Deep Ear Recognition Pipeline

Žiga Emeršič, Janez Križaj, Vitomir Štruc, and Peter Peer

Abstract Ear recognition has seen multiple improvements in recent years and still remains very active today. However, it has been approached from recognition and detection perspective separately. Furthermore, deep-learning-based approaches that are popular in other domains have seen limited use in ear recognition and even more so in ear detection. Moreover, to obtain a usable recognition system a unified pipeline is needed. The input in such system should be plain images of subjects and the output identities based only on ear biometrics. We conduct separate analysis through detection and identification experiments on the challenging dataset and, using the best approaches, present a novel, unified pipeline. The pipeline is based on convolutional neural networks (CNN) and presents, to the best of our knowledge, the first CNN-based ear recognition pipeline. The pipeline incorporates both, the detection of ears on arbitrary images of people, as well as recognition on these segmented ear regions. The experiments show that the presented system is a state-of-the-art system and, thus, a good foundation for future real-word ear recognition systems.

Key words: Ear biometrics, Ear recognition pipeline, Ear detection, Ear segmentation, Convolutional neural networks.

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

Janez Križaj · 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: {janez.krizaj, vitomir.struc@fe.uni-lj.si}

1 Introduction

General recognition pipelines based on specific biometric modalities consist of detecting and segmenting appropriate parts and then performing analysis of the detected regions to distinguish subjects and identify them. The detection or segmentation of images is therefore a necessary step towards biometric-based person recognition. Furthermore, obtaining good detection results and segmenting regions of interest directly impacts the recognition system's performance. Despite ear biometrics domain's large improvements in the recent year and increased popularity there are, to the best of our knowledge, still no deep-learning-based ear recognition pipelines. However, treated separately, there have been contributions in ear detection and ear recognition, such as [10, 20, 25, 26, 28, 29, 41]. This is expected, since ear biometrics offers numerous application possibilities in forensics, security and surveillance [2, 36]. As far as ear recognition by itself goes, the proposed approaches in literature range from geometric and holistic techniques [3, 11] to more recent descriptor- [14, 15, 51, 54, 61] and deep-learning-based [25, 26, 28, 35, 85] methods. While, descriptor-based methods have dominated the field over the last years, research is moving away from these methods and is now focusing increasingly on deep-learning-based models, which recently brought about considerable advancements in various areas of computer vision and beyond.

However, overall the field of ear biometrics still lags behind the research of other biometric modalities, such as faces or fingerprints. Recent surveys on ear recognition attribute this fact to the lack of efficient detection techniques, which are capable of determining the location of the ear(s) in the input images and represent a key component of automatic ear recognition systems [2, 36, 62]. In fact, the authors of a recent survey [36] argue that the absence of automatic ear detection approaches is one the most important factors hindering a wider deployment of ear recognition technology.

Despite the progress in the area of ear detection over the recent years, most of the existing work is limited to laboratory-like settings and controlled image acquisition conditions, where the appearance variability of ear images is limited and not representative of real-world imaging conditions [62], with some exceptions [29]. In unconstrained settings, on the other hand, ear detection is less well explored and remains challenging due to appearance changes caused by shape, size, and color variations, occlusions by hair strains or accessories and imaging conditions, which often vary due to different illumination and viewing angles. The problem of ear detection has only recently been considered [29]. The main shortcoming of [29] is bad detection performance under bad conditions, the approach in some cases fails completely. Furthermore, even when ears do get detected, the detected regions are not always accurate.

In this chapter, the performance of ear detection is improved upon, by using RefineNet [52] as opposed to the previous state-of-the-art ear detection that was achieved using PED-CED architecture [29]. The detection part is joined up with the ResNet as the recognition part, resulting, to the best of our knowledge, in the first ever CNN-based ear recognition pipeline. This is important, since it enables

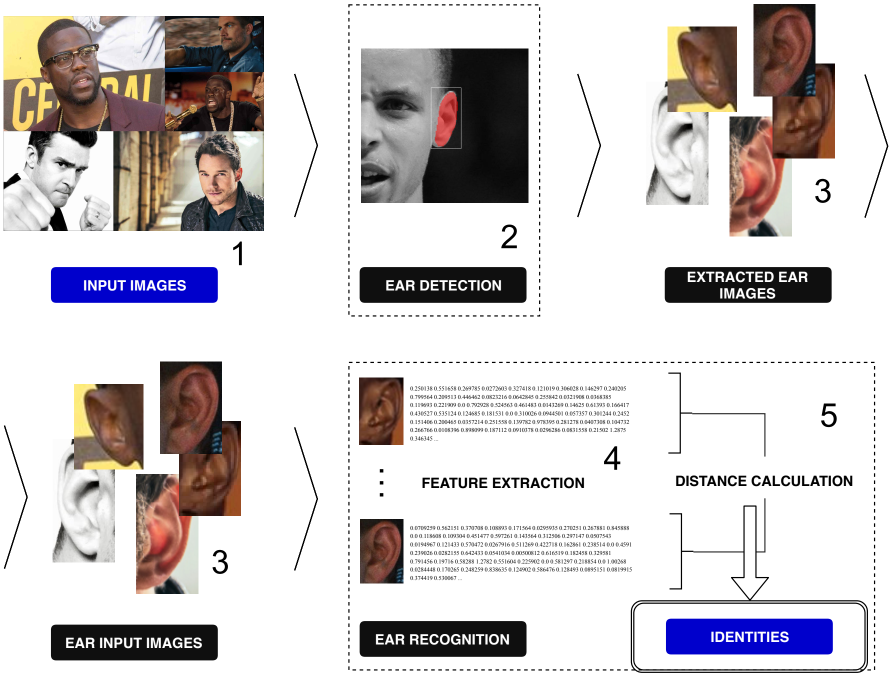


Fig. 1: Diagram of the proposed unified ear recognition pipeline. The inputs are arbitrary images of subjects (1), the output of the pipeline are identities based only on ear biometrics (5). Ear detection (2) is performed on images of subjects, which outputs cropped ear images (3) that serve as the input into ear recognition step (4 and 5).

ear recognition on plain, untreated images of subjects, without any preprocessing. A diagram of the proposed pipeline that predicts identities solely on ear biometrics, is shown in Fig. 1.

To summarize, the following contributions are presented in this chapter:

- a novel ear recognition pipeline based on a convolutional neural network that performs detection of ears, as well as recognition and works well on image data captured in completely unconstrained settings,
- an improvement upon the previous best ear detection approach,
- a detailed analysis of the proposed techniques for detection and recognition separately, as well as joint analysis.

The rest of the chapter is structured as follows. In Section 2, we overview the related work from ear detection and ear recognition perspective. In Section 3, the proposed deep pipeline is described. In Section 4, experiments and results are presented. In Section 5, conclusions and future work is described.

2 Prior Work

In this section, state-of-the-art approaches to ear detection and ear recognition are presented. A brief description of the approaches is provided to establish a foundation for the work presented in the continuation of this chapter. Although detection and recognition in this part are treated separately, the joint pipeline performs both.

2.1 Ear Detection

In this section, the most important techniques for ear detection are surveyed with the goal of providing the reader with the necessary context for our work. A more comprehensive review on existing ear detection approaches (from 2D as well as 3D imagery) can be found in recent surveys on this topic [62, 69] and our previous work [29], where also a CNN-based PED-CED approach for pixel-wise ear detection was presented. The main discrepancy of earlier works is inability to detect ears in a pixel-wise fashion under variable conditions. The approaches that manage that, such as [29], still lack high accuracy rates. Furthermore, approaches for ear detection mainly stand on their own with no direct applications to ear recognition.

Comparing existing approaches among each other is often difficult, since no standard benchmarks and evaluation methodology exists for ear detection. Authors typically report different performance metrics and rely on self compiled evaluation protocols in their experiments – even when same datasets are used. Furthermore, since face detection is commonly assumed to have been used on the images before ear detection is performed, the term ear detection is typically used interchangeably with ear localization or even ear enrollment.

One of the earliest groups of approaches towards ear detection consists of fitting ellipses to the possible ear candidates using the Hough Transform [5]. In [4, 7], the Canny edge detector is used to extract edges from ear images and the ears outer helix curves are used as features for the localization process. In the work of [46], a cascaded-AdaBoost-based ear detection approach is proposed. Another approach to ear detection based on the distance transform and template matching is proposed in [70]. In [71], the connected component analysis of a graph constructed using the edge map of the image and then the regions are bounded by rectangles. In [72], the same authors, Prakash et al., approach the ear detection problem by segmenting skin-colored regions.

Haar features arranged in a cascaded Adaboost classifier, better known as Viola-Jones [82], are used in [1] for ear detection. The authors manually annotate the UND-F [57], UMIST [80], WV HTF [1] and USTB [27] datasets with rectangles around ears and use the annotated data for training and testing. This approach is capable of handling a wide variety of image variability and operating in real-time. The work is interesting since Viola-Jones detection was very popular prior 2012 and the rise of deep-learning approaches. In [21], an approach based on image ray transform is used, which highlights the tubular structures of the ear as an enrollment technique. The approach presented in [67] makes use of the edge map of the side face images. An edge connectivity graph build on top of the edge map serves as the basis for ear candidate calculation.

A case of geometrical approach to ear detection was presented in [84] with the approach named HEARD. This ear detection method is based on three main shape features of the human ear: the height-to-width ratio of the ear, the area-to-perimeter ratio of the ear, and the fact that the ear's outline is the most rounded outline on the side of a human face. To avoid occlusions caused by hair and earrings, the method looks for the inner part of the ear instead of the outer part. The ear detection algorithm proposed in [64] uses texture and depth information to localize ears in profile-face images and images taken at different angles. Details on the ear surface and edge information are used for finding the ear outline in an image. The algorithm utilizes the fact that the surface of the outer ear has a delicate structure with high local curvature. The ear detection procedure returns an enclosing rectangle of the best ear candidate.

