Received: 3 October 2022 | Revised: 20 December 2022 | Accepted: 4 February 2023

IET Biometrics

DOI: 10.1049/bme2.12109

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Efficient ear alignment using a two-stack hourglass network

Anja Hrovatič[1] | Peter Peer[1] | Vitomir Štruc[2] | Žiga Emeršič[1]

[1]Computer Vision Lab, Faculty of Computer and Information Science University of Ljubljana, Ljubljana, Slovenia

[2]Laboratory for Machine Intelligence, Faculty of Electrical Engineering University of Ljubljana, Ljubljana, Slovenia

Correspondence
Žiga Emeršič.
Email: ziga.emersic@fri.uni-lj.si

Funding information
ARRS Research Programme P2-0250; ARRS Research Programme P2-0214

Abstract

Ear images have been shown to be a reliable modality for biometric recognition with desirable characteristics, such as high universality, distinctiveness, measurability and permanence. While a considerable amount of research has been directed towards ear recognition techniques, the problem of ear alignment is still under-explored in the open literature. Nonetheless, accurate alignment of ear images, especially in unconstrained acquisition scenarios, where the ear appearance is expected to vary widely due to pose and view point variations, is critical for the performance of all downstream tasks, including ear recognition. Here, the authors address this problem and present a framework for ear alignment that relies on a two-step procedure: (i) automatic landmark detection and (ii) fiducial point alignment. For the first (landmark detection) step, the authors implement and train a Two-Stack Hourglass model (2-SHGNet) capable of accurately predicting 55 landmarks on diverse ear images captured in uncontrolled conditions. For the second (alignment) step, the authors use the Random Sample Consensus (RANSAC) algorithm to align the estimated landmark/fiducial points with a pre-defined ear shape (i.e. a collection of average ear landmark positions). The authors evaluate the proposed framework in comprehensive experiments on the AWEx and ITWE datasets and show that the 2-SHGNet model leads to more accurate landmark predictions than competing state-of-the-art models from the literature. Furthermore, the authors also demonstrate that the alignment step significantly improves recognition accuracy with ear images from unconstrained environments compared to unaligned imagery.

KEYWORDS
convolutional neural nets, ear biometrics

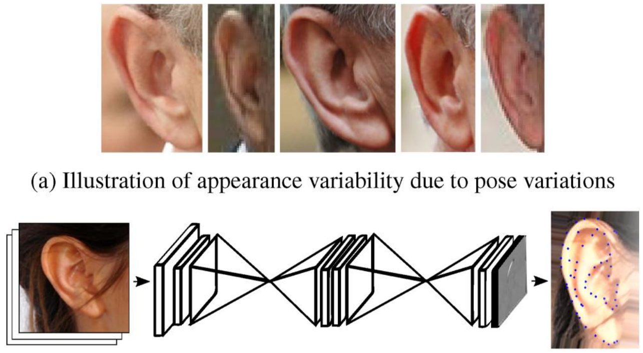
1 | INTRODUCTION

Ear recognition techniques have seen considerable improvements over the years and consequently contributed towards increased interest in automated ear recognition systems [1, 2]. While many powerful recognition approaches, mostly based on deep learning, have been proposed in the literature recently, the problem of ear alignment has received comparably less attention, as also emphasised in recent surveys in this field. [2, 3], Despite its importance and (potentially beneficial) impact on all downstream tasks, efficient ear alignment in diverse settings is still largely unsolved.

In general, the problem of ear alignment emerges in loosely constrained acquisition scenarios, where ear images are commonly captured under various head orientations and poses. Despite the use of ear detection/segmentation procedures, these acquisition scenarios typically lead to significant pose variability in the final ear images, as illustrated in Figure 1a. Such pose variability introduces noticeable differences in the ear appearance and has a considerable impact on all components of ear recognition systems, including the representation-calculation step and template comparison procedure. Minimising errors induced by poorly aligned ear images is, therefore, critical for recognition systems relying on ear biometrics.

