Journals
The Institution of Engineering and Technology

Convolutional Encoder-Decoder Networks for Pixel-wise Ear Detection and Segmentation

ISSN 1751-8644
Received: date / Accepted: date
doi: 0000000000
www.ietdl.org

Žiga Emeršič[1], Luka L. Gabriel[2], Vitomir Štruc[3], Peter Peer[1]*

[1] Faculty of Computer and information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia
[2] KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden
[2] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
** E-mail: ziga.emersic@fri.uni-lj.si*

Abstract: Object detection and segmentation represents the basis for many tasks in computer and machine vision. For instance, in biometric recognition systems the detection of the region-of-interest (ROI) is one of the most crucial steps in the overall processing pipeline, significantly impacting the performance of the entire recognition system. Existing approaches to ear detection, for example, are commonly susceptible to the presence of severe occlusions, ear accessories or variable illumination conditions and often deteriorate in their performance if applied on ear images captured in unconstrained settings. To address these shortcomings, we present in this paper a novel ear detection technique based on convolutional encoder-decoder networks (CEDs). We formulate the problem of ear detection as a two-class segmentation problem and design and train a CED-network architecture to distinguish between image-pixels belonging to the ear and the non-ear class. We post-processed the output of the network to refine the segmentation and determine the final locations of the ears in the input image. Unlike competing techniques, our approach does not simply return a bounding box around the detected ear, but provides detailed, pixel-wise information about the location of the ears in the image. Experiments on a dataset gathered from the web (a.k.a. in the wild) show that the proposed technique ensures good detection results in the presence of various covariate factors and significantly outperforms competing methods from the literature.

1 Introduction

Ear recognition is gaining on popularity over the the last few years [1–3]. One of the reasons could be the numerous application possibilities in forensics, security and surveillance. However, despite its potential, only a limited number of fully automatic techniques has been proposed and presented in the literature so far. Many recent surveys on ear recognition ascribe 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 [4–6]. In fact, the authors of a recent survey [6] argue that the absence of automatic ear detection approaches is one the most important factors hindering a wider deployment of ear recognition technology.

While there has been progress in the area of ear detection over the 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 [5]. 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. In these conditions, ear detection is still an unsolved problem and no widely adopted solution has been (to the best of our knowledge) proposed yet in the literature [7].

In this paper, we try to address this gap and introduce a novel ear detection approach based on convolutional neural networks (CNNs). We pose the ear detection problem as a two-class semantic segmentation task, where the goal is to assign each image pixel to either the ear or to the non-ear class. We design a convolutional encoder-decoder network (CED) for this problem by incorporating ideas from successful recent architectures such as SegNet [8, 9], U-Net [10] and the Hourglass model [11] to classify the image pixels into one of the two classes and use the trained network to generate an initial segmentation result from the input image. We then refine the result through a post-processing procedure that takes anthropometric assumptions into account and removes spurious image regions from the final output. Different from existing solutions to ear detection which typically return only bounding rectangles or ellipses for the detected ears in each image, our approach provides information about the locations of the ears at the pixel level. Such information is particularly useful for ear recognition, as it allows to exclude background pixels from the feature extraction and classification stages, which is not possible with standard detection techniques. However, certain restrictions apply: ears that are captured under severe angles or that are too small make ear recognition difficult or in the worst case impossible [12, 13]. Nevertheless, we expect that even with such images, recognition capabilities will improve over time.

We evaluate the pixel-wise ear detection (PED) approach based on our convolution encoder-decoder network (CED), called PED-CED for short, in experiments on the AWE dataset [6], which is a recent dataset of ear images gathered from the web with the goal of studying unconstrained ear recognition technology. The dataset is in our opinion and according to recent work [6, 12] one of the most challenging datasets currently available in the field of ear recognition which makes it a suitable choice for the experiments. In contrast to other datasets typically used to assess ear detection performance, images from the AWE dataset are not limited to perfect profile face images (i.e., faces rotated close to 90° yaw angles), but feature significant variations in yaw angles as well. We present a detailed experimental analysis of the proposed PED-CED approach and study the influence of various covariates on the detection performance. We also report comparative experiments with state-of-the-art segmentation and detection techniques from the literature. Our results indicate that the PED-CED technique is a viable option for ear detection in unconstrained settings and provides competitive detection performance even in the presence of different covariate factors.

To summarize, we make the following contributions in this paper:

- we present a novel ear detection technique based on a convolutional encoder-decoder (CED) network that works well on image data

Table 1 Summary of the surveyed work on ear detection using 2D images. While some works evaluate their methods on multiple datasets, we report only results for the dataset with the highest detection accuracy. * Not all accuracies were calculated using the same equation. Furthermore, different datasets and protocols were used, so conclusions about the relative performance of the listed techniques should be avoided.

Year	Detection Method	Dataset	# Test images	Accuracy [%]
2007	Hough Transform [15]	UND [16]	942	91.0
	Canny Edge Detector [17]	IITK [14]	700	93.3
2008	Canny Edge Detector and Line Tracing [18]	USTB [19]	308	98.1
	Adaboost [21]	UND [16]	203	100
	Distance Transform and Template Matching [22]	IITK [14]	150	95.2
2009	Connected component [23]	IITK [14]	490	95.9
	Skin-Color [24]	IITK [14]	150	94.0
2010	Viola-Jones [25]	UND-F [16]	940	88.7
	Ray Transform [26]	XM2VTS [27]	252	98.4
2012	Skin Color and Graph Matching [28]	IITK [14]	1780	99.3
	HEARD [29]	UND-E [16]	200	98.0
2013	Feature Level Fusion and Context Information [30]	UND-J2 [16]	1776	99.0
2014	EBPSO [31]	FERET [32]	258	95.0
2015	Entropy Hough transform [33]	UMIST [34]	564	100
2016	Modified Hausdorff Distance [35]	UND-E [16]	464	94.5

captured in completely unconstrained settings and returns pixel-wise segmentations of the ears in the input images,

- we provide a detailed analysis of the proposed techniques with respect to various covariates and identify open problems that need to be addressed to further improve its performance, and
- we report comparative results with state-of-the-art segmentation and ear detection techniques from the literature.

2 Related Work

In this section we survey the most important techniques for ear detection 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 [5, 14].

It needs to be noted that no standard benchmarks and evaluation methodology exists for ear detection technology, which makes it difficult to compare existing approaches among each other. Authors typically report different performance metrics and rely on self compiled evaluation protocols in their experiments. Furthermore, since face detection is commonly assumed to have been run on the images before ear detection is performed, the term ear detection is typically used interchangeably with ear localization or even ear enrollment.

In [15] the authors propose an ear enrollment algorithm that fits an ellipse to the ear using the Hough Transform. The approach is sufficiently tolerant to noise and occlusions and achieves a 91.0% enrollment success rate on the UND dataset [16] and 100% on XM2VTS when no occlusions are present. The authors do not explicitly state what constitutes a successful enrollment attempt.

In [17] 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. On the IITK dataset [14] the authors report the localization accuracy of 93.3%, where the accuracy is defined as:

$$\text{accuracy} = \frac{\text{\# of correct detections/localizations}}{\text{\# of all annotated ears}}, \qquad (1)$$

In another work using the Canny edge detector [18], the authors report the localization accuracy of 98.1% on the USTB [19] and 97.1% on the Carreira-Perpinan dataset [20], but similar to [17] do not provide information on how a correct ear detection/localization is defined (i.e., the nominator of Eq. 1).