In [37], Ganesh et al. present a method called Entropic Binary Particle Swarm Optimization (EBPSO), which generates an entropy map, which together with background subtraction is exploited to detect ears in the given face image. Prajwal et al. [19] propose an ear detection approach that relies on the entropy-Hough transform. A combination of a hybrid ear localizer and an ellipsoid ear classifier is used to predict locations of ears. Sarangi et al. [75] present a new scheme for automatic ear localization relying on template matching with the modified Hausdorff distance. The benefit of this technique is that it does not depend on pixel intensities and that the template incorporates various ear shapes. Thus, this approach is reported to be invariant to illumination, pose, shape and occlusion of the ear images.

In the majority of cases, authors evaluate their approaches on the USTB [27], UND dataset [57], Carreira-Perpinan dataset [17], CMU PIE [76], Pointing Head Pose [38], FERET [65], UMIST [80], XM2VTS [55] dataset and on the IITK dataset [69]. Arguably, in many cases these dataset are not challenging enough and not applicable to the real-life scenarios. Similarly to other fields in computer vision, and nonetheless ear recognition, deep-learning based approaches are starting to emerge and present new state of the art. In [29] for example a novel, modified SegNet architecture for a pixel-wise ear detection is applied to ear detection.

2.2 Ear Recognition

Ear recognition has seen in recent year even more imperative contributions as ear detection. Only a couple of years ago descriptor-based recognition techniques were the state-of-the-art in this field [2, 36, 62], recently deep-learning-based approaches prevail [25, 35, 41]. This is on pair with other biometric modalities, where the biometric domain shifted towards deep learning. Nevertheless, these two groups of techniques approach the ear recognition in fundamentally different ways.

Descriptor-based techniques, for example, extract information from local image areas and use the extracted information for identity inference. As emphasized in the recent survey [36], two groups of techniques can in general be considered descriptor-based: *i)* techniques that first detect interest points in the image and then compute descriptors for the detected interest points, and *ii)* techniques that compute descriptors densely over the entire images based on a sliding window approach (with or without overlap). Examples of techniques from the first group include [6, 16] or more recently [68]. A common characteristic of these techniques is the description of the interest points independently one from the other, 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 [12,18,50,83]. These techniques also capture the global properties of the ear in addition to the local characteristics, which commonly result in a higher recognition performance, but the dense descriptor-computation procedure comes at the expense of the robustness to partial occlusions. Nonetheless, recent trends in ear recognition favor dense descriptor-based techniques primarily due to their computational simplicity and high recognition performance.

Deep-learning-based methods, on the other hand, typically process the input images in a holistic manner and learn image representations (features, descriptors) directly from the training data by minimizing some suitable loss at the output of the recognition model. The most popular deep-learning models, CNNs, commonly process the data through a hierarchy of convolutional and pooling layers that can be seen as stacked feature extractors and once fully trained can be used to derive highly discriminative data representations from the input images that can be exploited for identity inference. While these representations commonly ensure formidable recognition performance, the CNN-training procedure typically requires a large amount of training data, which may not always be available and is not needed with descriptor based methods. In the field of ear recognition, deep-learning based methods are relatively new [25,26,28,35,60,85], but are already outperforming local descriptor based methods [30,34,35,79].

3 Deep Ear Recognition

In this section, the ear recognition pipeline based completely on deep learning models is presented. First, the overall structure of the pipeline is described, followed by the details of the detection and recognition parts. Second, the characteristics of the pipeline in comparison to existing approaches is discussed.

3.1 Proposed pipeline overview

A block diagram of the proposed pipeline is shown in Fig. 2 and a more detailed diagram with each step described is shown in Fig. 3. Arbitrary images of subjects

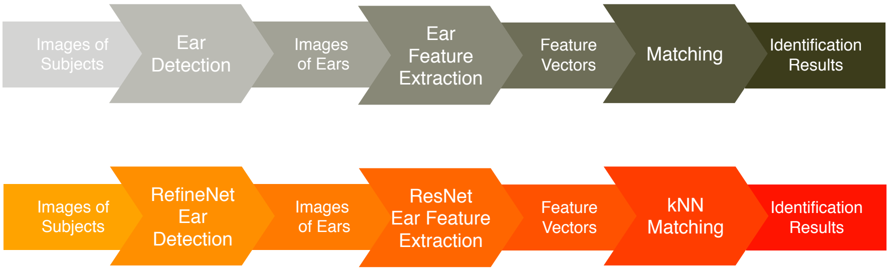


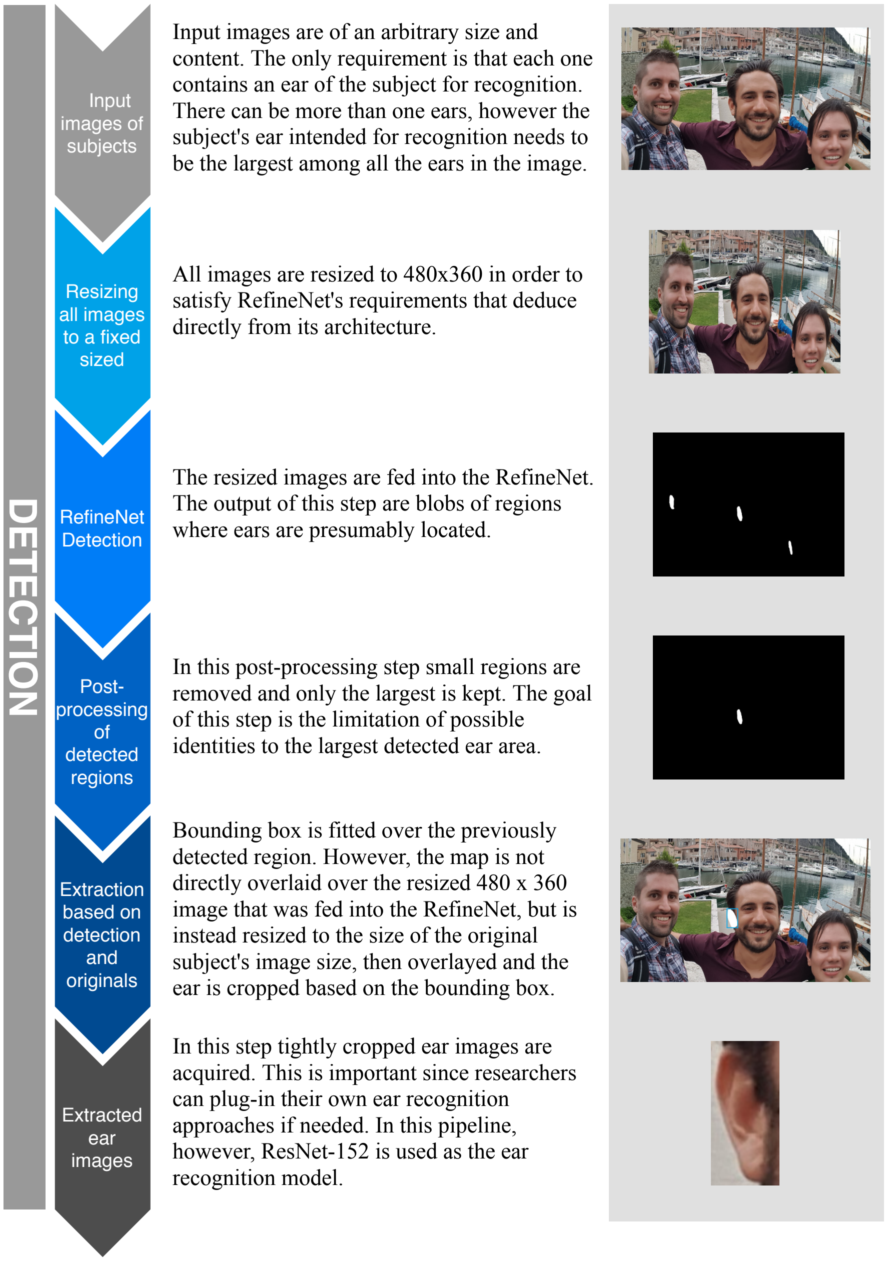
Fig. 2: A block diagram of a general ear recognition pipeline at the top, and the pipeline presented in this chapter at the bottom.

serve as the input into the pipeline. RefineNet-152 [52] is used in the first part of the pipeline for extracting ear images. The model produces maps of detected ear regions on the supplied input images of subjects. Images are resized to a fixed dimensions 480×360 prior inputting them into the network. In the detection post-processing step, all but the largest detected regions are removed. The single remaining region serves as the basis for cropping the ear out of the originally sized images. The reason only one ear is kept (and preferably the pipeline is supplied with only one-eared-images) is that there are no issues regarding the identity when preparing the evaluation. However, researchers are free to lift this limitation, since the pipeline is capable of detecting multiple ears, as illustrated in Fig. 3.

These extracted ear images are then fed into the recognition network – ResNet-152 without the last fully-connected layer. The output are feature vectors and are used the same way as any traditionally prepared feature vectors. This way the pipeline is able to predict identities on data and identities the trained models have never seen before. This is the so-called open-set prediction. The produced feature vectors are compared using χ^2 distance measure. However, this distance comparison could be replaced by some other approaches such as [60].

After the acquisition of these distances identification experiments are performed. Results are reported through rank scores and plot cumulative match curves (CMC). These measures are described in more detail in Section 4.2. The identification mode means that for each sample a prediction is made to which class the sample belongs. This is opposed to verification experiments, where for each sample we only predict whether it belongs to the observed class or not, and typically report equal error rates, verification rates etc. and also typically visualize results using Receiver operating characteristic (ROC) curves [36].

RefineNet and ResNet-152 were selected for the CNN-based ear recognition pipeline based on their superior performance reported in literature [40,43,44,53,81]. Furthermore, ResNet despite its superior performance compared to e.g. VGG, even in its deepest implementation, contains fewer parameters needed to set during training [42]. Both architectures are described Section 3.2 and Section 3.3, respectively. However, the reader is refered to [30] to see performance evaluation also for some other CNN architectures, such as SqueezeNet and VGG.



