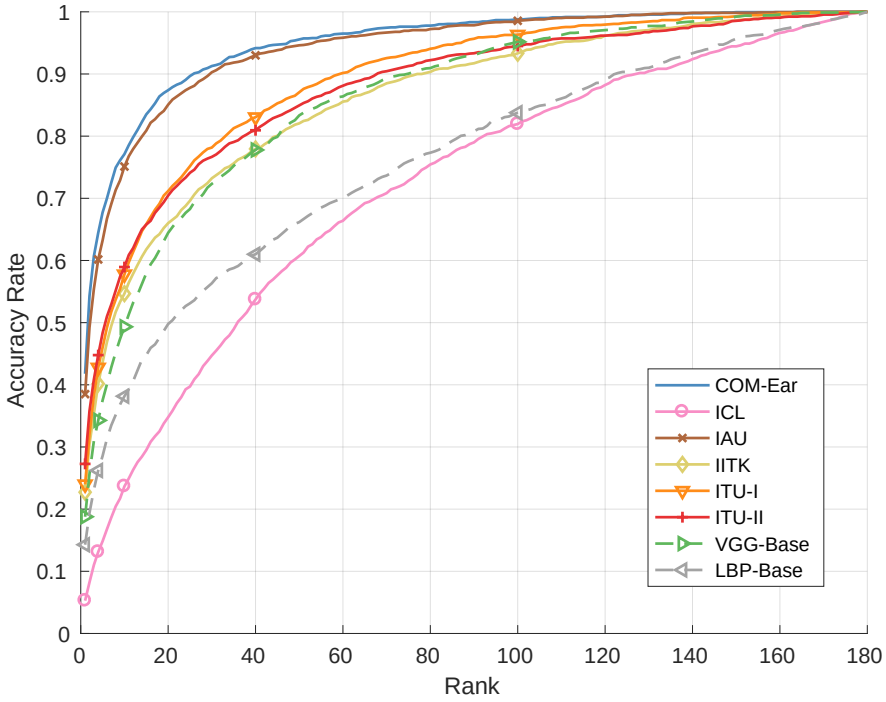


Fig. 4: CMC curves of the direct comparison to the UERC 2017 [28] recognition results. The results were generated on the testing split of the AWEx dataset and are shown in linear scale.

recognition results. The best performing entry from Islamic Azad University (IAU) in UERC 2017 was built around a VGG-16 architecture [63] and transfer learning. The idea of the IAU approach was to leave part of the pretrained VGG-16 model as is (i.e., with frozen weights), while retraining other parts of the model that are relavant for transfering to the new domain, i.e. ear recognition. Specifically, the authors added two fully-connected layers on top of the 7th layer of the pretrained VGG model. Only the newly added FC layers were trained on the UERC data. Learning only certain layers while leaving other layers untouched (e.g.. learned only on ImageNet) is beneficial especially in the case of smaller datasets like the one used for UERC 2017, as it prevents overfitting and thus results in features that generalize better to the new task. In our case we used center loss to make the learned features more discriminative and learn a descriptive model using limited training data.

In the next experiment, we evaluate how the proposed COM-Ear model scales with larger probe and gallery sets. For this experiments we use the scale experimental protocol employed for the scale experiments in UERC 2017. The results of this test are shown in the form of CMC curves in Fig. 5. The curves were generated using 7,442 probe images belonging to 1,482 subjects and 9,500 gallery images of 3,540