In the work of [21], a cascaded-AdaBoost-based ear detection approach is proposed. The authors report the detection rate of 100% with the false positive rate of 5×10^{-6} on 203 profile images from the UND dataset [16]. Again no formal criterion is given about the process of establishing the detection and false positive rates, though it is suggested that the results were examined manually.

Another approach to ear detection based on the distance transform and template matching is proposed in [22]. The authors report the detection accuracy (using Eq. (1)) of 95.2% on the IIT Kanpur ear database. The authors define a correct detection as one that exhibits a sufficient level (i.e., above some predefined threshold) of similarity with a generic ear template.

In [23] the connected component analysis of a graph constructed using the edge map of the image and is evaluated on a data set consisting of 2361 side face images from the IITK dataset [14]. The authors fit rectangular regions to the ear images and achieve the detection accuracy of 95.9% on 490 test images and the detection accuracy of 94.7% on another test set of 801 images, when at most 15% more pixels are detected around the annotated ground truth.

In [24] the same authors approach the ear detection problem by segmenting skin-colored regions. Using 150 side face images of the IITK dataset [14] the approach achieves 94.0% detection accuracy. The accuracy is again measured through Eq. (1) and similarly to [22] the correctness of the detection is established based on the similarity between the detected region and a generic ear template.

Haar features arranged in a cascaded Adaboost classifier, better known as Viola-Jones [36], are used in [25] for ear detection. The authors manually annotate the UND-F [16], UMIST [34], WV HTF [25] and USTB [19] datasets with rectangles around ears and use the annotated data for training and testing. The authors achieve 95% detection accuracy on the combined images and 88.7% on the UND-F dataset. This approach is capable of handling a wide variety of image variability and operating in real-time.

The authors of [26] propose an ear enrollment technique using the image ray transform, which highlights the tubular structures of the ear. Using 252 images from the XM2VTS [27] dataset the authors achieve a 99.6% enrollment rate and consider an image as successfully enrolled if after the enrollment/localization process, the entire ear is contained in the localized image area.

The approach presented in [28] 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. The detection performance is evaluated on the IITK, UND-E and UND-J2 datasets, achieving 99.3% accuracy on IITK. As suggested by the authors, the detection accuracy is defined by Eq. (1), but no criterion defining a correct detection is given by the authors.

The HEARD [29] 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

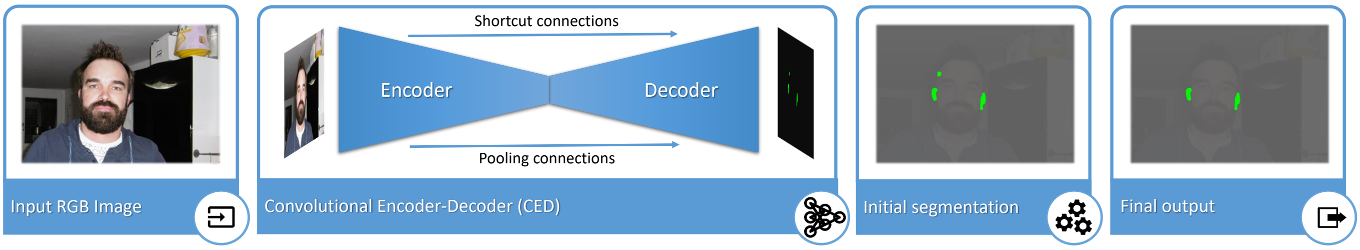


Fig. 1: Overview of the PED-CED ear detection approach. Ear detection is posed as a segmentation problem and solved using a convolutional encoder-decoder network (i.e., the segmentation network). The segmentation network takes an RGB-image (containing a face) as input and returns ear candidate regions as a result. The output is then post-processed and (at most) two largest areas are retained.

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 authors use the UND [16], CVL [37] and IITK [14] datasets. The method is able to detect 98.0% of ears in the UND-E dataset [16]. However, no information is given by the authors on how the reported detection accuracy is calculated.

The ear detection algorithm proposed in [30] 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 with a detection rate of 99.0%. A detection was considered successful when the overlap O between the ground truth pixels G (i.e., the annotated area) and the pixels in the detected region R is at least 50%. The overlap O is calculated by the following equation:

$$O = \frac{2\,|G \bigcap R|}{|G| + |R|}, \tag{2}$$

where $\bigcap$ stands for the intersection operator and $+$ for the union.

In [31] the authors 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. The authors evaluate the detection accuracy using the CMU PIE [38], Pointing Head Pose [39], FERET [32] and UMIST [34] datasets. On FERET the authors report the detection rate of 95.0%, where the detection rate is defined by Eq. (1) and a detection attempt is considered successful if at least part of the ear is contained in the detected area.

The authors of [33] 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. The authors achieve 100.0% detection rate on the UMIST [34] and FEI [40] datasets and 74.0% on FERET. The detection rate is computed with Eq. (1) and a detection attempt is considered successful if the center of the detected region is close enough to the center of the annotated ground truth (i.e., the distance is below some threshold) and the detected area contains the entire ear.

The authors of [35] 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 invariant to illumination, pose, shape and occlusion of the ear images. The detection accuracy of the technique was tested on two datasets, i.e., the CVL face database [37] and the UND-E database [16], on which accuracies of 91.0% and 94.5% were obtained, respectively. The accuracy is calculated by Eq. (1), but no criteria for a correct detection are reported.

A summary of the surveyed work is presented in Table 1. Note again that reported accuracies are not directly comparable, as different datasets, performance metrics and evaluation protocols were used by the authors.

3 Pixel-wise Ear Detection with CEDs

In this section we present our Pixel-wise Ear Detection technique based on Convolutional Encoder-Decoder networks (PED-CED). We start the section with some motivation and a high-level overview of the proposed technique and then describe the segmentation network and the post-processing step used to generate the final detection results.

3.1 Overview and Motivation

Deep encoder-decoder architectures are widely used today for various vision problems, such as image translation [10], image restoration and denoising [41, 42], contour detection [43] and semantic segmentation [44]. The main idea of these architectures is to first produce an abstract high-level encoding of the input image through a hierarchy of convolutional and pooling layers and then decode the generated representation (encoding) into the targeted output format with a series of deconvolutions and unpoolings. Such architectures are particularly suitable for conditional generative tasks, such as semantic segmentation, where an output image with specific target characteristics needs to be generated based on the provided input.

Due to the recent success of encoder-decoder architectures and the availability of pre-trained recognition models that can be exploited for the encoding, we design our PED-CED technique around this class of deep models. Our detection pipeline builds on the assumption that a single face is present in the input image and that the goal is to detect at most two ears. This assumption is reasonable given the fact that the input image is typically subjected to a face detection procedure prior to ear detection. No other assumptions regarding the input images are made, they are free to vary in terms of appearance, imaging conditions and alike, which is not the case for many competing methods, e.g., [15, 17, 18, 26, 28], which often require images of sufficient profile or rely on the visibility of certain ear-parts.

A high-level overview of our PED-CED detection approach is shown in Fig. 2. To detect ears in the image, we exploit a convolutional encoder-decoder (CED) network similar to [8, 9], but design the network to include connections that propagate information from the encoding layers to corresponding decoding layers. These connections ensure that we do not loose resolution (i.e., detailed information) at the output of the network due to the compression of information in the encoder. Different from other CED architectures from the literature, we exploit so-called pooling connections (introduced in [44]) as well as shortcut connections between convolutional layers [10, 11] to retain detail in the outputs of the CED model. As we show in the experimental section, these combined sources of information significantly contribute to the overall performance of the CED segmentation network and allow us to label each pixel in the input image with either the ear or non-ear class label with high efficiency. Because the segmentation procedure sometimes also returns spurious labels, we post-process the segmentation results and retain (at most) the two largest ear regions. This step corresponds to our assumption that a single face is present in the image and, hence, at most two ears may be found by the detection procedure.