DETECTION

Input images of subjects

Input images are of an arbitrary size and content. The only requirement is that each one contains an ear of the subject for recognition. There can be more than one ears, however the subject's ear intended for recognition needs to be the largest among all the ears in the image.

Resizing all images to a fixed sized

All images are resized to 480x360 in order to satisfy RefineNet's requirements that deduce directly from its architecture.

RefineNet Detection

The resized images are fed into the RefineNet. The output of this step are blobs of regions where ears are presumably located.

Post-processing of detected regions

In this post-processing step small regions are removed and only the largest is kept. The goal of this step is the limitation of possible identities to the largest detected ear area.

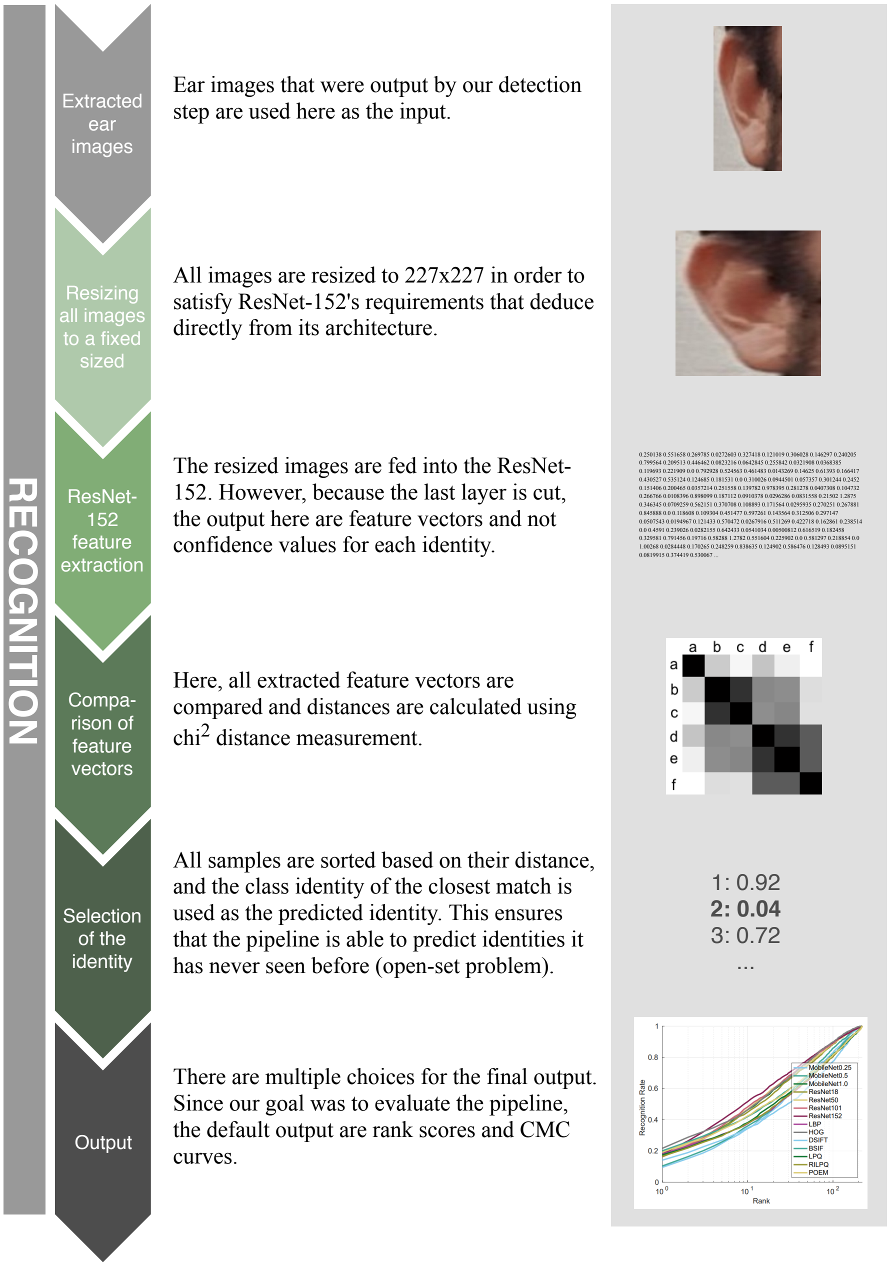
Extraction based on detection and originals

Bounding box is fitted over the previously detected region. However, the map is not directly overlaid over the resized 480 x 360 image that was fed into the RefineNet, but is instead resized to the size of the original subject's image size, then overlayed and the ear is cropped based on the bounding box.

Extracted ear images

In this step tightly cropped ear images are acquired. This is important since researchers can plug-in their own ear recognition approaches if needed. In this pipeline, however, ResNet-152 is used as the ear recognition model.

[The diagram continues on the next page.]


RECOGNITION
Extracted ear images

Ear images that were output by our detection step are used here as the input.

Resizing all images to a fixed sized

All images are resized to 227x227 in order to satisfy ResNet-152's requirements that deduce directly from its architecture.

ResNet-152 feature extraction

The resized images are fed into the ResNet-152. However, because the last layer is cut, the output here are feature vectors and not confidence values for each identity.

```
0.250138 0.551658 0.269785 0.0272603 0.327418 0.121019 0.306028 0.146297 0.240205
0.799564 0.209513 0.446462 0.0823216 0.0642845 0.255842 0.0321908 0.0368385
0.119693 0.221909 0.0 0.792928 0.524563 0.461483 0.0143269 0.14625 0.61393 0.166417
0.430527 0.535124 0.124685 0.181531 0.0 0.310026 0.0944501 0.057357 0.301244 0.2452
0.151406 0.200465 0.0357214 0.251558 0.139782 0.978395 0.281278 0.0407308 0.104732
0.266766 0.0108396 0.898099 0.187112 0.0910378 0.0296286 0.0831558 0.21502 1.2875
0.346345 0.0709259 0.562151 0.370708 0.108893 0.171564 0.0295935 0.270251 0.267881
0.845888 0.0 0.118608 0.109304 0.451477 0.597261 0.143564 0.312506 0.297147
0.0507543 0.0194967 0.121433 0.570472 0.0267916 0.511269 0.422718 0.162861 0.238514
0.0 0.4591 0.239026 0.0282155 0.642433 0.0541034 0.00500812 0.616519 0.182458
0.329581 0.791456 0.19716 0.58288 1.2782 0.551604 0.225902 0.0 0.581297 0.218854 0.0
1.00268 0.0284448 0.170265 0.248259 0.838635 0.124902 0.586476 0.128493 0.0895151
0.0819915 0.374419 0.530067 ...
```

Here, all extracted feature vectors are compared and distances are calculated using chi^2 distance measurement.

All samples are sorted based on their distance, and the class identity of the closest match is used as the predicted identity. This ensures that the pipeline is able to predict identities it has never seen before (open-set problem).

There are multiple choices for the final output. Since our goal was to evaluate the pipeline, the default output are rank scores and CMC curves.

Fig. 3: A diagram of the unified pipeline. The detection part of the pipeline (in blue) is assembled mainly from RefineNet detector and the recognition part of the pipeline (in green) is assembled mainly from ResNet-152. The inputs are arbitrary images of subjects, the outputs are the subjects' identities.

3.2 Ear detection with RefineNet

The ear detection part: The goal of this first part of our pipeline is to extract ear images. These images contain tightly cropped areas of ears. With this criteria satisfied there is as small amount of non-ear biometric data as possible. The steps of detection are described in the first part of the Fig. 3. The experiments were set up with the requirement that only one ear per image is recognized. Although the pipeline is capable of dealing with multiple ears, this limitation was set, as already emphasized, in order to guarantee the correct experimental evaluation. However, the pipeline is set in such a way, that this limitation can be lifted, and all the detections from RefineNet can be freely used.

RefineNet: a generic multi-path refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections [52]. This is enabled by fusing high-level features with low-level features. Combining such coarse and fine features results in a high-resolution maps of features, where both large-scale locations and fine details are captured well.

In its core, RefineNet is exploits ResNet as building blocks. In its multipath architecture ResNet is split into four blocks and directly onto each output a RefineNet unit is plugged on. One such block is illustrated in Fig. 4 and it consists of (from left to right respectively):

- Residual Convolution Units (RCU), which are simplified versions of the original ResNet's convolution unit.
- Multi-Resolution Fusion, which fuses multiple inputs into a high-resolution map.
- Chained Residual Pooling, which uses a high-resolution map to capture background context.
- Output convolutions to introduce non-linearity to the fusions of feature maps.

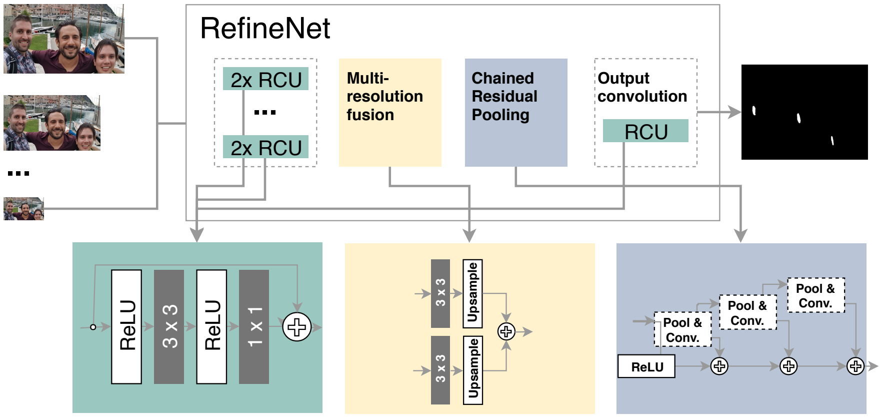


Fig. 4: Illustration of the RefineNet architecture [52].