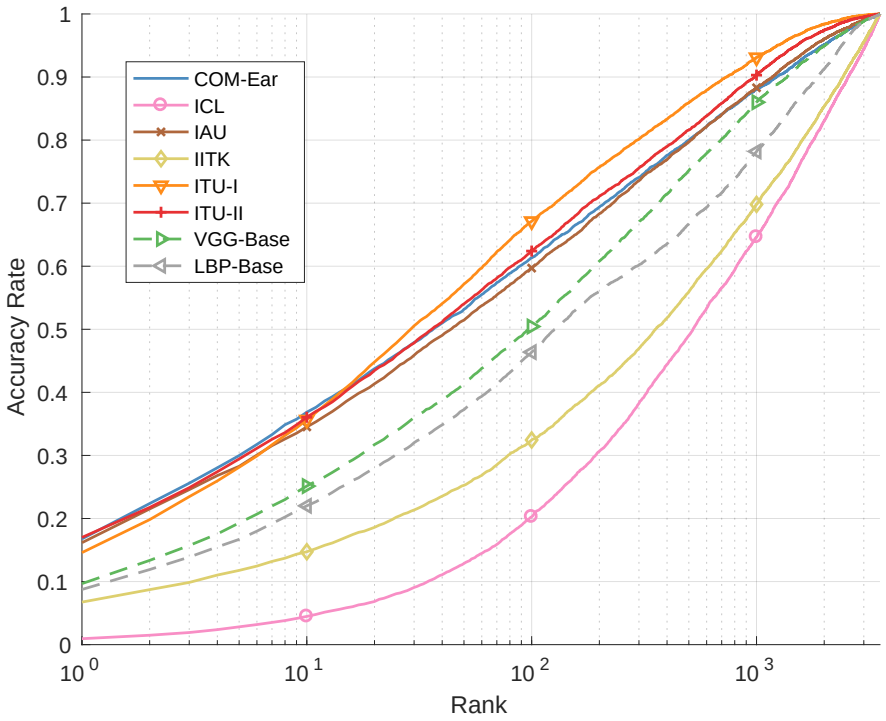


Fig. 5: CMC curves of the comparison to the UERC 2017 [28] recognition results. The results were generated on the complete test dataset of UERC containing all 3,540 subjects of the AWEx dataset and multiple distractor identities. The results are again shown in logarithmic scale to highlight the performance differences at the lower ranks. The figure is best viewed in color.

subjects. The gallery also contained identities that were not in the probe set. These samples act as distractors for the recognition techniques [28, 41]. The numerical results in Table 6 show that the proposed model perform comparable to the ITU-II approach in terms of rank-1 and rank-5 recognition rates and is very competitive even when a large number of distractor samples are introduced to the experiments. It also needs to be noted that the ITU-II technique combined two complementary CNN models and hand-crafted features to achieve this performance, COM-Ear, on the other hand, is a coherent model that relies on the same feature representation but considers aggregates global and local information about the appearance of the ears for identitiy inference.

Table 6: Comparison with results from the 2017 Unconstrained Ear Recognition Challenge (UERC) [28]. The results were generated on the complete test dataset containing all 3,540 subjects of the UERC dataset.

Method	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
ICL [28]	0.9	2.8	73.8
IAU [28]	16.2	28.3	90.5
IITK [28]	6.7	11.8	77.5
ITU-I [28]	14.6	28.1	**93.6**
ITU-II [28]	**17.0**	29.4	91.9
LBP-Base [28]	8.7	16.7	84.3
VGG-Base [28]	9.7	19.3	88.3
COM-Ear	16.8	**30.0**	90.2

4.6 Robustness to Occlusions

In our last experiment we evaluate the robustness of our model to occlusions of the ear. We use the same protocol as during the ablation study and run two types of experiments: with and without occlusions. The experiments without occlusions are equivalent to the experiments already presented above. For the experiments with occlusions we simulate the presence of ear accessories and place images of accessories on random places over the cropped ear images. The added accessories are of different shapes and color and simulate a broad spectrum of real-world accessories. Some of the generated images are presented in Fig. 6. As we can see, the accessories mostly cover a smal area of the image, but may be as big as 20% of the image area.

Results for this series of experiments are presented in Table 7 for the COM-Ear model as well as the baseline ResNet-18 model. We see that both models deteriorate in performance, but the degradation is worse for ResNet-18. These results suggest that the local processing path that encodes local image details is indeed beneficial and contributes not only to state-of-the-art performance on unconstrained ear images, but also improves robustness when accessories are present in ear images.

Table 7: Results with accessory-based presentation attack. The upper values show baseline values without the attack, and the values below the delimiting line show the results when attacked with earring images.

Experiment	Method	Rank-1 [%]
Without occlusions	ResNet-18	24.5
	COM-Ear	31.1
With occlusions	ResNet-18	16.1
	COM-Ear	**22.3**

Fig. 6: Example of some of the inputs with added computer generated accessories. The ear accessories were generated in different shapes, positions and with varying sizes.

4.7 Qualitative Evaluation

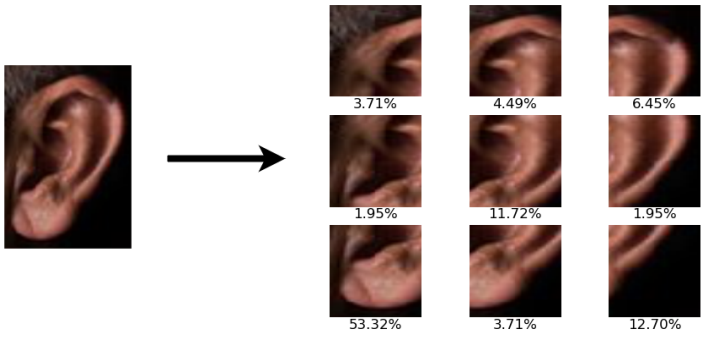
In this section we show some qualitative results related to the COM-Ear modeland also with regard to approaches from UERC 2017. As discussed earlier, COM-Ear aggregates local features with the proposed PRI-pooling operation, which takes a maximum over the patch dimension to produce the 512-dimensional feature vector (in the case of a ResNet-18 backend) from the set of local feature vectors extracted from the image patches. The COM-Ear model allows us to determine, which patch is represented in what proportion in the aggregated (local) feature vector if we examine where each value of the aggregated feature vector came from (i.e. argmax operation).