A detailed description of the PED-CED approach is given in the following sections.

Table 2 High-level summary of the layers used in our CED architecture. A convolutional layer is always followed by a BN and a ReLU layer.

Layer Number/Label	Type of Layer	Number of Filters	Filter/(Un)pooling Size	Data/Output Size	Note
-	Data/input	-	-	480 × 360 × 3	RGB image
1, 2	Convolutional	64	3 × 3	480 × 360 × 64	No shortcut connections
(a)	Max pooling	-	2 × 2	240 × 180 × 64	Pooling indices forwarded to (j)
3, 4	Convolutional	128	3 × 3	240 × 180 × 128	No shortcut connections
(b)	Max pooling	-	2 × 2	120 × 90 × 128	Pooling indices forwarded to (i)
5, 6, 7	Convolutional	256	3 × 3	120 × 90 × 256	Shortcut connection from layer # 6 to (after) #21
(c)	Max pooling	-	2 × 2	60 × 45 × 256	Pooling indices forwarded to (h)
8, 9, 10	Convolutional	512	3 × 3	60 × 45 × 512	Shortcut connection from layer #9 to (after) #18
(d)	Max pooling	-	2 × 2	30 × 23 × 512	Pooling indices forwarded to (g)
11, 12, 13	Convolutional	512	3 × 3	30 × 23 × 512	Shortcut connection from layer #12 to (after) #15
(e)	Max pooling	-	2 × 2	15 × 12 × 512	Pooling indices forwarded to (f)
(f)	Upsampling/unpooling	-	2 × 2	30 × 23 × 512	Pooling indices forwarded from (e)
14, 15, 16	Convolutional	512	3 × 3	30 × 23 × 512	Layer #16 has 1024 input channels - shortcut from #12
(g)	Upsampling/unpooling	-	2 × 2	60 × 45 × 512	Pooling indices forwarded from (d)
17, 18	Convolutional	512	3 × 3	60 × 45 × 512	Activation maps of layer #18 combined with maps of #9
19	Convolutional	256	3 × 3	60 × 45 × 256	Layer #19 has 1024 input channels
(h)	Upsampling/unpooling	-	2 × 2	120 × 90 × 256	Pooling indices forwarded from (c)
20, 21	Convolutional	256	3 × 3	120 × 90 × 256	Activation maps of layer #21 combined with maps of #6
22	Convolutional	128	3 × 3	120 × 90 × 128	Layer #22 has 512 input channels
(i)	Upsampling/unpooling	-	2 × 2	240 × 180 × 128	Pooling indices forwarded from (b)
23	Convolutional	128	3 × 3	240 × 180 × 128	No shortcut connections
24	Convolutional	64	3 × 3	240 × 180 × 64	No shortcut connections
(j)	Upsampling/unpooling	-	2 × 2	480 × 360 × 64	Pooling indices forwarded from (a)
25	Convolutional	64	3 × 3	480 × 360 × 64	No shortcut connections
26	Convolutional	2	3 × 3	480 × 360 × 2	No shortcut connections
-	Softmax	-	-	480 × 360 × 1	Outputs initial segmentation

3.2 The segmentation network

The main component of the PED-CED detection technique is the convolutional encoder-decoder segmentation network, which we build around the pre-trained VGG-16 model [45] similarly to [8, 9, 44]. The pre-trained VGG-16 model represents a powerful deep model trained on over 1.2 million images of the ImageNet dataset [46] (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 CED models also used in our approach. The decoding part of 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 whole architecture is summarized in Table 2 and Fig. 2.

Similarly to related models, such as [44] or [10], our overall CED model consists of a sequence of nonlinear processing layers (encoders) and a corresponding set of decoders with a pixel-wise classifier on top. As shown in Fig. 2, a single encoder features several convolutional layers, a batch normalization layer, a rectified-linear-unit ReLU layer (shown in gray), and a max-pooling layer (shown in green). The goal of the encoders is to produce low-resolution feature maps that compress the semantic information in the input image and can be fed to the sequence of decoders for upsampling and ultimately segmentation. Similarly to the encoders, each decoder is also composed of several convolutional layers, a batch normalization layer and a ReLU layer (shown in gray on the right side of Fig. 2) followed by an upsampling layer (shown in red). The final layer of the segmentation network is a pixel-wise softmax layer, which in our case, assigns each pixel a label corresponding to one of two classes (i.e., ear or non-ear).To ensure that high frequency details are retained in the segmented images (which is a known problem with CED architectures) we modify the described architecture and add two types of additional connections to the model to propagate information from the encoding layers to the corresponding decoding layers:

- *Pooling connections*: We forward the max-pooling indices from the encoders to the corresponding unpooling layers of the decoders through pooling connections (see lower part of Fig. 2). These connections allow us exploit the information from the size- and resolution-reducing pooling layers during the unpooling operation and consequently to upsample the low-resolution feature maps in a non-linear manner. The pooling connections contribute towards retaining high-frequency details in the segmented images and have initially been introduced for SegNet in [44].

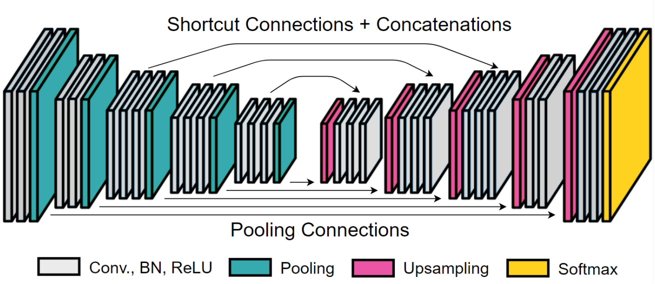


Fig. 2: Illustration of the network structure used for our PED-CED ear detection approach. The network has a encoder-decoder architecture and ensures detailed segmentation results by exploiting propagation of pooling indices from the max-pooling to the unpooling (upsampling) layers similar to [44], but provides an additional source of information by introducing shortcut connections that forward feature maps from the lower convolutional layers of the encoder to the corresponding convolutional layers of the decoder similar to [10, 11, 42]. The layer groups in the figure are marked as: Convolution layers (Conv.), Batch Normalization layers (BN) and ReLU layers in gray; Pooling layers in green; Upsampling/Unpooling layers in purple; and the Softmax layer in yellow (the figure is best viewed in color).

- *Shortcut connections*: We add shortcut connection between the convolutional layers of the encoder and decoder. Specifically, we forward the feature maps 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. We introduce the shortcut connections 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 (i.e., we shortcut layers #6 and #21, layers #9 and #18, and layers #12 and #15). While we also experimented with shortcut connections between other layers, these had a less beneficial effect on performance as shown by the results in Section 5 and were dropped from the final model. In general, the impact of the shortcut connections is two-fold: *i)* they propagate high-resolution information from the encoding layers to the decoding layers and help preserve detail in the segmentation outputs, and *ii)* they contribute towards a more efficient training procedure by reducing the problem of vanishing gradients [47, 48]. Note that shortcut connections have successfully been used for many related problems, e.g., [10, 11, 41].

3.3 Postprocessing

Once the segmentation network is trained it can be used to generate initial segmentation results from the input images. However, these results are not always perfect and despite the fact that only images with a single face are expected as input to our PED-CED detection procedure, several ear candidate regions may be present in the segmentation output. Since the only possible correct result of the segmentation network is the detection of one or two regions (corresponding to either one or two visible ears in the image), we apply an additional post-processing step and clean the initial segmentation output. Thus, we retain only the two largest regions (or one, if only one was detected) and discard the rest.