3.3 Ear recognition with ResNet

The recognition part: In the recognition part, the extracted ear images are used to deduce information about the identity of the person. However, during training the whole architecture is used nevertheless, to produce confidence values for each identity. During testing and final prediction the last fully-connected layer is removed in order to produce the aforementioned feature vectors. Images of ears from the previous step, the RefineNet detection part, serve here as an input for feature vector calculation. Each step of the recognition part of the pipeline is described in the second part of the Fig. 3.

ResNet: a member of the so-called Deep Residual Networks [42], meaning it consists of many stacked Residual Units. Instead of learning unreferenced functions, here the layers are reformulated as learning residual functions with reference to the layer inputs. Authors show that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth as compared to VGG or AlexNet. The most obvious difference between these residual networks compared to e.g. VGG is in its shortcut connections, typical for residual learning building blocks, as illustrated in Fig. 5.

The identity shortcuts can be used directly when the input and the output are of the same dimensions. Otherwise there are two possibilities: zero padding the data or performing a 1×1 convolution. In the latter case, however, the number of trainable parameters is increased and thus the footprint of a model. The authors report that using identity shortcuts without introduction additional parameters improves the training performance.

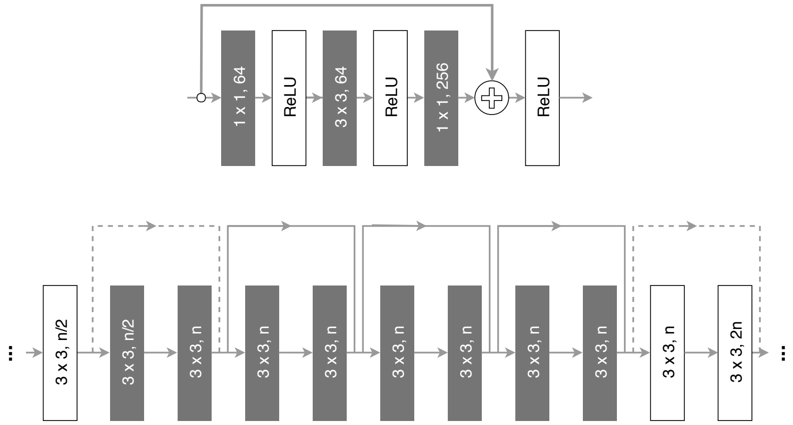


Fig. 5: Illustration of the ResNet's bottleneck building block at the top, a section of residual network at the bottom, where each box presents a convolution and n-values go from 64 through 128 etc. Note that number of convolutions within one section is not always 8. For the full architecture the reader is referred to to [42].

3.4 Characteristics

The pipeline consists solely of CNNs. Everything is set by training from the data. Furthermore, both RefineNet and ResNet-152 present state-of-the-art from the field of detection and recognition, respectively. The presented ear recognition pipeline is one of the few available pipelines for ear recognition, and to the best of our knowledge, the only CNN-based one. The pipeline is capable of handling new identities without retraining.

Covariate analysis of comparing performance of images of specific characteristics is not reported in this chapter (such as [32]), however, visual inspection can reveal that the pipeline is robust to bad illumination conditions, high degrees of which the images were taken and, at least the detection part, high levels of occlusions. The reason lies in the way RefineNet and ResNet acknowledge data. In the traditional approaches, knowledge of how to describe and how to respond to data specific for the domain is embedded by the expert's knowledge that is developing the system. This is in direct opposition to approaches based on convolutional neural networks, where no such prerequisites are set. Instead, the models themselves deduce what is important according to the data supplied during the training stages. And since the training data is comparable with the test data (as far as difficulty goes), it is expected that the models pickup and deduce what is important and what it is not. Without dwelling excessively into the experiments, a speculation can be made that the higher performance scores of both recognition and segmentation could be attributed to these described robustness factors.

One of the major characteristic of these CNN-based approaches compared to the more traditional dense-descriptor-based is the fact that due to their self adaptation during training stages, they learn to use parts that are important for recognition and ignore the rest. This is important, since, under correct training this makes them robust to occlusions or otherwise missing data. Traditional dense approaches on the other hand, while providing relative high success rates and computational simplicity, suffer from these characteristics.

Another important characteristic connected with prediction models is the footprint of the models. While ResNet-152 was selected for the pipeline due to its superior performance, its footprint with model size of over 234.1 MB is larger compared to the full MobileNet model size of 13.4 MB. Nonetheless, this is still significantly smaller than e.g. VGG model with size of over 500 MB (depending on the variation).

4 Experiments and Results

In this section, the performance of the proposed ear recognition pipeline is evaluated. Experiments highlight the main characteristics of both the ear detection and recognition parts. The section begins with a description of the experimental data and performance measures used to report results. The results are presented separately for detection and recognition, as well as for the entire ear recognition pipeline. All code, data and models are made available publicly to the research community on the website (http://awe.fri.uni-lj.si) to ensure transparency and reproducibility of the experiments. In order to display multiple possibilities to the readers, other approaches are also evaluated and also presented separately.

Detection: In the detection part, three architectures are evaluated: SegNet [9], PED-CED [29] and RefineNet [52]. The latter is used in the pipeline and is described in Section 3.2. Here, however, only a brief description of SegNet and PED-CED is provided. Both SegNet and PED-CED (encoder-decoder architecture) approaches contain the convolutional encoder-decoder segmentation network as the main component for detection. This segmentation network is built around the pre-trained VGG-16 model [77] similarly to [8, 9]. The pre-trained VGG-16 model represents a powerful deep model trained on over 1.2 million images of the ImageNet dataset [74] (there are over 14 million images in the whole ImageNet dataset) for the task of object recognition and is publicly available. It is comprised of 13 convolutional layers interspersed with max pooling layers and is a common choice for the encoding part of such models. The decoding part of SegNet and PED-CED has a similar (but inverted) architecture to VGG-16, but instead of max pooling layers contains unpooling layers that upsample the feature maps generated by the encoders to a larger size.The main difference of PED-CED compared to SegNet and also its biggest advantage are the shortcuts connections. Shortcuts are made between the convolutional layers of the encoder and decoder. Specifically, the feature maps are forwarded from the encoders composed of blocks of three convolutional layers and concatenate the forwarded feature maps with the feature maps produced by the convolutional layers of the corresponding decoder. These shortcut connections are introduced only between a single convolutional layers of a given encoder block and the corresponding decoder layer to reduce redundancy as well as the computational burden.

Recognition: For the recognition part of the proposed pipeline, seven descriptor-based techniques and 7 CNN-based approaches are evaluated. The latter seven, are based on two architectures: MobileNet [45] and ResNet [42]. ResNet was selected due to its superior performance in literature and due to its use in [32], where it was shown that it outperforms other evaluated architectures. MobileNet was selected as a representative of the so-called lightweight architectures [45]. The premise here is that ear recognition pipeline should work in real life, where speed and small footprint of the model is important. ResNet was selected for the pipeline and is described in Section 3.3, here MobileNet is briefly covered. The MobileNet architecture was developed with mobile and embedded vision applications deployment in mind. This is also the main reason why it was selected for the evaluation in this work. The architecture uses two main hyper-parameters that efficiently trade off between latency and accuracy [45]. These hyper-parameters allow to tweak the size of the model with accordance with the problem domain and use-case scenarios. In this work an evaluation of three such versions with different width multipliers is provided. Lower the value, less parameters there are to train, the more lightweight is the model. Higher the value (highest being 1), more parameters there are to train, heavier the footprint (space- and time-wise) of the model. Although, the main goal of our ear recognition pipeline is the accuracy, the ability to plug-in such a lightweight model may be useful for some readers. During the experiments three levels of multipliers were used: $\frac{1}{4}$, $\frac{1}{2}$ and 1. For the descriptor-based methods, a dense-descriptor computation is considered, generating d-dimensional feature vectors needed for recognition. Specifically, the methods based on the following approaches are used for the analysis: Local Binary Patterns (LBPs) [13,36,39,63,66], (Rotation Invariant) Local Phase Quantization Features (RILPQ and LPQ) [58,59], Binarized Statistical Image Features (BSIF) [36,49,63], Histograms of Oriented Gradients (HOG) [22,23,36,63], Dense Scale Invariant Feature Transform (DSIFT) [24,36,50], and Patterns of Oriented Edge Magnitudes (POEM) [36,83].

4.1 Dataset and experimental protocol

Ear images from several datasets are used for the experiments. Specifically, the experiments are performed on images from the latest version of the Annotated Web Ears (AWE) [36] and the Unconstrained Ear Recognition Challenge (UERC) [35] as two main sources of data. The final experimental dataset contains 4,004 images of 336 distinct subjects (with a variable number of images per subject). Because all images were gathered from the web, they exhibit a large amount of appearance variability across ear rotations (also in-plane), illumination, age, gender, race, occlusion, and other factors. The outlined characteristics make this dataset one of the most challenging ear datasets publicly available to the research community. Because the first (detection) part of the pipeline is based on a segmentation network, manual annotations (at the pixel level) of 1000 images of 100 subjects were used for the training procedure [29]. Such annotations allow learning segmentation-based ear detectors, and computation of bounding boxes that are commonly returned by standard ear (and object) detectors. Additionally, bounding boxes for the whole test set were prepared.

The two main sources of images were the datasets presented in [29,35]. However, we not only joined the images, but are also releasing for the first time original images from which the ear images were cropped out for the UERC competition, with the annotated locations of ears. This makes the dataset, to the best of our knowledge, one of the largest and the most challenging datasets freely available for both detection and recognition tasks. The data is split into two parts:

- Train part: 1804 images of 116 subjects. 1000 images of 100 subjects are available with pixel-wise annotations of ear locations intended for training ear detection and segmentation models. These images are already presented in [29]. For the recognition part, also additional 804 images of 16 subjects are supplied from the CVL dataset [33].
- Test part: 2200 images of 220 subjects intended for the ear pipeline evaluation. These images contain bound-box ground truth locations of ears and are not appropriate for training pixel-wise detectors, but nevertheless useful for evaluation or for use in ear recognition. This makes this dataset a perfect match for the ear pipeline evaluation.

All of the data is made freely available on the website (http://awe.fri.uni-lj.si). Some of the sample images from the dataset are shown in Fig 6.