We show some example ear images from the AWEx dataset and their corresponding image patches in Fig. 7. Here, the fraction of features each patch contributes into the aggregated feature vector is shown below the patches.

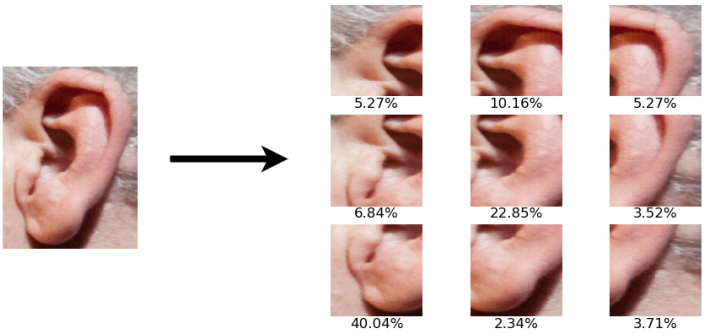
The examples in Fig. 7(a), Fig. 7(b) and Fig. 7(c) represent the same subject with images captured in different conditions and ears in slightly different positions. We can see that similar image features are considered important for all three examples and that the importance of all patches are very similar. Especially important seems to be the bottom-left patch which has a distinct ear shape.

The images in Fig. 7(d) and Fig. 7(e) represent input samples with earrings. We can see that patches with earrings are not weighted heavily as one might expect. This is because the training set contains data with and without earings, so the model can learn that earrings are not necessarily important - this may also be one of the reasons why COM-Ear is much more robust with to the presence of accessories than ReNet-18.

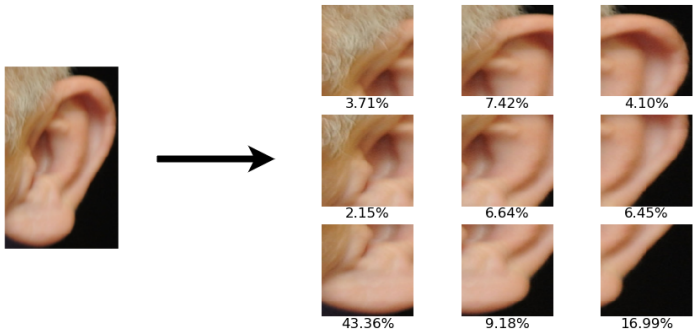
The examples in Fig. 7(f) and Fig. 7(g) show images with large occlusions and presence of earrings which makes it difficult to perform recognition, as many distinct


3.71%
4.49%
6.45%
1.95%
11.72%
1.95%
53.32%
3.71%
12.70%

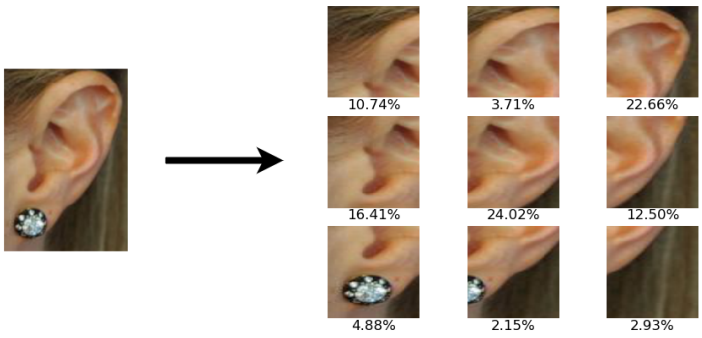
(a)


5.27%
10.16%
5.27%
6.84%
22.85%
3.52%
40.04%
2.34%
3.71%

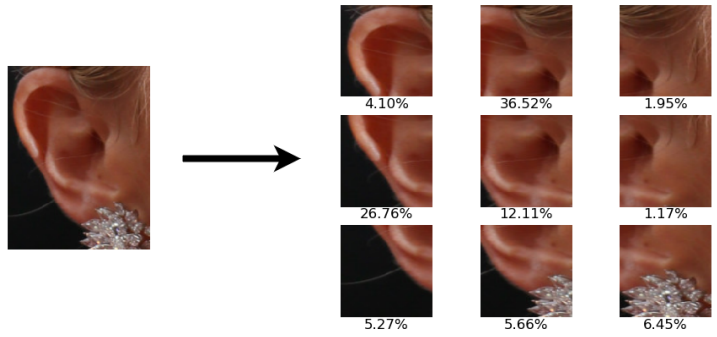
(b)