4 Experimental Setup

In this section we describe the data, experimental protocols and performance metrics used to assess the efficacy of PED-CED ear detection approach.

4.1 Data, experimental protocol and network training

The dataset used in our experiments comprises the original (uncropped) images of the Annotated Web Ears (AWE) dataset [6] and is freely available (both the uncropped version with annotations and the cropped version) *. The dataset contains a total of 1000 annotated images from 100 distinct subject, with 10 images per subject. All images from the dataset were gathered from the web using a semi-automatic procedure and were labeled according to yaw, roll and pitch angles, ear occlusion, presence of accessories, ethnicity, gender and identity. However, new pixel-wise annotations of ear locations had to be created in order to evaluate the performance of the PED-CED approach. To this end, a trained annotator manually marked the ear locations in each image at the pixel level and stored the resulting annotations in the form of binary masks for later processing. Fig. 4 shows a few of the original images that we used for our experiments together with the newly created ground truth. Note that the annotations provide more detailed information about the locations of the ears than simple rectangular bounding boxes and were already used in prior publications, such as [49, 50].

From the available 1000 images in the AWE dataset [6], we use 750 randomly-selected images for training and 250 randomly-selected images for testing purposes. The training images are used to learn the parameters of our CED segmentation network, while the testing images are reserved solely for the final performance evaluation. The images were gathered from the web for the goal of studying ear recognition technology in unconstrained settings and therefore exhibit a high-degree of variability unprecedented in other datasets.

The hardware used for experimentation is a desktop PC with an Intel(R) Core i7-6700K CPU with 32 GiB system memory and an Nvidia GeForce GTX 980 Ti GPU with 6 GiB of video memory running Ubuntu 16.04 LTS. On this hardware the training of our segmentation network takes around one hour and was completed with the network parameters converging after approximately 10000 iterations, when stable loss and accuracy values are reached. Figure 3 shows the loss values, collected in steps of 20 iterations throughout the course of the training process.

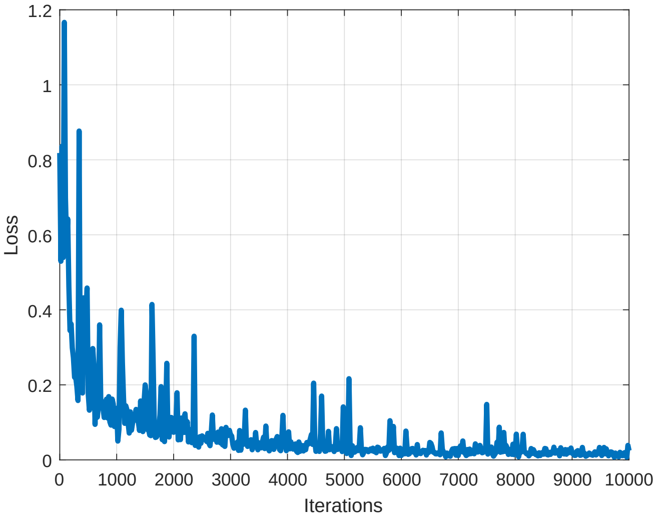


Fig. 3: Graphical representation of the convergence of the loss during network training. The values were sampled every 20 training iterations.

To train the segmentation network we use stochastic gradient descent and set the learning rate to the value of 0.0001, the momentum to 0.9 and the weight decay to 0.005 [51], [52], [53]. We use the publicly available Caffe implementation of SegNet * to initialize our network, but modify the last convolutional and softmax layer and introduce our shortcut connections. We set the number of outputs of the last convolutional layer to 2 (ear and non-ear) and calculate new class weights that we apply to the softmax layer to ensure stable network training. As advocated in [8, 9, 44], we also use median frequency balancing [54] to compute the class weights for our cross-entropy loss, which compensates for the fact that pixels from the ear class cover only a small portion of the input image, while the the remaining pixels belong to the non-ear class. Without frequency balancing, the network would likely converge to a trivial solution, where all pixels would be assigned to the dominant (over-represented) non-ear class. All training images are resized to the resolution of 480 × 360 pixels to reduce the needed graphical memory prior to training.

4.2 Performance metrics

A typical way of measuring the performance of ear detection technology is to use the detection accuracy, which is typically defined as the ratio between the number of correct detections and the overall number of annotated ear areas. However, as already pointed out in Section 2, what is considered a correct detection is usually defined by the authors is not used consistently from paper to paper. Since a single profile face is typically presumed in the input image, the general assumption is that only a single ear needs to be detected (localized or enrolled), so false positives are not considered in the reported results, and the decisive criterion is whether the correct ear was found or not.

In this work, we measure the performance of our detection approach by comparing the manually annotated ground-truth locations and the output of our PED-CED approach during testing. We report accuracy values for our approach, which are computed as follows:

$$Accuracy = \frac{TP + TN}{All}, \tag{3}$$

http://awe.fri.uni-lj.si

https://github.com/alexgkendall/caffe-segnet

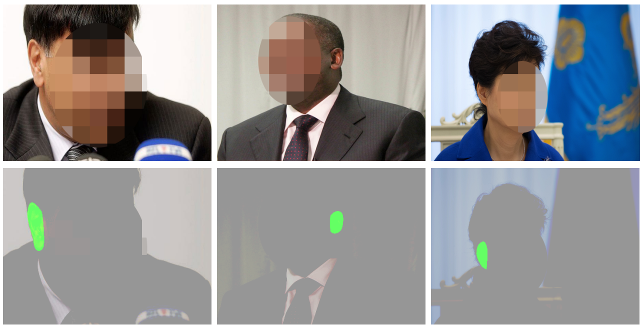


Fig. 4: Sample uncropped images from the AWE dataset. The top row shows the input images and the bottom row shows the annotated ear locations. The ears are annotated at the pixel level in all 1000 images of the AWE dataset (in the images given here, faces were pixelated in order to guarantee anonymity).

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 experiments is the the Intersection over Union (IoU), which is calculated as follows:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{4}$$

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 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:

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}. \tag{6}$$

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 measure we report for our 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 [55] and is defined as the average of the false positive ($FPR = FP/All$) and false negative ($FNR = FN/All$) rates, i.e.:

$$E_2 = \frac{FPR + FNR}{2}. \tag{7}$$

5 Results and Discussion

We now present the experimental results aimed at demonstrating the merits of our PED-CED ear detection approach. We show comparative results with state-of-the-art techniques from the literature, analyze the impact of various covariates on the performance of our technique and show qualitative examples of our detections.

5.1 Assessment and comparison to the state-of-the-art

We evaluate the performance of the PED-CED approach on the entire test set of 250 AWE images and compute all performance metrics presented in Section 4.2. The average values of each metric together with the corresponding standard deviations over the test set are shown in Table 3. Here, we also report results for 3 competing approaches to put our technique into perspective and show comparative results with a state-of-the-art techniques from the literature:

- The Haar-based ear detector from [25] (Haar hereafter), which exploits the established object detection framework proposed by Viola and Jones [36]. For the Haar-based ear detector we use the detection cascades for the left and right ear that ship with OpenCV [56] and optimize the open hyper-parameters for optimal performance. We select this approach, as it makes minimal assumptions regarding the input images similarly to our approach and an open-source implementation is publicly available.
- The SegNet model trained for ear detection (SegNet hereafter). We select this model, because of its state-of-the-art performance on segmentation tasks [44] and the architectural similarities with our model. Thus, we are able to demonstrate the performance of our PED-CED approach and the impact of our architectural design choices through direct comparisons on the same dataset.
- The PED-CED approach with additional connections (PED-CED-alt hereafter), where a fourth shortcut connection is added between the #4 and #24 convolutional layer of the encoder and decoder, respectively. The goal of including this model is to show how different connections affect performance.