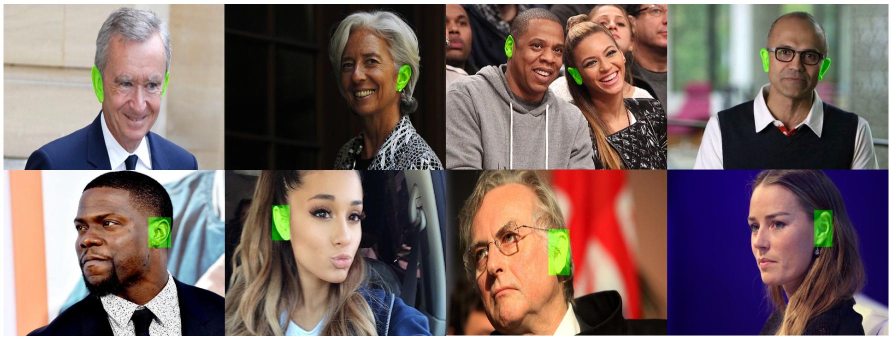


Fig. 6: Some sample images from the dataset with the corresponding annotations. The train part of the dataset shown in the top row contains pixel-wise annotations and the test set in the bottom row contains bounding-boxes. All images are also subject-annotated, meaning they are useful for recognition tasks as well. The images in this figures are resized to a fixed resolution to better present them, in the dataset they are available in the original resolution of different aspect ratios.

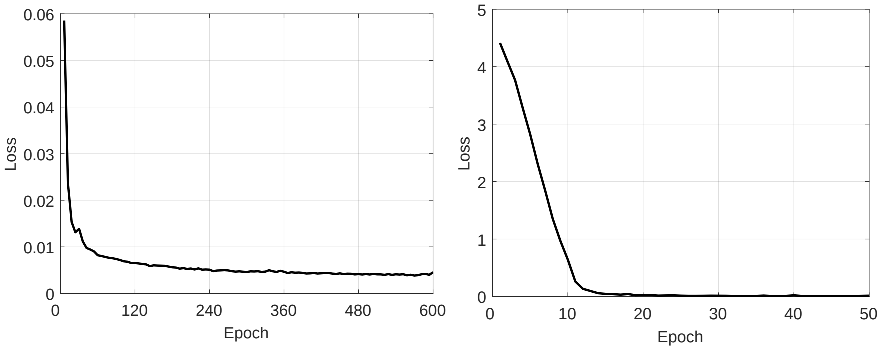
Both the detection network architectures and recognition network architecture are trained on the same train set with some data omitted during training the detection models, because only 1000 out of 1804 images are pixel-wise annotated. However, all images in the test set are used for evaluating the whole pipeline. The set contains annotations of both ear locations and subject identities as described in Section 4.1. For preliminary results the base train set of 1000 images is split into the true training data and the validation set with ratio of 3:1 for the detection experiments and 4:1 for the recognition experiments. The reason for the different ratio for recognition is to ensure as large amounts of train data for recognition as possible. These findings are in line with the literature [34], where the authors emphasize the importance of large amounts of data for CNN-recognition tasks. Especially problematic is a number of images per class when dealing with recognition problems.

The training of CNNs is conducted as a closed set problem, due to the nature of CNNs. However, this is not applicable to real life, where it is coveted to predict identities that the model has never seen before. But using pipeline that would require CNNs to re-learn is not desirable. Therefore, after training, the last fully-connected layers from all the recognition networks are cut and used as feature extractors. This also enabled us to use them on pair with the traditional feature extractors. Note that this so-called open-set problem is much more challenging, but at the same time makes our pipeline a suitable foundation for ear recognition that is deployable in real-life scenarios.

Due to the fact that in the literature many experiments conducted using CNNs report a closed-set results, in Section 4.3 closed-set results on the validation set are reported as well. However, it is important to emphasize that these scores are only for the readers' big picture overview – to get a better picture of how well recognition actually works. For the final scores, readers are referred to the open-set results. For the detection the following parameters were set experimentally for SegNet and PEDCED: the learning rate to the value of 0.0001 [78], the momentum to 0.9 [47] and the weight decay to 0.005 [56]. For the RefineNet the parameters are left to the default as the preliminary tests showed satisfactory results. The learning rate is set to 0.00005 for 600 epochs. However, the number of prediction classes was changed to ear and non-ear. For the traditional feature extractors the default values set in the AWE Toolbox [36] are used. For the CNNs used for recognition the parameters shown in Table 1 were set experimentally according to the video memory available on the Titan Xp GPU and by fine-tuning the hyper-parameters on the training data. Loss values during training for the detection and the recognition part is shown in Fig 7.

Table 1: Experimentally set and used parameters for training recognition models.

	Learning Rate	Batch Size	Momentum	Weight Decay
ResNet Group	0.01	16	0.85	0.001
MobileNet Group	0.01	32	0.75	0.005



(a) RefineNet.

(b) ResNet.

Fig. 7: Plots of losses for RefineNet shown in (a) and ResNet shown in (b). The training of RefineNet converges after approximately 250 epochs and the training of ResNet converges after approximately 15 epochs.

Training of the detection CNNs: SegNet, PEDCED and RefineNet was performed on randomly initialized weights. Training of the detection models was therefore done from scratch. Training the recognition CNNs: the group of ResNet models and the group of MobileNet models was, however, performed after loading ImageNet weights. Training of the recognition CNNs was therefore performed as a transfer learning. The reason for this decision is that preliminary tests showed these two options are the most promising.

For the detection part of the pipeline RefineNet implementation in Matlab (and MatConvNet) available online (https://github.com/guosheng/refinenet) is used. For the SegNet and PED-CED the Caffe framework [29, 48] is used. As a basis for SegNet, the code available online (https://github.com/alexgkendall/caffe-segnet) is used and for the PED-CED the implementation from [29] is used. For the recognition part, for both MobileNet and ResNet, GluonCV (https://gluon-cv.mxnet.io/) is used because more consistent results are achieved more easily compared to architectures written in Keras and Tensorflow. For the traditional feature extraction methods AWE Toolbox [36] is used.

4.2 Performance metrics

For the presentation of the ear recognition pipeline performances on recognition and detection are reported. The recognition experiments are organized as the identification problem, whereas the detection is a two-class segmentation problem.

4.2.1 Detection metrics

Five types of measurements are used to report detection scores. The first one, accuracy is defined as:

$$Accuracy = \frac{TP+TN}{All}, \tag{1}$$

where TP stands for the number of true positives, i.e., the number of pixels that are correctly classified as part of an ear, TN stands for the number of true negatives, i.e., the number of pixels that are correctly classified as non-ear pixels, and All denotes the overall number of pixels in the given test image. This accuracy value measures the quality of the segmentation, but is dominated by the non-ear pixels (i.e., the majority class), which commonly cover most of the test image. Thus, our accuracy measure is expected to have large values (close to 1) even if most pixels are classified as belonging to the non-ear class.

The second performance metric used for our detection experiments is the the Intersection over Union (IoU), which is calculated as follows:

$$IoU = \frac{TP}{TP+FP+FN}, \tag{2}$$

where FP and FN denote the number of false positives (i.e., ear pixels classified as non-ear pixels) and number of false negatives (i.e., non-ear pixels classified as ear pixels), respectively. IoU represents the ratio between the number of pixels that are present in both the ground-truth and detected ear areas and the number of pixels in the union of the annotated and detected ear areas. As such it measures the quality (or tightness) of the detection. A value of 1 means that the detected and annotated ear areas overlap perfectly, while a value of 0 indicates a completely failed detection, i.e. no detection at all or a detection outside the actual ear area.

The third and the fourth performance metrics reported for our experiments are recall and precision respectively, defined as:

$$Precision = \frac{TP}{TP+FP}, \tag{3}$$

$$Recall = \frac{TP}{TP+FN}. \tag{4}$$

Precision measures the proportion of correctly detected ear-pixels with respect to the overall number of true ear pixels (i.e., how many detected pixels are relevant), while recall measures the proportion of correctly detected ear-pixels with respect to the overall number of detected ear pixels (i.e., how many relevant pixels are detected).

The last reported measure for the experiments is E_2, which considers both type-I and type-II error rates. A lower value of E_2 implies better performance and $E_2 = 0$ means maximum precision and maximum recall (i.e., no false negatives and no false positives). The performance measure E_2 compensates for the disproportion in the apriori probabilities of the ear and non-ear classes [73] and is defined as the average of the false positive ($FPR = FP/All$) and false negative ($FNR = FN/All$) rates, i.e.:

$$E_2 = \frac{FPR+FNR}{2}. \tag{5}$$

4.2.2 Recognition metrics

For the recognition part of the pipeline, identification experiments are performed. This means that for each sample the subject identity (class) is predicted by selecting the sample whose feature vector is the closest. If such closest sample belongs to the same class as the observed sample the classification is regarded as correct. After this is repeat over all samples rank values can be calculated.

Rank-1 and rank-5 measures are reported, where rank-*n* means observation of a top-*n* set of the closest samples, where *n* is the number of samples of the observed class. For the visual inspection, all ranks are plotted – from 1 to the number of classes in the test set, i.e. 220 into a Cumulative Match-score Curves (CMCs). Based on this curve, Area Under the CMC (AUCMC) is also reported. The latter gives good estimate on how well the algorithm orders (classifies) all the samples and not only for the top one or the top five classes.

4.3 Evaluation of the ear detection model

The results presented in Table 2 show the superior performance of RefineNet over the other two approaches, SegNet and PED-CED. This is on pair with literature, where RefineNet achieved remarkable results [40, 81]. These results were obtained using subset of train set. The reason for this is, because it is preferable to have a strict pixel-wise evaluation of all of the approaches before selecting one for the final ear recognition pipeline. The results show that RefineNet with Intersection Over Union (IOU) of 84.8% significantly outperforms the second best approach: PED-CED with 55.7%. However, in order to get a more in-depth view of detection scores the histograms of IOU metrics shown in Fig. 8 need to be observed. The distribution scores emphasize one of the largest differences between RefineNet and the other two – a small number of completely failed detections. Both SegNet and PED-CED approaches have a significant share of missed detections. This makes a RefineNet a perfect candidate for the pipeline and is also the approach used in the final experiments. Furthermore, space- and time-complexity wise these three models are close, as opposed to traditional feature extraction approaches and CNN-based approaches for ear recognition evaluated in Section 4.4.