There are several issues that make landmark detection with ear images particularly challenging, that is,

- When captured in unconstrained settings, ear images exhibit a considerable level of appearance variability caused by pose



(a) Illustration of appearance variability due to pose variations

(b) High-level overview of the proposed landmarking model

FIGURE 1 This paper introduces a novel Two-Stack Hourglass Network (2-SHGNet) for landmark detection in ear images and integrates it with an alignment procedure to minimise the impact of misalignment on ear recognition performance.

variations, occlusions by hair and other accessories, illumination changes and other similar factors.
- The ears typically cover only a small portion of the image, leading to input samples for the landmarking procedure that are of mostly low-resolution.
- The cartilage structures of the ear that are shared among individuals (e.g., the helix, antihelix etc.) correspond to smooth curvature in the ear images with indistinct appearance, which makes it difficult to identify common anchor points (landmarks) for alignment across different subjects.

The outlined challenges and importance of alignment for ear recognition provide a strong motivation for research into ear landmarking models that are capable of providing reliable results with diverse input images captured in unconstrained settings. Such landmarking models can provide information on point-to-point correspondences across ear images and be used with standard registration schemes to align the images in accordance with some pre-defined canonical shape. While existing solutions in this area were shown to yield competitive results, for example, Refs. [4, 5], robust landmarking across ear images with challenging characteristics with respect to pose, resolution, head orientation, occlusion and other similar nuisance factors, induced by the unconstrained nature of the acquisition procedure, still represents an open problem.

To address this gap, we present in this paper a novel landmarking model tailored towards the task of ear alignment, as illustrated in Figure 1b. At the core of the solution is a Two-Stack Hourglass Network (2-SHGNet), a convolutional neural network (CNN), capable of identifying a markup of 55 landmarks of the ear. The landmarks generated by 2-SHGNet are then used with an alignment procedure based on the Random Sample Consensus (RANSAC) algorithm to align the ear images with a canonical shape template of the ear. We evaluate the 2-SHGNet model on two publicly available datasets, that is, ITWE (In-The-Wild Ear) [4] and AWEx (Extended Annotated Web Ears) [2, 6] dataset, and compare it to several state-of-the-art landmark detection methods from the literature. To further demonstrate, the merits of ear alignment with our approach, we also evaluate the impact of alignment on the performance of various recent ear recognition techniques. The experimental results show that the proposed model yields highly accurate landmark detection results with diverse input images and that the alignment if beneficial for recognition performance of all tested models.

A summary of the key contributions of this paper is given in the following bulleted list:

- *A novel framework landmark-based ear alignment*: We describe a deep learning model, abbreviated 2-SHGNet, for 2D landmark detection in ear images. 2-SHGNet consists of two stacked hourglass models that sequentially process the input image and extract landmark locations that serve as the basis for ear alignment. As we elaborate in the methodology section, the 2-SHGNet model is able to capture and consolidate information across different image scales, resulting in a powerful and reliable ear landmarking procedure that is used as the basis for ear alignment with a RANSAC-based alignment step.
- *A comprehensive analysis*: We illustrate the performance and characteristics of the proposed ear alignment framework through quantitative as well as qualitative results and explore the impact of ear alignment on ear recognition performance.
- *Publicly available source code*: We make all source code, including the model definitions, weights and training scripts freely available to the research community to foster reproducibility. The source code is available from https://github.com/Anjdroid/ear_alignment_stacked_hourglass.

The rest of this paper is structured as follows: in Section 2 we describe related work on ear alignment as well as ear and face based landmark detection. In Section 3 we present the methodology and in Section 4 we present the experimental evaluation of the proposed alignment framework. Finally, we conclude the paper with directions for future work and some closing remarks in Section 6.

2 | RELATED WORK

In this section, we review existing work on ear alignment and landmarking as well as literature on alignment techniques developed for facial images. The goal of this section is to provide factual background and context for the proposed alignment framework.