3.71%
7.42%
4.10%
2.15%
6.64%
6.45%
43.36%
9.18%
16.99%

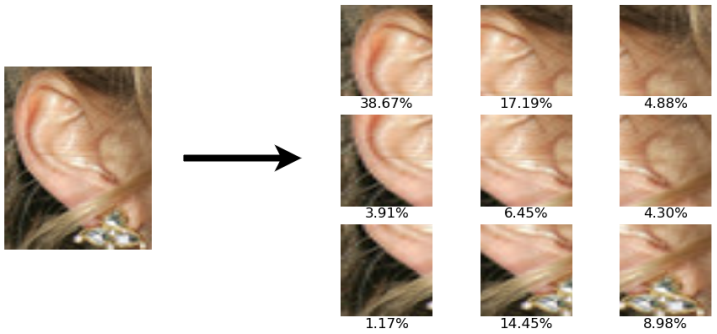
(c)


10.74%
3.71%
22.66%
16.41%
24.02%
12.50%
4.88%
2.15%
2.93%

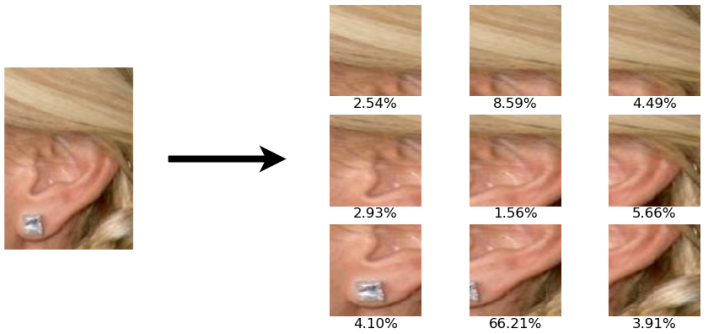
(d)



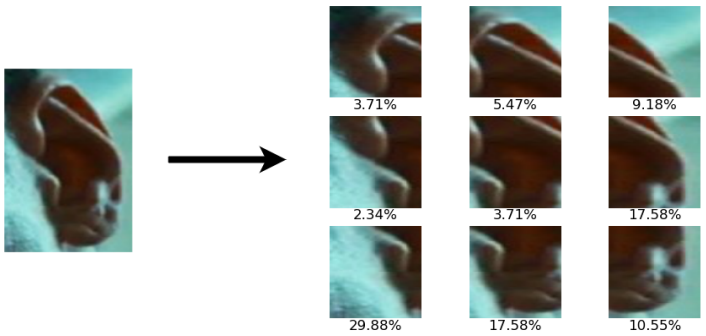
(e)



(f)



(g)



(h)

Fig. 7: Example of input patches and their importance in the aggregated feature vector of the local processing path of the COM-Ear model.

feature are not clearly visible. In the example Fig. 7(f) we see that the top-left patch is the most important as there is almost no occlusion. Similarly for the Fig. 7(g) where the upper sections of the ear are completely occluded, the patch that has no occlusions is chosen to be the most discriminative. Patches with occlusions such as hair are down-weighted as hair is highly variable and the model learns this fact during training.

Fig. 7(h) has no occlusions or accessories but the image is captured in low light conditions and at an difficult angle. The distribution of the importance of the patches is, therefore, much more equal as there are more relevant features present along the whole ear area and one does not dominate.

In order to compare the proposed COM-Ear to others qualitatively, we show what type of images the model and the deep learning approaches from UERC 2017 retrieve from the gallery as the first (rank-1) and as the second match (rank-2) for a given probe images. We also show the first correct prediction (note that there are multiple images of the correct subject in the gallery) and provide the rank, at which it was retrieved. The first correct prediction is considered to be the image that is closest in the ranking and has the same identity as the probe image. For this experiment we again use the entire UERC test data with 9,500 images in the gallery set. The described qualitative analysis is shown in Figure 8 for five randomly selected probe images - shown on the left.

With the top performing approaches, the images retrieved at rank 1 and 2 exhibit a high visual similarity to the probes, as expected. Thus, even when predictions fail, the closest matches visually resemble the probe image. However, with the second, fourth and fifth probe image all 5 evaluated techniques fail. For the fifth probe image the low-resolution of the probe is likely the reason for the failure. The fourth probe image contains high contrast illumination that could be the cause of the error. In the second example (the second probe), the image looks fairly easy to recognize, since illumination is good, the ear is well visible and there are no ear accessories. However, we assume that the cause for the bad performance in this case could be attributed to the fact that there are many images form other subjects in the dataset that look similar. The visual similarity of images found as the closest matches seems to confirm this observation.