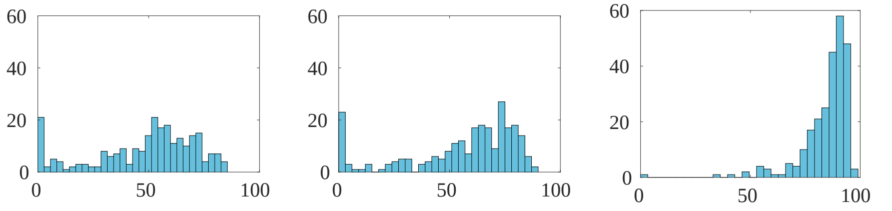
Table 2: Comparison of the pixel-wise detection approaches. The table shows the average accuracy of the detections (Accuracy), the Intersection Over Union (IOU), the average precision and recall values and the E_2 error measure over the test images. Standard deviations are also reported for all techniques. The metrics are computed over 250 test images. Note that all of the approaches were evaluated using strict pixel-wise annotation.

Approach	Accuracy [%]	IOU [%]	Precision [%]	Recall [%]	E2 [%]
SegNet	99.2	48.3	60.8	75.9	25.8
PED-CED	99.4	55.7	67.7	77.7	22.2
RefineNet	**99.8**	**84.8**	**91.7**	**91.6**	**7.6**

In order to show the typical failures and visually inspect the performance, some samples are shown in Fig. 9. However, these examples had to be cherry-picked. Refinenet proved to be really successful. In Fig. 10, arguably some of the most difficult cases are shown to be correctly detected.

4.4 Evaluation of the ear recognition model

For the recognition results, a two-fold evaluation is presented – preliminary results on the validation set (closed-set experiments) and the final results on the test set (open-set experiments). For the test set, two sets of results are presented: results on the manually cropped ear images and results on the images segmented by RefineNet



(a) SegNet. (b) PED-CED. (c) RefineNet.

Fig. 8: Histograms for the Intersection-over-Union (IOU) metric for the three evaluated detection approaches. The histograms for the RefineNet approach shows a much better distribution than the other two approaches with most of the mass concentrated at the higher IOU values.

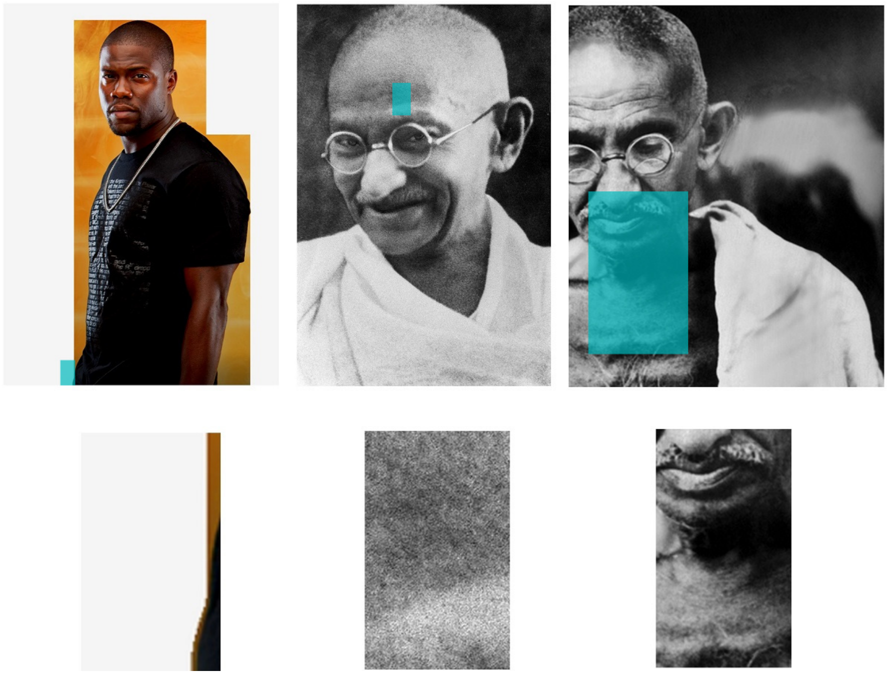


Fig. 9: Three examples of failed detections.

in the detection step. The reason is to isolate the effects of detection on recognition and to evaluate recognition itself as well. The close-set experiments mean that the identities are shared between the training and the validation set. This means, that the fully connected layer is used and the final output of the CNN are the identities prediction. On the open-set however, completely new identities (subjects) are used. This makes tests like this rigorous, but also useful, as in the real-life deployment,

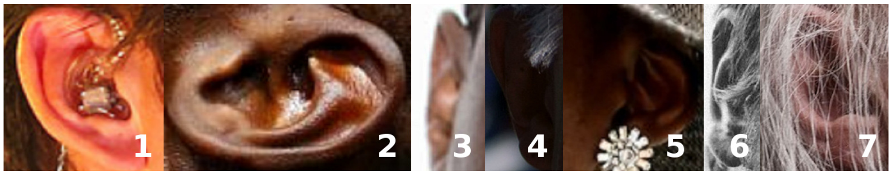


Fig. 10: Examples of a successful detections despite being arguably difficult to detect. The first image contains large ear accessory, the second has large rotation angle, the third high angle, the fourth image is very dark, the fifth contains protruding ear accessory and hat on top, the sixth is in grey-scale, has significant angle and a lot of hair, and the seventh contains large amount of hair covering the ear. Note that these images were not carefully selected, RefineNet truly proved to be very robust.

retraining the whole network only to perform classification tests is not always a viable option. However, the final ranking numbers appear to be lower, but arguably the open-set also presents a much more challenging task.

To capture properties of ear recognition section, also a comparison between space complexities of the approaches in Table 3 and time complexities of the approaches measured in milliseconds in Table 4 is made. The latter is, of course, machine-dependent; using different machine will yield different values. However, relative differences between the approaches are what is important and should stay the same across the systems. In Fig. 11, typical samples that achieved high recognition results and low recognition results are shown. The most problematic samples proved to be the ones where detection part failed in the first place. However, in samples that contained correct detections, the ones containing accessories, high occlusions, bad lightning and high angles proved to be the most problematic, which is on pair with findings in [31]. In Fig. 11, two samples (images 15 and 16) that contain perturbing ear accessories are one of the samples that were correctly detected, but still proved to be problematic for recognition. Images 9-13 do not contain enough (or none at all) ear biometric data for ear recognition.

4.4.1 Closed-set experimental evaluation

In Table 5, a closed-set results on validation set are shown. Note that these numbers server as a representation only. Using closed-set protocol in real life applications is not useful, as this means that during enrollment stage, for each new subject, recognition CNN models need to be retrained. The final scores are presented in Section 4.4.2.

Table 3: Space complexity. The table shows a comparison of all considered techniques with respect to different characteristics such as the model size, number of parameters to train, feature vector size, training time and average test time.

Method	Model size (in *MB*)	# Parameters to train	Feature vector size
BSIF	0	0	9,216
DSIFT	0	0	12,800
HOG	0	0	8,712
LBP	0	0	9,971
LPQ	0	0	9,216
POEM	0	0	11,328
RILPQ	0	0	9,216
ResNet-152	234.1	25,636,712	2,048
MobileNet (1)	19.8	3,347,764	1,024

Table 4: Time complexity. The table shows a comparison of all considered techniques with respect to different characteristics such as the model size, number of parameters to train, feature vector size, training time and average test time.

Method	Training time (in *min*)	Average test time - per image (in *ms*)
BSIF	0	8
DSIFT	0	8
HOG	0	4
LBP	0	18
LPQ	0	6
POEM	0	25
RILPQ	0	25
ResNet-152	~ 10	7
MobileNet (1)	~ 2	2

Table 5: Closed-set intermediate recognition results after the training set. RNet denotes ResNet and MNet MobileNet, respectively.

	RNet-18	RNet-50	RNet-101	RNet-152	MNet ($\frac{1}{4}$)	MNet ($\frac{1}{2}$)	MNet (1)
Rank-1 [%]	68.1	72.4	72.4	**74.6**	45.7	50.4	72.8

	LBP	HOG	DSIFT	BSIF	LPQ	RILPQ	POEM
Rank-1 [%]	12.5	13.8	11.6	11.2	10.8	10.3	13.8

4.4.2 Open-set experimental evaluation

Table 6 and Figs. 12 and 13 show results on the manually cropped ear images. These separate results from the combined detection scores serve as a representation of how well each separate recognition approach works. In Fig. 12, all approaches are plotted – traditional feature extractors on the left and CNN-based extractors on the right. The best performing are then compared and plotted in Fig. 13. Here ResNet-152