2.1 | Ear alignment

Early work on ear alignment focussed mostly on the correction of in-plane rotations, for example, [1, 7, 8], where the global oval shape of the ears was exploited to align different samples. While these procedures produced reasonable results for high-quality images captured in (semi-) controlled conditions, they were found to be less suitable for more challenging data acquired in-the-wild, where the assumption about the oval shape may not apply well due to occlusions and in-place rotations of the ear regions. To address such issues, landmark-based approaches have been presented in the literature. Zhou et al. [4], for example, first presented a dataset of 2D ear images captured in the wild with annotated landmarks and then explored the feasibility of state-of-the-art methodologies for 2D and 3D landmark detection. Specifically, they investigated the performance of the Supervised Descent Method (SDM) [9], Constrained Local Models (CLMs) [10] and Active Appearance Models (AAMs) [11]. The authors built two different kinds of AAMs, holistic and patch-based, obtaining different deformation models and appearance representations. Although they reported impressive landmarking performance with holistic and patch-based AAMs in predicting 2D landmarks, issues with detecting landmarks on ears with higher degrees of pose variations (in terms of pitch, roll or yaw angles) still remained. Another line of research in ear landmarking and alignment is looking at deep learning methodologies. In Refs. [5, 12], the authors explored the use of CNN-based features for automatic 2D ear landmark detection in unconstrained imaging scenarios and studied the impact of ear normalisation (i.e. alignment) on the performance of several different recognition methods. In these works, the detected landmarks acted as the basis for geometric ear normalisation with the use of Principal Component Analysis (PCA). While less relevant to our research, 3D ear landmark detection has also been addressed within 3D Point Clouds (PTC) in Ref. [13, 14]. The idea here was to extend the existing state-of-the-art 2D landmark localization algorithms to 3D.

In this work, we built on the advances outlined above and present a solution for ear alignment that follows the landmark-based framework and combines the landmarking procedure with a RANSAC-based alignment step. As we show in the experimental section, the proposed approach leads to highly competitive landmarking performance when compared to the current state-of-the art, while the overall alignment procedure helps to improve recognition accuracy.

2.2 | Landmark-based alignment with facial images

Landmark detection is a common step towards alignment with other biometric modalities as well. Especially with facial images, landmarking techniques have been very popular, as evidenced by the tremendous amount of work done in this area [15–19]. Here, the literature has long been focussed on Constrained Local Models (CLMs) [10] and Active Appearance Models (AAMs) [11]. More recently, however, deep learning solutions started dominating this task due to their superior characteristics. In Ref. [20], Chen et al. presented a kernel density deep neural network for face alignment and reported competitive results. The authors of Ref. [21] presented a boundary-aware face alignment method using stacked dense U-Nets. They employed dual transformers to make the stacked dense U-Nets spatially invariant and reported improved performance of facial landmark detection on unconstrained data. Yang et al. [22] proposed a deep CNN, called Stacked Hour-Glass Network that followed the idea of Cascaded Shape Regression (CSR) [23] by refining landmark predictions over a cascade of regression models. The authors performed extensive experiments on several challenging face datasets to validate their model. Despite the progress in facial landmarking presented above, landmark prediction in facial images in still a challenging problem, where encoder–decoder solutions (based, for example, on U-Nets) have also shown significant promise recently [24–26].

Motivated by the impressive performance of facial landmarking techniques and particularly the success of stacked/cascaded models, such as Refs. [22, 27–29], we describe in this paper a Two-Stack Hourglass Network capable of locating landmarks in ear images through a coarse-to-fine landmarking strategy, which results in highly competitive performance and represents the basis for ear alignment in our experiments.

2.3 | Alternatives to landmark-based alignment in biometrics

Recent work on aligning regions-of-interest (ROIs) in biometrics has explored alternative solutions to landmark-based alignment that often times include end-to-end-models. Matkowski et al. [30], for example, described a ROI-alignment module (CNN) capable of registering palmprint regions from images, captured in unconstrained scenarios, to a pre-defined shape. Yin et al. [31] proposed a generative face frontalisation approach, capable of synthesising frontal faces from off-pose headshots, while also ensuring alignment of salient facial feature points. Similar solutions were also presented in Refs. [32–34]. Reddy et al. [35] proposed a ROI detector for periocular images that normalised the geometric characteristics of the data to a pre-defined form. The detector used a spatial transformer network and was again learnt end-to-end without intermediate landmarks.