5 Conclusion

In this chapter we introduced the first deep constellation model for ear recognition, termed COM-Ear. We evaluated the model in extensive experiments on the Extended Annotated Web Ears (AWEx) dataset and improved upon state-of-the-art results by a large margin. We showed that with the COM-Ear constellation model we not only achieve state-of-the-art results, but also contribute towards more stable recognition performance in challenging setting when parts of the ears are occluded or ear accessories are present in the images. The design of the COM-model and the novel pooling procedure, proposed in this chapter allowed us to visualize certain

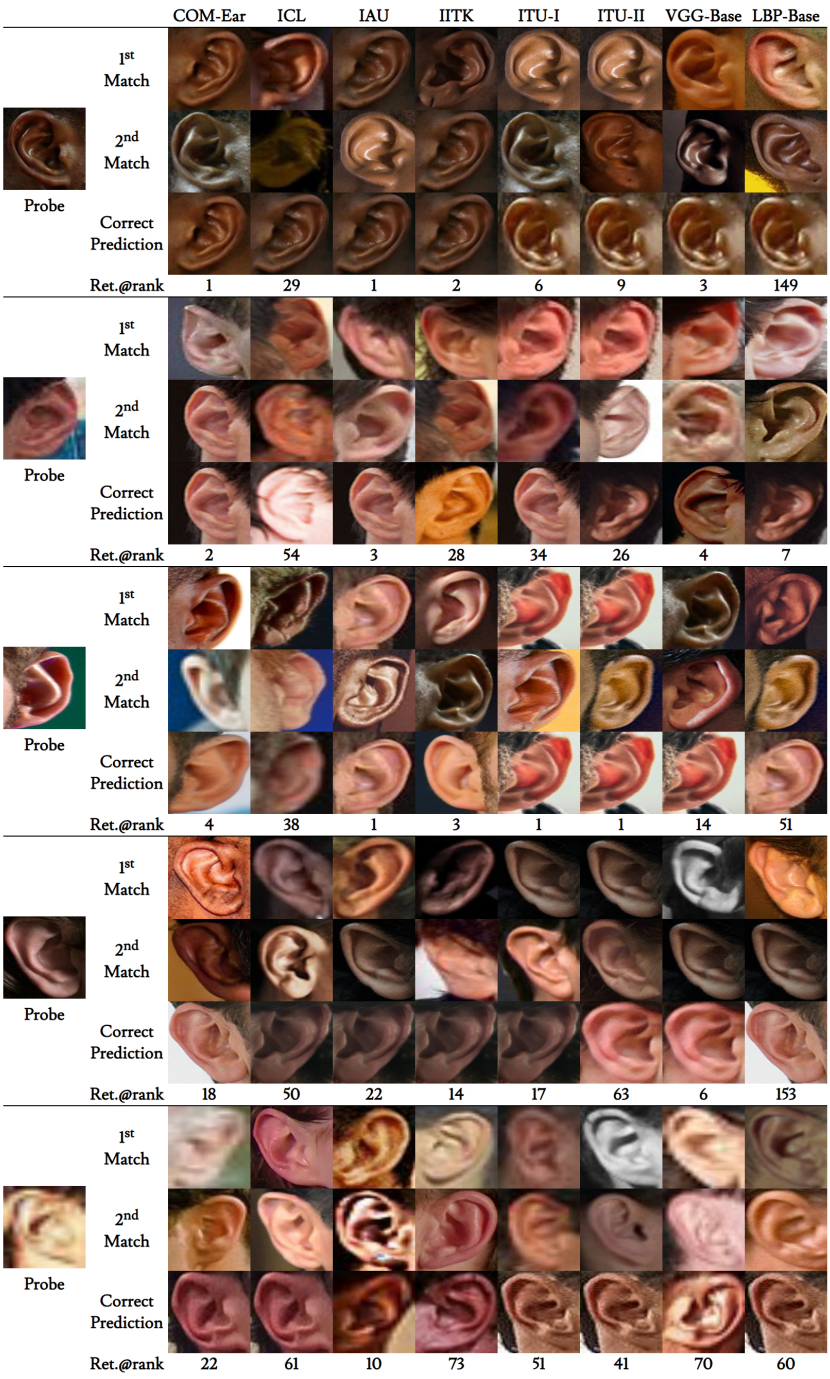


Fig. 8: Qualitative analysis with selected probe images. The figure shows selected probe images (on the left) and the first and second match generated by the evaluated approaches. The first retrieved image with the correct identity is also shown together with the corresponding rank, at which it was retrieved.

aspect of the learned ear representations and resulted in a level of interpretability not seen with competing models. With reversing the the aggregation operation (i.e. the proposed PRI-pooling) we were able to obtain patch level importance which presents an additional novelty of our proposed model. This has important implications for future research a similar concepts could be integrated into other models and offer to better understand the inner workings of deep learning based methods.

6 Acknowledgements

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Program P2-0250 (B) Meteorology and Biometric Systems, the ARRS Research Program P2-0214 (A) Computer Vision. The authors thank NVIDIA for donating the Titan Xp GPU that was used in the experiments.

This work was also partially supported by the European Commission through the Horizon 2020 research and innovation program under grants 688201 (M2DC) and 690907 (IDENTITY).