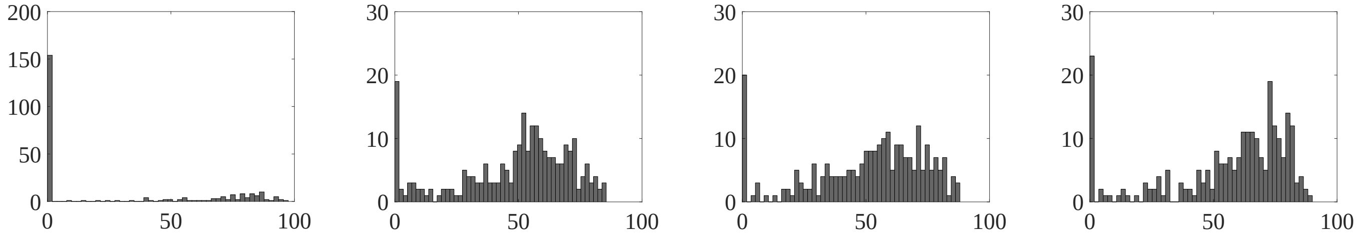
To be able to compare the segmentation techniques (i.e., PED-CED, SegNet and PED-CED-alt) and the detection method (i.e., Haar) on equal footing, we calculate bounding rectangles for our ground-truth annotations and then compute our performance metrics for the Haar-based approach by comparing the (corrected) ground truth rectangles to the bounding boxes returned by the detector. For the segmentation techniques we compare detections to the ground truth at the pixel level, making the comparison stricter for this group of techniques. For the first series of experiments we threshold the softmax output of the three CNN-based approach using the threshold of 0.5, for Haar we use OpenCV's default values.

The results of our assessment are shown in Table 3. While the accuracy measure shows how many pixels are correctly classified in the whole image, it needs to be noted that the ear and non-ear class are not balanced (there is significantly more pixels belonging to the non-ear class (the majority class) than to the ear class (the minority class)), so care needs to be taken when interpreting the presented results. For our test data, the majority class covers 98.9% of all pixels and the minority class covers the remaining 1.1%. This means that a classifier/detector assigning all image pixels over the entire test set would show an overall accuracy of 98.9%. Our PED-CNN detection approach achieves the average accuracy of 99.4% compared to the 98.8% of the Haar-based detector and the 99.2% and 99.2% achieved by SegNet and PED-CED-alt, respectively. The E_2 measure (lower is better) is related to the accuracy but is not affected by the a priori class distribution. As can be seen from Table 3 the proposed PED-CED approach is the best performer in term of E_2 with an average value of 22.2%, while the PED-CED-alt, SegNet and Haar approaches achieve an E_2 value of 24.6%, 25.8% and 36.4%, respectively.

The Intersection over Union (IoU) better reflects the actual performance of the evaluated detectors, and is also not affected by the distribution of pixels among the majority and minority classes. The average IoU for our PED-CED detection approach is 55.7%, whereas

Table 3 Comparison of the proposed PED-CED approach and and competing techniques. 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 the Haar-based approach was evaluated using bounding rectangles, whereas the remaining three segmentation techniques were evaluated more strictly using pixel-wise comparisons.

	Accuracy [%]	IoU [%]	Precision [%]	Recall [%]	E_2 [%]
Haar [25]	98.8 ± 1.1	27.2 ± 36.5	36.7 ± 46.6	28.5 ± 38.4	36.4 ± 18.2
SegNet [44]	99.2 ± 0.6	48.3 ± 23.0	60.8 ± 26.0	75.9 ± 33.1	25.8 ± 11.5
PED-CED-alt (ours)	99.2 ± 0.6	50.8 ± 23.6	62.5 ± 25.9	$\mathbf{78.5 \pm 32.2}$	24.6 ± 11.8
PED-CED (ours)	$\mathbf{99.4 \pm 0.6}$	$\mathbf{55.7 \pm 25.0}$	$\mathbf{67.7 \pm 25.7}$	77.7 ± 32.8	$\mathbf{22.2 \pm 12.5}$



(a) Haar (b) SegNet (c) PED-CED-alt (d) PED-CED

Fig. 5: Histograms for the Intersection-over-Union (IoU) metric for the four evaluated detection approaches. The histograms for the PED-CED approach shows a much better distribution than the Haar-based approach with most of the mass concentrated at the higher IoU values. The two competing segmentation approaches perform better, but PED-CED again exhibits a more favorable distribution.

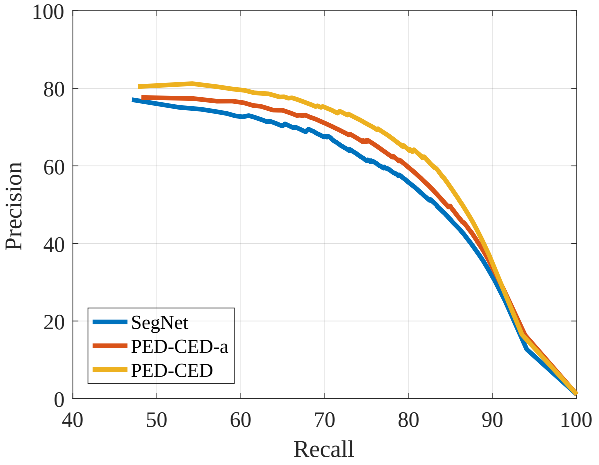


Fig. 6: Precision-recall curves for our experiments. The graphs shows that the proposed PED-CED approach clearly outperforms the considered competing methods.

Table 4 The average time for the PED-CED detection procedure and competing methods on images (of size 480×360 pixels) on our hardware setup.

Approach	Detection time
Haar [25]	178 ms
SegNet [44]	85 ms
PED-CED-alt (ours)	89 ms
PED-CED (ours)	89 ms

the Haar-based detector achieves an average IoU of 27.2%. SegNet and PED-CED-alt perform better, but are still inferior to the proposed PED-CED detector capitalizing on the importance of our design choices for the segmentation model. The high standard deviations can be attributed to some cases where the detector completely misses the ears. This could be improved by using more diverse images during training, as detection errors typically happen on images with bad illumination, extreme viewing angles and in the presence of severe occlusions (see Section 5.3).

The average precision value is 67.7% for the PED-CED, 36.7% for the Haar-based detector, 60.8% for SegNet and 63.5% for the alternative model PED-CED-alt. The high standard deviation in the case of the Haar detector points to a high number of complete detection failures. The ranking of the techniques is similar when recall is considered, but now PED-CED-alt is overall the top performer.