While the alignment procedures describe above provide for efficient and well-performing registrations/alignment schemes, they are usually also computationally expensive and often need to generate complete output images. Landmark-based approaches, on the other hand, need to only output sparse (i.e. a limited number of coordinates) landmark points, allowing for lighter parameterisation of the utilised models. While evaluated in this paper in conjunction with a RANSAC-based alignment procedure, landmarking models are also useful for other tasks beyond alignment. Other application scenarios include 3D ear morphing, ear-shape/orientation based head-pose estimation, ear reconstruction or extraction of ear sub-regions—similar to what has been done with facial images [36–38]. These characteristics also apply to the Two-Stack Hourglass Network, described in this work.

3 | METHODOLOGY

Ear alignment has been an integral pre-processing step of early ear recognition techniques [39, 40]. However, with the shift to unconstrained acquisition settings and the emergence of deep learning models, ear recognition has been increasingly approached within end-to-end solutions that tried to explicitly capture the variability caused by unaligned ear images instead of introducing separate (potentially error-prone) normalisation procedures [2, 41–43]. Nonetheless, recent studies related to various visual recognition tasks, including ear recognition [4], have shown that prior alignment can still contribute to better overall recognition performance. Inspired by these observations, we study in this paper the task of ear alignment and its impact on ear recognition performance.

Specifically, we experiment with the overall framework illustrated in Figure 2, which given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ first predicts a set of n landmarks $\hat{\mathbf{x}} = \left[\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \ldots, \hat{x}_n, \hat{y}_n\right]^T \in \mathbb{R}^{2n}$ that jointly define the shape of the ear in X and then normalises the ear image in accordance with a pre-defined canonical shape x_S. The main component of this framework is a powerful landmarking model, called 2-Stack Hourglass Network (or 2-SHGNet for short), trained using a heatmap regression loss. Details on the framework and the 2-SHGNet model are given in the following sections.

3.1 | Landmark detection and learning objective

Our framework leverages a stacked hourglass network architecture to predict landmark positions in ear images through heatmap regression. Formally, the network ϕ accepts an ear image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ as input and outputs $n = 55$ so-called heatmaps arranged into the tensor $\hat{\mathbf{Y}} \in \mathbf{R}^{H' \times W' \times n}$. Here, each of the n channels $\hat{\mathbf{Y}}_i$ in $\hat{\mathbf{Y}}$ acts as a probability map, where the location of the largest value defines the ith landmark location, that is,

$$\left(\hat{x}_i, \hat{y}_i\right) = argmax_{(x,y)}\left(\hat{\mathbf{Y}}_i\right), \tag{1}$$

where $i \in \{1, 2, \ldots, n\}$. Once the above expression is evaluated over all n channels, the n landmark locations $\left(\hat{x}_i, \hat{y}_i\right)$ jointly define the predicted overall ear shape $\hat{\mathbf{x}}$.

To learn the model, we minimise the following Mean Square Error (MSE) based learning objective over a given (and labelled) dataset:

$$\mathscr{L}_{mse} = \|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2 = \|\phi(\mathbf{X}) - \mathbf{Y}\|_2^2, \tag{2}$$

where each channel Y_i of the reference heatmaps in Y is constructed by evaluating a Gaussian centred at the ith ground truth location (x_i, y_i) over an $H' \times W'$ lattice, that is,

$$\mathbf{Y}_i = \frac{1}{\sqrt{2\pi\sigma}} \exp \ - \frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}, \tag{3}$$

where (x, y) denote the spatial coordinates in $\mathbf{Y}_i \in \mathbb{R}^{H' \times W'}$ and σ defines the shape of the Gaussian and is set to $\sigma = 1$ to ensure well-localised reference heatmaps for the training procedure. The above learning process is referred to as heatmap regression in the open literature and has been applied successfully to different problems, ranging from face alignment to human pose estimation [5, 20–22].