References

1. Abate, A.F., Nappi, M., Riccio, D., Ricciardi, S.: Ear recognition by means of a rotation invariant descriptor. In: International Conference on Pattern Recognition. vol. 4, pp. 437–440. IEEE, IEEE (2006)
2. Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.: A Survey on Ear Biometrics. ACM Computing Surveys 45(2), 1–22 (2013)
3. Abdel-Mottaleb, M., Zhou, J.: Human ear recognition from face profile images. In: Advances in Biometrics, pp. 786–792. Springer (2006)
4. Akin, C., Kacar, U., Kirci, M.: A Multi-Biometrics for Twins Identification Based Speech and Ear. CoRR abs/1801.09056 (2018)
5. Almisreb, A.A., Jamil, N., Din, N.M.: Utilizing AlexNet Deep Transfer Learning for Ear Recognition. In: International Conference on Information Retrieval and Knowledge Management. pp. 1–5. IEEE (Mar 2018)
6. Alshazly, H.A., Hassaballah, M., Ahmed, M., Ali, A.A.: Ear Biometric Recognition Using Gradient-Based Feature Descriptors. In: Hassanien, A.E., Tolba, M.F., Shaalan, K., Azar, A.T. (eds.) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018. pp. 435–445. Advances in Intelligent Systems and Computing, Springer International Publishing (2019)
7. Arbab-Zavar, B., Nixon, M.S.: Robust log-Gabor filter for ear biometrics. In: International Conference on Pattern Recognition. pp. 1–4. IEEE, IEEE (2008)
8. Baoqing, Z., Zhichun, M., Chen, J., Jiyuan, D.: A robust algorithm for ear recognition under partial occlusion. In: Chinese Control Conference. pp. 3800–3804 (2013)
9. Basit, A., Shoaib, M.: A human ear recognition method using nonlinear curvelet feature subspace. International Journal of Computer Mathematics 91(3), 616–624 (2014)
10. Benzaoui, A., Kheider, A., Boukrouche, A.: Ear description and recognition using ELBP and wavelets. In: International Conference on Applied Research in Computer Science and Engineering. pp. 1–6 (2015)

11. Burge, M., Burger, W.: Ear biometrics. In: Jain, A.K., Bolle, R., Pankanti, S. (eds.) Biometrics, chap. Ear Biometrics, pp. 273–285. IEEE (1996)
12. Bustard, J.D., Nixon, M.S.: Toward unconstrained ear recognition from two-dimensional images. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 40(3), 486–494 (2010)
13. Chan, T.S., Kumar, A.: Reliable ear identification using 2-D quadrature filters. Pattern Recognition Letters 33(14), 1870–1881 (2012)
14. Chang, K., Bowyer, K.W., Sarkar, S., Victor, B.: Comparison and combination of ear and face images in appearance-based biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9), 1160–1165 (2003)
15. Choraś, M.: Perspective methods of human identification: Ear biometrics. Opto-electronics review 16(1), 85–96 (2008)
16. Choras, M., Choras, R.S.: Geometrical algorithms of ear contour shape representation and feature extraction. In: International Conference on Intelligent Systems Design and Applications. pp. 451–456. IEEE, IEEE (2006)
17. Chowdhury, D.P., Bakshi, S., Guo, G., Sa, P.K.: On Applicability of Tunable Filter Bank Based Feature for Ear Biometrics: A Study from Constrained to Unconstrained. Journal of Medical Systems 42(1), 11/1–20 (Nov 2017)
18. Chowdhury, M., Islam, R., Gao, J.: Robust ear biometric recognition using neural network. In: Conference on Industrial Electronics and Applications. pp. 1855–1859. IEEE (Jun 2017)
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition. pp. 886–893. IEEE, IEEE (2005)
20. Damar, N., Fuhrer, B.: Ear Recognition Using Multi-Scale Histogram of Oriented Gradients. In: Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 21–24 (2012)
21. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
22. Dewi, K., Yahagi, T.: Ear photo recognition using scale invariant keypoints. In: Computational Intelligence. pp. 253–258 (2006)
23. Dodge, S., Mounsef, J., Karam, L.: Unconstrained Ear Recognition Using Deep Neural Networks. IET Biometrics 7(3), 207–214 (Jan 2018)
24. Emeršič, Ž., Playà, N.O., Štruc, V., Peer, P.: Towards Accessories-Aware Ear Recognition. In: International Work Conference on Bioinspired Intelligence. pp. 1–8. IEEE (Jul 2018)
25. Emeršič, Ž., Križaj, J., Štruc, V., Peer, P.: Deep Ear Recognition Pipeline. In: Hassaballah, M., Hosny, K.M. (eds.) Recent Advances in Computer Vision, vol. 804, pp. 333–362. Springer (2019)
26. Emeršič, Ž., Meden, B., Peer, P., Štruc, V.: Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements. Neural Computing and Applications pp. 1–16 (2018)
27. Emeršič, Ž., Štepec, D., Štruc, V., Peer, P.: Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild. In: International Conference on Automatic Face and Gesture Recognition – Workshop on Biometrics in the Wild. pp. 987–994. IEEE, IEEE (2017)
28. Emeršič, Ž., Štepec, D., Štruc, V., Peer, P., George, A., Ahmad, A., Omar, E., Boult, T.E., Safdari, R., Zhou, Y., Zafeiriou, S., Yaman, D., Eyiokur, F.I., Ekenel, H.K.: The Unconstrained Ear Recognition Challenge. In: International Joint Conference on Biometrics. pp. 715–724. IEEE/IAPR (2017)
29. Emeršič, Ž., Štruc, V., Peer, P.: Ear Recognition: More Than a Survey. Neurocomputing 255, 26–39 (2017)
30. Eyiokur, F.I., Yaman, D., Ekenel, H.K.: Domain adaptation for ear recognition using deep convolutional neural networks. IET Biometrics 7(3), 199–206 (Dec 2017)