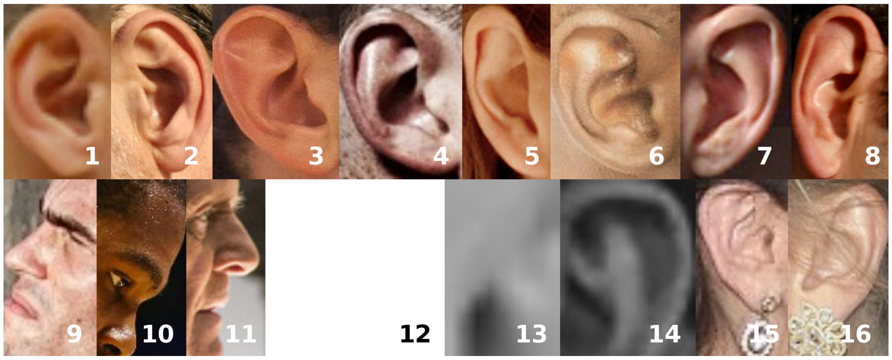
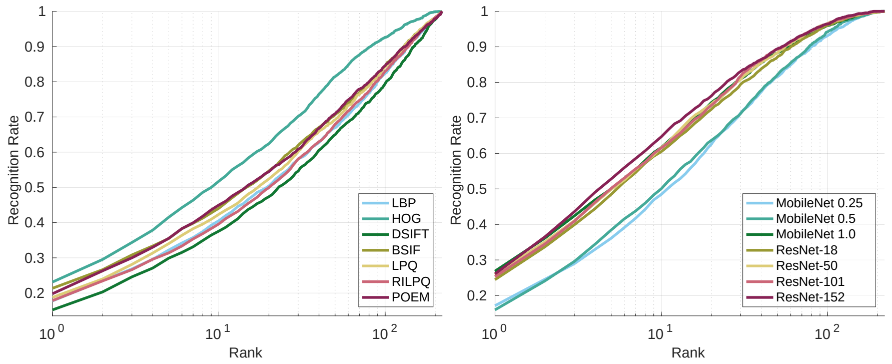


Fig. 11: Some selected examples of successful recognitions (first row) and bad recognition performance (second row). Images 9-11 contain faces that are a cause of bad detections by RefineNet. Image 12 contains complete mis-detection and is therefore impossible to use in recognition. Images 13 and 14 are cropped too tightly. Images 15 and 16 were correctly detected and cropped, but contain protruding ear accessories, that proved to be challenging for the recognition procedure.

with 92.6% AUCMC and MobileNet (1) with 26.9% rank-1 significantly outperform traditional feature extraction methods, such as HOG and BSIF with rank-1 of 23.1% and 21.4%, respectively, although the models have never seen any samples from the subjects. The only two CNN-based approaches that achieve lower scores compared to non-CNN approaches are MobileNet ($\frac{1}{4}$) and MobileNet ($\frac{1}{2}$). Presumably, the reason for this is that these two architectures do not capture the complexity of ear features in depth enough. MobileNet (1) and all three ResNet setups do that significantly better.

Table 6: Open-set recognition results using manually cropped ear images.

	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
MobileNet ($\frac{1}{4}$)	17.1	36.1	88.0
MobileNet ($\frac{1}{2}$)	16.0	38.5	88.5
MobileNet (1)	**26.9**	50.0	91.8
ResNet-18	24.5	48.5	91.4
ResNet-50	25.9	49.9	92.0
ResNet-101	25.3	50.2	92.1
ResNet-152	26.1	**52.8**	**92.6**
LBP	17.8	32.2	79.6
HOG	23.1	41.6	87.9
DSIFT	15.2	29.9	77.5
BSIF	21.4	35.5	81.6
LPQ	18.8	34.1	81.0
RILPQ	17.9	31.4	79.8
POEM	19.8	35.6	81.5



(a) Traditional feature extractors. (b) CNN-based approaches.

Fig. 12: CMC plot in logarithmic scale showing the recognition performance on manually cropped ear images. On the left traditional dense-feature-extraction approaches, on the right CNN-based.

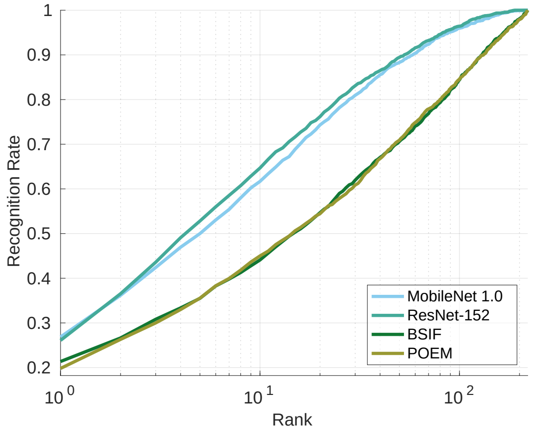


Fig. 13: CMC plot in logarithmic scale showing the comparison of recognition performance on manually cropped ear images. The contrast between the two top performing traditional feature-extraction approaches, BSIF and POEM, and two CNN models, ResNet-152 and MobileNet, is big. This difference makes CNN-based approaches the obvious choice for the pipeline.

4.4.3 Evaluation of the complete pipeline

In Table 7 and Fig. 14 open-set results using RefineNet as an ear detector are shown. All numbers as expected drop compared to the first case, where manually cropped images were used. The reason is because RefineNet detections are not completely accurate and the recognition approaches then need to distinguish subjects using some

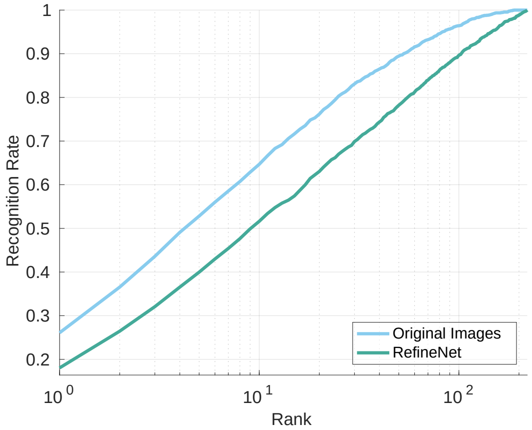


Fig. 14: CMC plot in logarithmic scale comparing the performance of ResNet-152 on annotated ear images vs. ears detected using RefineNet.

arbitrary pieces of information; or in some small number of cases images are missing all together (failed detections) as illustrated in Fig 9. Furthermore, here traditional non-CNN-based approaches perform well compared to the CNN-based ones. Rank-1 scores remain high, with BSIF approach even surpassing ResNet. Nevertheless, the AUCMC scores are still significantly higher for the CNN-based approaches and arguably this is the indicator showing that the overall performance is still better.

Table 7: Final open-set recognition results using RefineNet as the ear detector.

	Rank-1 [%]	Rank-5 [%]	AUCMC [%]
MobileNet ($\frac{1}{4}$)	9.5	24.9	79.5
MobileNet ($\frac{1}{2}$)	10.4	25.8	80.7
MobileNet (1)	17.0	35.5	85.0
ResNet-18	16.3	34.6	84.1
ResNet-50	17.5	36.0	84.8
ResNet-101	16.4	36.0	85.1
ResNet-152	18.0	**40.0**	**85.9**
LBP	18.0	30.6	78.1
HOG	**21.7**	37.4	85.6
DSIFT	14.1	27.0	75.5
BSIF	20.1	34.0	79.6
LPQ	17.5	30.4	78.4
RILPQ	16.5	29.7	78.2
POEM	19.2	33.8	79.3

5 Conclusions

In this chapter, the first freely available, CNN-based ear recognition pipeline is presented. This joint pipeline makes it possible to use arbitrary images of subjects taken in an uncontrolled environment and recognize subjects (predict identity) based only on ears, with no prior knowledge of ear locations. With the use of RefineNet for the ear extraction from unconstrained images of subjects and ResNet for the feature extraction, the pipeline achieves state-of-the-art results. RefineNet detection part achieves 84.8% IOU, when measured with a strict pixel-wise criteria. The recognition scores with ResNet-152 on a closed set go up to 74.6%. A remarkable result, considering the difficulty of the dataset with various levels of occlusions, variable illumination conditions, different poses, different image resolutions etc. On the open-set 26.1% rank-1 and 92.6% AUCMC are achieved using ResNet-152. The final scores for the whole pipeline using RefineNet for detection and ResNet-152 for recognition, where the input consists of an arbitrary images of subjects is 18.0% rank-1 recognition rate and 85.9% AUCMC. The input consists of 2200 images of 220 subjects that both, the detection CNN and the recognition CNN network, have never seen before. CNN outputs are treated as feature vectors in order to make it robust towards new identities.

Nevertheless, many possible improvements still remain. One of them is the use of pixel-wise annotations in the recognition process as well, instead of plainly using bounding-boxes (cropped ear images). Furthermore, feature extraction process could further be improved by modifying the CNN architecture, possibly adding shortcut connections or deepening it. Another possible aspect addressed in the future, as a part of the pipeline, is accessories-aware ear recognition, where ear accessories are first detected and then appropriately addressed during ear recognition stages, making ear recognition more robust. Hopefully this new joint pipeline will help researchers in the future and help progress the field of ear biometrics even further. The ear recognition pipeline could also be used as a complement to some existing face recognition pipelines, making biometric recognition as a whole more accurate and thus widening the impact of ear detection and recognition.

Acknowledgements

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Program P2-0250 (B) Metrology 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 and our colleague Blaž Meden for his help with RefineNet's Matlab scripts.