3.2 | Model architecture: the 2-Stack Hourglass Model

Stacked deep learning models have been reported to be highly efficient in solving human pose estimation and face alignment problems [20–22]. Such stacked models are capable of processing input data through a sequence of simpler models, where each next model provides additional capabilities to the overall sequence, resulting in powerful computational architectures. If implemented with suitable base models (e.g., Hourglass models, U-nets), the stacked topology can also ensure multi-scale processing. Inspired by the success of such models in various problem areas, we implement and train a 2-Stack Hourglass Model (2-SHGNet) for ear landmarking in this work, as shown in Figure 3. The model consists of a backbone feature extractor and two hourglass modules stacked one after the other.

The goal of the *backbone feature extractor* is to process the given ear image and compute information-rich representations from the input data that can be used later by the hourglass modules in the heatmap regression task. The backbone model is a simple stack of convolutional layers, where each convolutional layer is followed by batch normalisation,

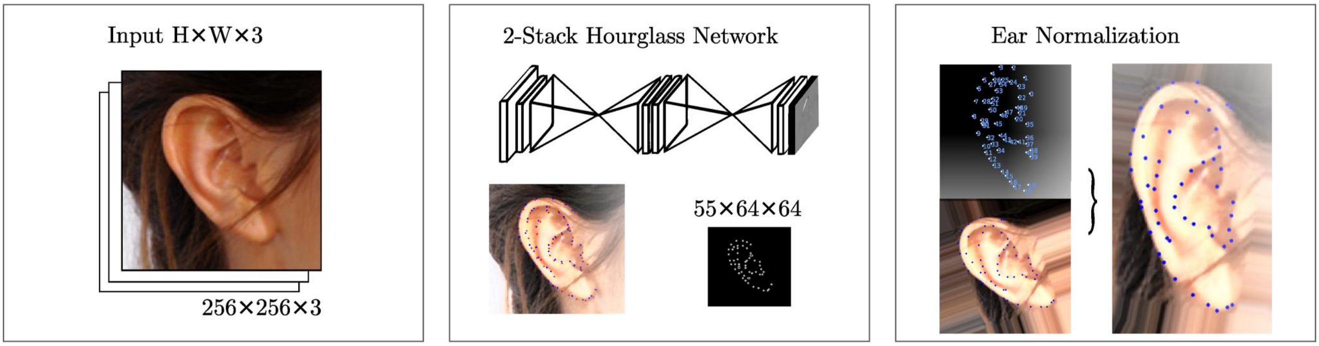


FIGURE 2 Visualisation of the ear alignment approach developed in this work. The key component of the approach is a 2-Stack Hourglass model that located 49 ear landmarks in the input image. These landmarks are then used in the geometric normalisation step to align the overall ear shape with a pre-defined ear-shape template using a RANSAC-based alignment procedure.