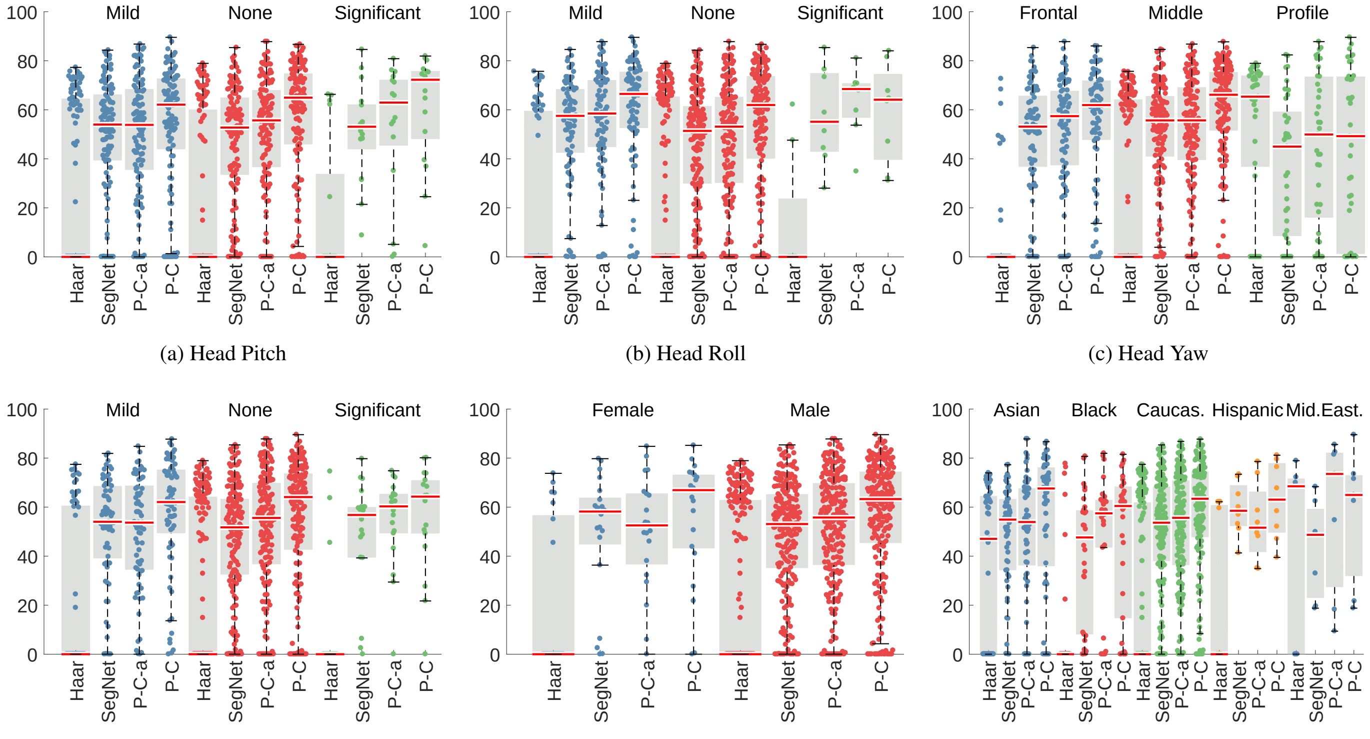
Next to the scalar values in Table 3, we show the complete distribution of (in our opinion) the most informative performance metric, the IoU, in Fig. 5. Although, in many cases our PED-CED approach fails completely, the majority of detections is still between 50% and 90% IoU. The Haar-based detector, on the other hand, exhibits a significant peak at the low IoU values with more than 150 images showing an IoU below 5%. The SegNet and alternative model PED-CED-alt are closer to PED-CED in terms of performance, but still have less mass around the highest IoU values.

In Fig. 6 we show the complete precision-recall curves for all three segmentation-based approaches. Among the tested methods, the proposed PED-CED approach results in the highest precision values at a given value of the recall. All in all, the presented results suggest that the proposed PED-CED approach is a viable option for ear detection (through segmentation) and that our CED architecture ensures efficient detection results.

The average processing time computed over the entire test set for the four tested approaches on images of size 480 × 360 pixels is shown in Table 4. The average time for the PED-CED (and PED-CED-alt) detection procedure is 89 ms, which is comparable to SegNet (85 ms), orders of magnitude faster from what was reported for the HEARD approach in [29] (2.48 s) and faster than the Haar-based detector from [25], which requires 178 ms for one image on average with our configuration that uses separate left and right ear detectors.

5.2 *Impact of Covariates*

Next, we evaluate the impact of various covariate factors on the performance of our PED-CED ear detection approach and competing methods. We use the existing annotations that ship with the AWE dataset and explore the impact of: *i)* head pitch, *ii)* head roll, *iii)* head yaw, *iv)* presence of occlusions, *v)* gender, and *vi)* ethnicity. It needs to be noted that the studied covariates are not necessarily unique to



(a) Head Pitch

(b) Head Roll

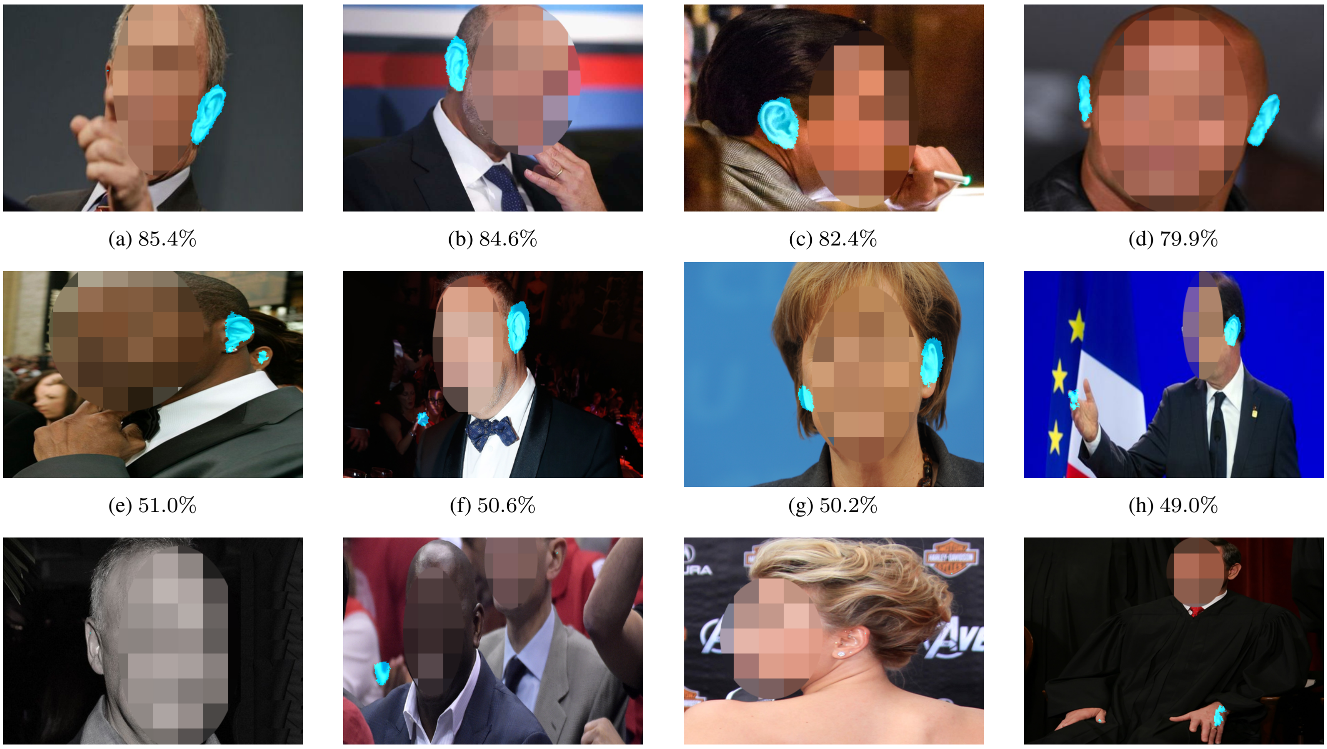
(c) Head Yaw

(d) Occlusions

(e) Gender

(f) Ethnicity

Fig. 7: Impact of various covariates on the performance. The box-plots show the distribution of IoU values computed over the corresponding test set. Actual IoU values (data points) are superimposed over the box-plots and are shown as black dots. The results are plotted in the following order: Haar, SegNet, PED-CED-alt (P-C-a) and PED-CED (P-C). The results show that PED-CED outperforms other approaches over all covariates. Note that median line in red is often at zero with Haar. This is due to the fact that in more than half cases Haar makes no predictions with IoU of more than 0.