31. Fan, T., Mu, Z., Yang, R.: Multi-modality recognition of human face and ear based on deep learning. In: International Conference on Wavelet Analysis and Pattern Recognition. pp. 38–42 (Jul 2017)
32. Galdámez, P.L., Raveane, W., González Arrieta, A.: A brief review of the ear recognition process using deep neural networks. Journal of Applied Logic 24, 62–70 (Nov 2017)
33. Guo, Y., Xu, Z.: Ear recognition using a new local matching approach. In: International Conference on Image Processing. pp. 289–292. IEEE, IEEE (2008)
34. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics 7(3), 215–223 (Jan 2018)
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 770–778. IEEE (2016)
36. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR abs/1704.04861 (2017)
37. Hurley, D.J., Nixon, M.S., Carter, J.N.: Automatic ear recognition by force field transformations. In: Colloquium on Visual Biometrics. pp. 7/1–5. IET, IET (2000)
38. Jain, A., Ross, A., Nandakumar, K.: Introduction to Biometrics. Springer Science & Business Media (2011)
39. Kacar, U., Kirci, M.: Ear Recognition With Score-Level Fusion Based On CMC In Long-Wave Infrared Spectrum. CoRR abs/1801.09054 (2018)
40. Kannala, J., Rahtu, E.: BSIF: Binarized statistical image features. In: International Conference on Pattern Recognition. pp. 1363–1366. IEEE, IEEE (2012)
41. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Conference on Computer Vision and Pattern Recognition. pp. 4873–4882. IEEE (2016)
42. Križaj, J., Štruc, V., Pavešić, N.: Adaptation of SIFT features for robust face recognition. In: Image Analysis and Recognition. pp. 394–404. Springer, Springer (2010)
43. Laptev, D., Savinov, N., Buhmann, J.M., Pollefeys, M.: TI-POOLING: Transformation-invariant pooling for feature learning in convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition. pp. 289–297. IEEE (2016)
44. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
45. Mahto, S., Arakawa, T., Koshinak, T.: Ear Acoustic Biometrics Using Inaudible Signals and Its Application to Continuous User Authentication. In: European Signal Processing Conference. pp. 1407–1411 (Sep 2018)
46. Mu, Z., Yuan, L., Xu, Z., Xi, D., Qi, S.: Shape and structural feature based ear recognition. In: Advances in Biometric Person Authentication, pp. 663–670. Springer (2004)
47. Naseem, I., Togneri, R., Bennamoun, M.: Sparse representation for ear biometrics. In: Advances in Visual Computing, pp. 336–345. Springer (2008)
48. Nejati, H., Zhang, L., Sim, T., Martinez-Marroquin, E., Dong, G.: Wonder Ears: Identification of Identical Twins from Ear Images. In: International Conference on Pattern Recognition. pp. 1201–1204. IEEE, IEEE (2012)
49. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: International Conference on Pattern Recognition, Computer Vision & Image Processing. vol. 1, pp. 582–585. IEEE/IAPR (1994)
50. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Image and Signal Processing. pp. 236–243. Springer (2008)
51. Ojansivu, V., Rahtu, E., Heikkilä, J.: Rotation invariant local phase quantization for blur insensitive texture analysis. In: International Conference on Pattern Recognition. pp. 1–4. IEEE, IEEE (2008)
52. Omara, I., Wu, X., Zhang, H., Du, Y., Zuo, W.: Learning pairwise SVM on hierarchical deep features for ear recognition. IET Biometrics 7(6), 557–566 (Feb 2018)