References

1. Abaza, A., Hebert, C., Harrison, M.A.F.: Fast learning ear detection for real-time surveillance. In: International Conference on Biometrics: Theory Applications and Systems. pp. 1–6. IEEE (2010)
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. Alaraj, M., Hou, J., Fukami, T.: A neural network based human identification framework using ear images. In: International Technical Conference of IEEE Region 10. pp. 1595–1600. IEEE (2010)
4. Ansari, S., Gupta, P.: Localization of ear using outer helix curve of the ear. In: International Conference on Computing: Theory and Applications. pp. 688–692. IEEE (2007)
5. Arbab-Zavar, B., Nixon, M.S.: On shape-mediated enrolment in ear biometrics. In: International Symposium on Visual Computing. pp. 549–558. Springer (2007)
6. Arbab-Zavar, B., Nixon, M.S.: Robust log-Gabor filter for ear biometrics. In: International Conference on Pattern Recognition. pp. 1–4. IEEE (2008)
7. Attarchi, S., Faez, K., Rafiei, A.: A new segmentation approach for ear recognition. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 1030–1037. Springer (2008)
8. Badrinarayanan, V., Handa, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv:1505.07293 (2015)
9. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481–2495 (2017)
10. Banerjee, S., Chatterjee, A.: Robust multimodal multivariate ear recognition using kernel based simultaneous sparse representation. Engineering Applications of Artificial Intelligence 64, 340–351 (2017)
11. 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)
12. Basit, A., Shoaib, M.: A human ear recognition method using nonlinear curvelet feature subspace. International Journal of Computer Mathematics 91(3), 616–624 (2014)
13. 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)
14. Benzaoui, A., Hezil, N., Boukrouche, A.: Identity recognition based on the external shape of the human ear. In: International Conference on Applied Research in Computer Science and Engineering. pp. 1–5. IEEE (2015)
15. Bourouba, H., Doghmane, H., Benzaoui, A., Boukrouche, A.H.: Ear recognition based on Multi-bags-of-features histogram. In: International Conference on Control, Engineering Information Technology. pp. 1–6 (2015)
16. Bustard, J.D., Nixon, M.S.: Toward unconstrained ear recognition from two-dimensional images. Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 40(3), 486–494 (2010)
17. Carreira-Perpinan, M.A.: Compression neural networks for feature extraction: Application to human recognition from ear images. Master's thesis, Faculty of Informatics, Technical University of Madrid, Spain (1995)
18. Chan, T.S., Kumar, A.: Reliable ear identification using 2-D quadrature filters. Pattern Recognition Letters 33(14), 1870–1881 (2012)
19. Chidananda, P., Srinivas, P., Manikantan, K., Ramachandran, S.: Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization. Machine Vision and Applications 26(2), 185–203 (2015)

20. 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 (2018)
21. Cummings, A.H., Nixon, M.S., Carter, J.N.: A novel ray analogy for enrolment of ear biometrics. In: International Conference on Biometrics: Theory Applications and Systems. pp. 1–6. IEEE (2010)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision and Patten Recognition. pp. 886–893. IEEE (2005)
23. 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)
24. Dewi, K., Yahagi, T.: Ear photo recognition using scale invariant keypoints. In: Computational Intelligence. pp. 253–258 (2006)
25. Dodge, S., Mounsef, J., Karam, L.: Unconstrained ear recognition using deep neural networks. IET Biometrics (2018)
26. Dogucan, Y., Fevziye, E., Ekenel, H.: Domain adaptation for ear recognition using deep convolutional neural networks. IET Biometrics 7(3), 199–206 (2018)
27. Ear Recognition Laboratory at the University of Science & Technology Beijing: Introduction to USTB ear image databases (2002), http://www1.ustb.edu.cn/resb/en/index.htm, visited on 2018-03-15
28. Earnest, H., Segundo, P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics 7(3), 215–223 (2018)
29. Emeršič, Ž., Gabriel, L.L., Štruc, V., Peer, P.: Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. IET Biometrics 7(3), 175–184 (2018)
30. 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
31. Emeršič, Ž., Meden, B., Peer, P., Štruc, V.: Covariate analysis of descriptor-based ear recognition techniques. In: Bioinspired Intelligence (IWOBI), 2017 International Conference and Workshop on. pp. 1–9. IEEE (2017)
32. 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)
33. Emeršič, Ž., Peer, P.: Ear biometric database in the wild. In: Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference on. pp. 27–32. IEEE (2015)
34. Emeršič, Ž., Štepec, D., Štruc, V., Peer, P.: Training convolutional neural networks with limited training data for ear recognition in the wild. In: 12th IEEE International Conference on Automatic Face and Gesture (FG 2017) (2017)
35. 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. International Joint Conference on Biometrics (IJCB) (2017)
36. Emeršič, Ž., Štruc, V., Peer, P.: Ear recognition: More than a survey. Neurocomputing 255, 26–39 (2017)
37. Ganesh, M.R., Krishna, R., Manikantan, K., Ramachandran, S.: Entropy based binary particle swarm optimization and classification for ear detection. Engineering Applications of Artificial Intelligence 27, 115–128 (2014)
38. Gourier, N., Hall, D., Crowley, J.L.: Estimating face orientation from robust detection of salient facial structures. In: FG Net Workshop on Visual Observation of Deictic Gestures. vol. 6 (2004)
39. Guo, Y., Xu, Z.: Ear recognition using a new local matching approach. In: International Conference on Image Processing. pp. 289–292. IEEE (2008)
40. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: A survey. IEEE Signal Processing Magazine 35(1), 84–100 (2018)

41. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. arXiv preprint arXiv:1710.07662 (2017)
42. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645. Springer (2016)
43. He, X., Yu, Z., Wang, T., Lei, B.: Skin lesion segmentation via deep refinenet. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 303–311. Springer (2017)
44. He, X., Yu, Z., Wang, T., Lei, B., Shi, Y.: Dense deconvolution net: Multi path fusion and dense deconvolution for high resolution skin lesion segmentation. Technology and Health Care pp. 1–10 (2018)
45. 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. arXiv preprint arXiv:1704.04861 (2017)
46. Islam, S.M., Bennamoun, M., Davies, R.: Fast and fully automatic ear detection using cascaded adaboost. In: Workshop on Applications of Computer Vision. pp. 1–6. IEEE (2008)
47. Jacobs, R.A.: Increased rates of convergence through learning rate adaptation. Neural networks 1(4), 295–307 (1988)
48. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM international conference on Multimedia. pp. 675–678. ACM (2014)
49. Kannala, J., Rahtu, E.: BSIF: Binarized statistical image features. In: International Conference on Pattern Recognition. pp. 1363–1366. IEEE (2012)
50. Križaj, J., Štruc, V., Pavešic, N.: Adaptation of SIFT features for robust face recognition. In: Image Analysis and Recognition. pp. 394–404. Springer (2010)
51. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recognition 45(3), 956–968 (2012)
52. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, p. 5 (2017)
53. Lu, Z., Jiang, X., Kot, A.: Deep coupled resnet for low-resolution face recognition. Signal Processing Letters 25(4), 526–530 (2018)
54. Meraoumia, A., Chitroub, S., Bouridane, A.: An automated ear identification system using Gabor filter responses. In: International Conference on New Circuits and Systems. pp. 1–4. IEEE (2015)
55. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. In: International conference on audio and video-based biometric person authentication. vol. 964, pp. 965–966 (1999)
56. Moody, J., Hanson, S., Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. Advances in neural information processing systems 4, 950–957 (1995)
57. of Notre Dame, U.: Face database (2015), https://sites.google.com/a/nd.edu/public-cvrl/datasets, visited on 2018-03-01
58. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: International conference on image and signal processing, pp. 236–243 (2008)
59. 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 (2008)
60. Omara, I., Wu, X., Zhang, H., Du, Y., Zuo, W.: Learning pairwise SVM on hierarchical deep features for ear recognition. IET Biometrics (2018)
61. Pflug, A., Busch, C., Ross, A.: 2D ear classification based on unsupervised clustering. In: International Joint Conference on Biometrics. pp. 1–8. IEEE (2014)
62. Pflug, A., Busch, C.: Ear biometrics: A survey of detection, feature extraction and recognition methods. IET Biometrics 1(2), 114–129 (2012)

63. 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 (2014)
64. Pflug, A., Winterstein, A., Busch, C.: Robust localization of ears by feature level fusion and context information. In: International Conference on Biometrics. pp. 1–8 (2013)
65. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and vision computing 16(5), 295–306 (1998)
66. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer vision using local binary patterns. Computational Imaging and Vision, Springer (2011)
67. Prakash, S., Gupta, P.: An efficient ear localization technique. Image and Vision Computing 30(1), 38 – 50 (2012)
68. Prakash, S., Gupta, P.: An efficient ear recognition technique invariant to illumination and pose. Telecommunication Systems 52(3), 1435–1448 (2013)
69. Prakash, S., Gupta, P.: Ear biometrics in 2D and 3D: Localization and recognition, vol. 10. Springer (2015)
70. Prakash, S., Jayaraman, U., Gupta, P.: Ear localization from side face images using distance transform and template matching. In: Workshops on Image Processing Theory, Tools and Applications. pp. 1–8 (2008)
71. Prakash, S., Jayaraman, U., Gupta, P.: Connected component based technique for automatic ear detection. In: International Conference on Image Processing. pp. 2741–2744. IEEE (2009)
72. Prakash, S., Jayaraman, U., Gupta, P.: A skin-color and template based technique for automatic ear detection. In: International Conference on Advances in Pattern Recognition. pp. 213–216. IEEE (2009)
73. Proença, H., Alexandre, L.A.: The NICE.I: Noisy iris challenge evaluation - part I. In: International Conference on Biometrics: Theory, Applications, and Systems. pp. 1–4. IEEE (2007)
74. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
75. Sarangi, P.P., Panda, M., Mishra, B.S.P., Dehuri, S.: An automated ear localization technique based on modified hausdorff distance. In: International Conference on Computer Vision and Image Processing. pp. 1–12 (2016)
76. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: International Conference on Automatic Face and Gesture Recognition. pp. 53–58. IEEE (2002)
77. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
78. Szkuta, B.R., Sanabria, L.A., Dillon, T.S.: Electricity price short-term forecasting using artificial neural networks. IEEE Transactions on Power Systems 14(3), 851–857 (1999)
79. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). pp. 437–441. IEEE (2016)
80. University of Sheffield: The sheffield (previously UMIST) face database (1998), https://www.sheffield.ac.uk/eee/research/iel/research/face, visited on 2016-05-01
81. Urooj, A., Borji, A.: Analysis of hand segmentation in the wild. In: Conference on Computer Vision and Pattern Recognition, IEEE. pp. 4710–4719 (2018)
82. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition. pp. I–I. IEEE (2001)
83. Vu, N.S., Caplier, A.: Face recognition with patterns of oriented edge magnitudes. European conference on computer vision pp. 313–326 (2010)
84. Wahab, N.K.A., Hemayed, E.E., Fayek, M.B.: HEARD: An automatic human ear detection technique. In: International Conference on Engineering and Technology. pp. 1–7 (2012)
85. Zhang, Y., Mu, Z., Yuan, L., Yu, C.: Ear verification under uncontrolled conditions with convolutional neural networks. IET Biometrics 7(3), 185–198 (2018)