(a) 85.4%

(b) 84.6%

(c) 82.4%

(d) 79.9%

(e) 51.0%

(f) 50.6%

(g) 50.2%

(h) 49.0%

(i) 0.1%

(j) 0.1%

(k) 0.1%

(l) 0.0%

Fig. 8: Sample detection results ordered in terms of decreasing values of IoU. The top row shows some of the best detections, the middle row shows average detections and the last row shows some of the worst detection examples. All listed percentage values represent IoU-s. (In the images given here, faces were pixelated in order to guarantee anonymity.)

each image, so effects of covariate cross talk may be present in the results.

We report results in the form of covariate-specific box-plots computed from IoU values in Fig. 7. Here, the actual data points (i.e., IoU values) are also superimposed over the box plots, which is important for the interpretation of the results, as some covariate classes (i.e., significant head roll, Hispanic or middle eastern ethnicity, etc.) contain only a few samples. The presented results show that the median value (red lines in the box-plots) of the IoU is reasonably close for all categories (or labels) of a all covariate factors for the PED-CED approach. No significant performance difference can be observed for any of the studied covariates, which suggests that the proposed PED-CED approach is robust to various sources of variability end ensures stable detection performance across different covariates. Similar observation can be made for the remaining tow segmentation-based techniques (i.e., PED-CED-alt and SegNet), which are also not affected significantly by any of the covariates, but overall exhibit lower IoU values than PED-CED. The Haar based approach, on the other hand, produces usable detections only on profile images and fails completely (with a median UoI value of 0) in the presence of even mild head rotations (in term of pitch, roll or yaw). As seen in Fig. 7(d) and (f), it is also affected considerably by the presence of occlusions and ethnicity.

5.3 Qualitative evaluation

We show a few qualitative examples of our detection results in Fig. 8. The first row of images shows the best detections with IoU values above 80%, the second detections with IoU values around 50%, which is close to the average value achieved on our test set (see Table 3) and the last row shows the worst detection results with IoU values close to 0%. The last row of examples represents complete failures from our test set.

It needs to be noted that the IoU values around 50% (middle row in Fig. 8) are achieved on difficult images with variations across pose, race, occlusion and so forth. These values are also more than sufficient to ensure good detections that can be exploited by fully automatic recognition systems. On the other hand, detections with IoU values close to 0% are of no use to biometric systems and represent cases where our method is in need of improvement. These cases occur with images captured at extreme viewing angles with respect to the ears and in images with limited contrast among others.

In Fig. 9 we show a comparison of the detection results of all considered approaches. The top row depicts images, where all tested methods perform well. The second row shows images where the segmentation-based methods perform well, but the Haar-based approach struggles - it detects only the person's right ear in the first image and misses the ear in the second image completely. In the first image of the bottom row the CNN-based approaches fail, while Haar produces a decent detection output, whereas in the second of the bottom row, Haar does not produce a detection output, while the CNN-based approaches falsely detect an ear on the persons hand.

6 Conclusion

Ear detection in unconstrained conditions is a difficult problem affected by: different angles from which images are taken, various skin color tones, changes illumination conditions, occlusions, accessories. In order to address the problem of ear detection in unconstrained environments successfully we proposed in this work a new ear detection approach, called PED-CED, based on a convolutional encoder-decoder network.

Our experiments showed that PED-CED detects ears with the average accuracy of 99.4%, the average Intersection over Union (IoU) of 55.7%, average precision of 67.7% and the average recall of 77.7%. All of these performance metrics were also shown to be significantly higher than those achieved by competing methods, such as the Haar-based ear detector and the SegNet segmentation network.

Our future work with respect to ear detection includes incorporating contextual information into the detection pipeline, looking for ears only in the vicinity of faces and in specific relation to other facial parts (such, as the nose, eyes, etc.).

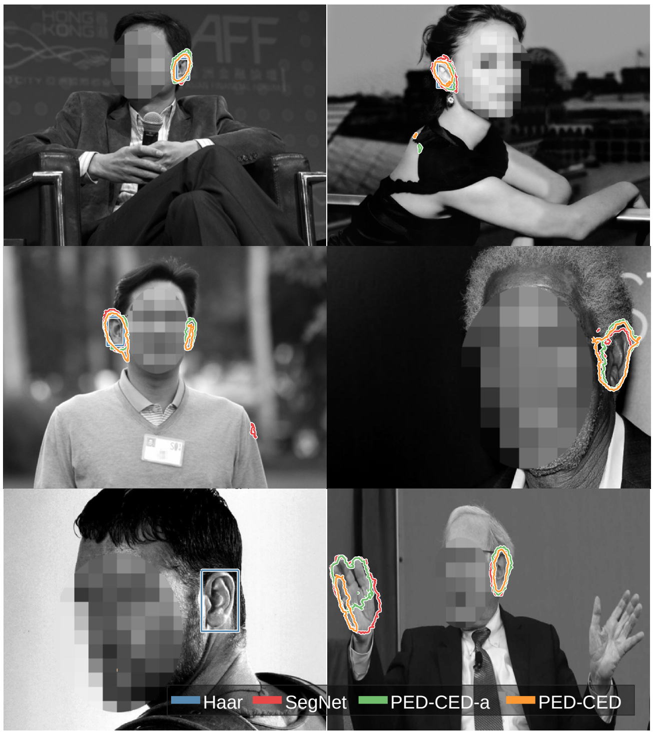


Fig. 9: Sample detections for all four tested approaches. In the images shown here faces were pixelated in order to guarantee anonymity and converted to gray-scale so that the detection/segmentation lines are easier to distinguish. The image is best viewed in colour.

Our long-term plan is to incorporate the presented detection method into a pipeline of ear recognition. A system like that will be able to recognize persons based on ears only by inputting plain images taken in unconstrained environments. Furthermore, considering the speed of the current implementation of the ear detector, the recognition should be able to operate in real-time or close to real-time at the very least.

Acknowledgement

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Program P2-0250 (B) Metrology and Biometric Systems as well as the ARRS Research Program P2-0214 (A) Computer Vision.