53. Omara, I., Xiao, G., Amrani, M., Yan, Z., Zuo, W.: Deep features for efficient multi-biometric recognition with face and ear images. In: International Conference on Digital Image Processing. vol. 10420, pp. 1–6. International Society for Optics and Photonics (Jul 2017)
54. Othman, R.N., Alizadeh, F., Sutherland, A.: A Novel Approach for Occluded Ear Recognition Based on Shape Context. In: International Conference on Advanced Science and Engineering. pp. 93–98. IEEE (Oct 2018)
55. Pflug, A., Busch, C.: Ear biometrics: A survey of detection, feature extraction and recognition methods. IET Biometrics 1(2), 114–129 (2012)
56. Pflug, A., Paul, P.N., Busch, C.: A comparative study on texture and surface descriptors for ear biometrics. In: International Carnahan Conference on Security Technology. pp. 1–6. IEEE, IEEE (2014)
57. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Computational Imaging and Vision, Springer (2011)
58. Prakash, S., Gupta, P.: An efficient ear recognition technique invariant to illumination and pose. Telecommunication Systems 52(3), 1435–1448 (2013)
59. Rahman, M., Islam, M.R., Bhuiyan, N.I., Ahmed, B., Islam, A.: Person identification using ear biometrics. International Journal of The Computer, the Internet and Management 15(2), 1–8 (2007)
60. Ribič, M., Emeršič, Ž., Štruc, V., Peer, P.: Influence of Alignment on Ear Recognition: Case Study on AWE Dataset. In: International Electrotechnical and Computer Science Conference. vol. 25-B, pp. 131–134. IEEE (2016)
61. Saeed, U., Khan, M.M.: Combining ear-based traditional and soft biometrics for unconstrained ear recognition. Journal of Electronic Imaging 27(5), 051220/1–10 (May 2018)
62. Sepas-Moghaddam, A., Pereira, F., Correia, P.L.: Ear recognition in a light field imaging framework: A new perspective. IET Biometrics 7(3), 224–231 (Jan 2018)
63. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556 (2014)
64. Sinha, H., Manekar, R., Sinha, Y., Ajmera, P.K.: Convolutional Neural Network-Based Human Identification Using Outer Ear Images. In: Bansal, J.C., Das, K.N., Nagar, A., Deep, K., Ojha, A.K. (eds.) Soft Computing for Problem Solving. pp. 707–719. Advances in Intelligent Systems and Computing, Springer Singapore (2019)
65. Theoharis, T., Passalis, G., Toderici, G., Kakadiaris, I.A.: Unified 3D face and ear recognition using wavelets on geometry images. Pattern Recognition 41(3), 796–804 (2008)
66. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. pp. 437–441. IEEE (Oct 2016)
67. Victor, B., Bowyer, K., Sarkar, S.: An evaluation of face and ear biometrics. In: International Conference on Pattern Recognition. vol. 1, pp. 429–432. IEEE, IEEE (2002)
68. Vidyasri, R., Priyalakshmi, B., Raja, M.R., Priyanka, S.: Recognition of ear based on partial features fusion. International Journal of Biometrics 10(2), 105–120 (Jan 2018)
69. Vu, N.S., Caplier, A.: Face recognition with patterns of oriented edge magnitudes. In: European Conference on Computer Vision. pp. 313–326. Springer, Springer (2010)
70. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision. pp. 499–515. Springer, Springer (2016)
71. Xie, Z., Mu, Z.: Ear recognition using LLE and IDLLE algorithm. In: International Conference on Pattern Recognition. pp. 1–4. IEEE, IEEE (2008)
72. Xu, X., Mu, Z., Yuan, L.: Feature-Level fusion method based on KFDA for multimodal recognition fusing ear and profile face. In: International Conference on Wavelet Analysis and Pattern Recognition. pp. 1306–1310. IEEE, IEEE (2007)
73. Ying, T., Shining, W., Wanxiang, L.: Human ear recognition based on deep convolutional neural network. In: Chinese Control And Decision Conference. pp. 1830–1835 (Jun 2018)
74. Yuan, L., Mu, Z.C.: Ear recognition based on 2D images. In: Conference on Biometrics: Theory, Applications and Systems. pp. 1–5. IEEE, IEEE (2007)

75. Yuan, L., Zhao, H., Zhang, Y., Wu, Z.: Ear Alignment Based on Convolutional Neural Network. In: Zhou, J., Wang, Y., Sun, Z., Jia, Z., Feng, J., Shan, S., Ubul, K., Guo, Z. (eds.) Biometric Recognition. pp. 562–571. Lecture Notes in Computer Science, Springer International Publishing (2018)
76. Zhang, H.J., Mu, Z.C., Qu, W., Liu, L.M., Zhang, C.Y.: A novel approach for ear recognition based on ICA and RBF network. In: International Conference on Machine Learning and Cybernetics. vol. 7, pp. 4511–4515. IEEE, IEEE (2005)
77. Zhang, Y., Mu, Z., Yuan, L., Yu, C.: Ear verification under uncontrolled conditions with convolutional neural networks. IET Biometrics 7(3), 185–198 (Jan 2018)
78. Zhang, Y., Mu, Z., Yuan, L., Zeng, H., Chen, L.: 3D Ear Normalization and Recognition Based on Local Surface Variation. Applied Sciences 7(1), 104/1–21 (Jan 2017)
79. Zhou, Y., Zaferiou, S.: Deformable Models of Ears in-the-Wild for Alignment and Recognition. In: International Conference on Automatic Face and Gesture Recognition. pp. 626–633. IEEE (May 2017)