7 References

1 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*, 2018, **42**, (1), pp. 11

2 Hansley, E.E., Segundo, M.P., Sarkar, S.: 'Employing Fusion of Learned and Handcrafted Features for Unconstrained Ear Recognition', *arXiv preprint arXiv:171007662*, 2017,

3 Banerjee, S., Chatterjee, A.: 'Robust multimodal multivariate ear recognition using kernel based simultaneous sparse representation', *Engineering Applications of Artificial Intelligence*, 2017, **64**, pp. 340–351

4 Abaza, A., Ross, A., Hebert, C., Harrison, M.A.F., Nixon, M.S.: 'A Survey on Ear Biometrics', *ACM computing surveys (CSUR)*, 2013, **45**, (2), pp. 22

5 Pflug, A., Busch, C.: 'Ear biometrics: a survey of detection, feature extraction and recognition methods', *IET biometrics*, 2012, **1**, (2), pp. 114–129

6 Emeršič, Ž., Štruc, V., Peer, P.: 'Ear Recognition: More Than a Survey', *Neurocomputing*, 2017, **255**, pp. 26–39

7 Emeršič, Ž., Gabriel, L.L., Štruc, V., Peer, P.: 'Pixel-wise Ear Detection with Convolutional Encoder-Decoder Networks', *arXiv preprint arXiv:170200307*, 2017,

8 Badrinarayanan, V., Handa, A., Cipolla, R.: 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling', *arXiv:150507293*, 2015,

9 Badrinarayanan, V., Kendall, A., Cipolla, R.: 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', *arXiv preprint*

arXiv:151100561, 2015,

10 Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: 'Image-to-Image Translation with Conditional Adversarial Networks', *arXiv preprint arXiv:161107004*, 2016,

11 Newell, A., Yang, K., Deng, J. 'Stacked Hourglass Networks for Human Pose Estimation'. In: European Conference on Computer Vision, 2016. pp. 483–499

12 Emeršič, Ž., Štepec, D., Štruc, V., Peer, P., George, A., Ahmad, A., et al.: 'The Unconstrained Ear Recognition Challenge', *International Joint Conference on Biometrics (IJCB)*, 2017,

13 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, 2017. pp. 1–9

14 Prakash, S., Gupta, P.: 'Ear Biometrics in 2D and 3D: Localization and Recognition'. vol. 10. (Springer, 2015)

15 Arbab.Zavar, B., Nixon, M.S. 'On Shape-Mediated Enrolment in Ear Biometrics'. In: International Symposium on Visual Computing, 2007. pp. 549–558

16 University of Notre Dame. 'Face Database', 2015? visited on 2016-05-01. Available from: `http://www.nd.edu/cvrl/CVRL/DataSets.html`

17 Ansari, S., Gupta, P. 'Localization of Ear Using Outer Helix Curve of the Ear'. In: International Conference on Computing: Theory and Applications, 2007. pp. 688–692

18 Attarchi, S., Faez, K., Rafiei, A. 'A New Segmentation Approach for Ear Recognition'. In: International Conference on Advanced Concepts for Intelligent Vision Systems, 2008. pp. 1030–1037

19 Ear Recognition Laboratory at the University of Science & Technology Beijing. 'Introduction to USTB ear image databases', 2002. visited on 2016-05-01. Available from: `http://www1.ustb.edu.cn/resb/en/doc/Imagedb_123_intro_en.pdf`

20 Carreira.Perpinan, M.A. 'Compression neural networks for feature extraction: Application to human recognition from ear images'. Faculty of Informatics, Technical University of Madrid. Spain, 1995

21 Islam, S.M., Bennamoun, M., Davies, R. 'Fast and Fully Automatic Ear Detection Using Cascaded Adaboost'. In: Workshop on Applications of Computer Vision, 2008. pp. 1–6

22 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, 2008. pp. 1–8

23 Prakash, S., Jayaraman, U., Gupta, P. 'Connected Component based Technique for Automatic Ear Detection'. In: International Conference on Image Processing, 2009. pp. 2741–2744

24 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, 2009. pp. 213–216

25 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, 2010. pp. 1–6

26 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, 2010. pp. 1–6

27 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, 1999. pp. 965–966

28 Prakash, S., Gupta, P.: 'An efficient ear localization technique', *Image and Vision Computing*, 2012, **30**, (1), pp. 38 – 50

29 Wahab, N.K.A., Hemayed, E.E., Fayek, M.B. 'HEARD: An automatic human EAR detection technique'. In: International Conference on Engineering and Technology, 2012. pp. 1–7

30 Pflug, A., Winterstein, A., Busch, C. 'Robust Localization of Ears by Feature Level Fusion and Context Information'. In: International Conference on Biometrics, 2013. pp. 1–8

31 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*, 2014, **27**, pp. 115–128

32 Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: 'The FERET database and evaluation procedure for face-recognition algorithms', *Image and vision computing*, 1998, **16**, (5), pp. 295–306

33 Chidananda, P., Srinivas, P., Manikantan, K., Ramachandran, S.: 'Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization', *Machine Vision and Applications*, 2015, **26**, (2), pp. 185–203

34 University of Sheffield. 'The Sheffield (previously UMIST) Face Database', 1998. visited on 2016-05-01. Available from: `https://www.sheffield.ac.uk/eee/research/iel/research/face`

35 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, 2016. pp. 1–12

36 Viola, P., Jones, M. 'Rapid Object Detection Using a Boosted Cascade of Simple Features'. In: Conference on Computer Vision and Pattern Recognition, 2001. pp. I–511

37 Peer, P. 'CVL Face Database', 2005. visited on 2016-05-01. Available from: `http://www.lrv.fri.uni-lj.si/facedb.html`

38 Sim, T., Baker, S., Bsat, M. 'The CMU Pose, Illumination, and Expression (PIE) Database'. In: International Conference on Automatic Face and Gesture Recognition, 2002. pp. 53–58

39 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, 2004.

40 Thomaz, C.E., Giraldi, G.A.: 'A new ranking method for principal components analysis and its application to face image analysis', *Image and Vision Computing*, 2010, **28**, (6), pp. 902–913

41 Mao, X., Shen, C., Yang, Y.B. 'Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections'. In: Advances in Neural Information Processing Systems, 2016. pp. 2802–2810

42 Liu, Z., Hu, Y., Xu, H., Nasser, L., Coquet, P., Boudier, T., et al. 'NucleiNet: A Convolutional Encoder-decoder Network for Bio-image Denoising'. In: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, 2017. pp. 1986–1989

43 Yang, J., Price, B., Cohen, S., Lee, H., Yang, M.H. 'Object Contour Detection with a Fully Convolutional Encoder-Decoder Network'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 193–202

44 Badrinarayanan, V., Kendall, A., Cipolla, R.: 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', *IEEE transactions on pattern analysis and machine intelligence*, 2017,

45 Simonyan, K., Zisserman, A.: 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv preprint arXiv:14091556*, 2014,

46 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al.: 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision*, 2015, **115**, (3), pp. 211–252. Available from: `https://doi.org/10.1007/s11263-015-0816-y`

47 Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A. 'Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning'. In: AAAI, 2017. pp. 4278–4284

48 He, K., Zhang, X., Ren, S., Sun, J. 'Deep Residual Learning for Image Recognition'. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. pp. 770–778

49 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, 2016.

50 Emeršič, Ž., Peer, P., Dimitrovski, I. 'Assessment of Predictive Clustering Trees on 2D-Image-Based Ear Recognition'. In: International Electrotechnical and Computer Science Conference, 2016.

51 Szkuta, B.R., Sanabria, L.A., Dillon, T.S.: 'Electricity Price Short-Term Forecasting Using Artificial Neural Networks', *IEEE transactions on power systems*, 1999, **14**, (3), pp. 851–857

52 Jacobs, R.A.: 'Increased Rates of Convergence Through Learning Rate Adaptation', *Neural networks*, 1988, **1**, (4), pp. 295–307

53 Moody, J., Hanson, S., Krogh, A., Hertz, J.A.: 'A Simple Weight Decay Can Improve Generalization', *Advances in neural information processing systems*, 1995, **4**, pp. 950–957

54 Eigen, D., Fergus, R. 'Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture'. In: ICCV, 2015. pp. 2650–2658

55 Proença, H., Alexandre, L.A. 'The NICE.I: Noisy Iris Challenge Evaluation - Part I'. In: Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on, 2007. pp. 1–4

56 Lienhart, R., Maydt, J. 'An Extended Set of Haar-like Features for Rapid Object Detection'. In: International Conference on Image Processing, 2002. pp. I–900