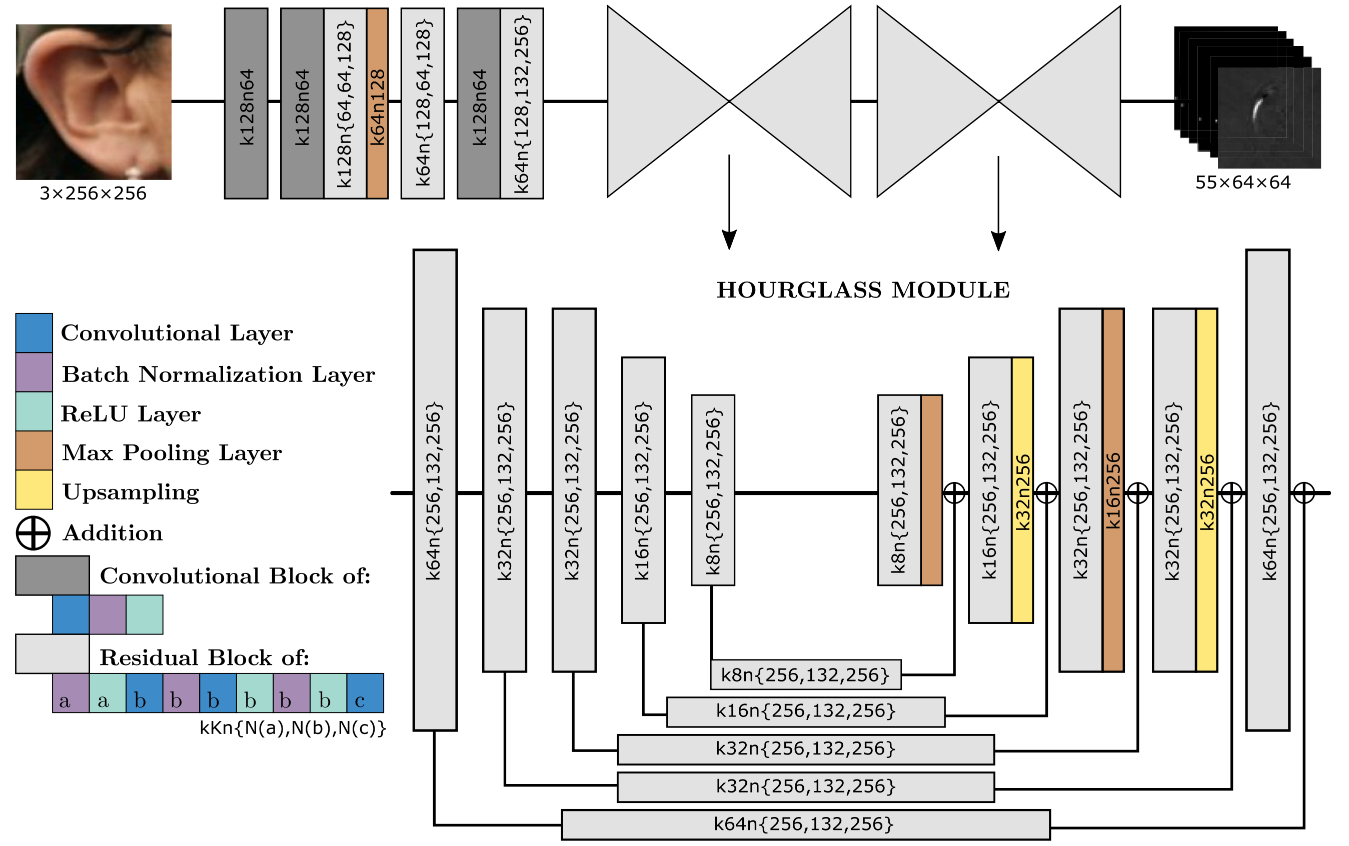


FIGURE 3 High-level overview of the 2-Stack Hourglass Network architecture. The model consists of a backbone feature extractor and two stacked hourglass modules that jointly generate the heatmaps needed for landmark prediction. The kKnN notation is used in the figure to denote a convolutional layer with N convolutional filters with $K \times K$ support.

aimed at reducing the internal covariate shift, and a ReLU activation. The extracted representations are then fed to the stacked hourglass modules.

The hourglass module (HGNet hereafter) has an encoder–decoder structure and resembles an hourglass shape, hence the name. The encoder–decoder structure allows for the manipulation (and/or translation) of the input data by first extracting informative features along the layers of the encoder and then decoding the extracted information into the desired form. The HGNet structure has a symmetric topology and (similarly as U-Net based models) consolidates information across different resolutions by employing down-sampling and up-sampling operations, which makes it possible to efficiently explore the relationships between landmarks at different scales.

As shown in the bottom right of Figure 3, the hourglass module consists of residual blocks (BN + ReLu + 3 × conv.) with interspersed max pooling layers in the encoder and corresponding up-sampling layers (using the nearest neighbour strategy [44]) in the decoder. We use ReLU activation functions in the modules because they enable sparse activations, better gradient propagation, scale-invariance and efficient computations [45]. Furthermore, the max pooling layers help to reduce the input's dimensionality, the number of learnable parameters, as well as the overall computational cost and also provide basic invariance to translations. The residual connections from the encoder to the decoder help to propagate information at various scales from the feature extraction side to the decoder side and facilitate the concatenation of low-level features with higher-level representations with semantic-awareness.

We use a stack of two hourglass modules to implement the final *two-stack hourglass landmarking network*, as presented in Figure 3. The implementation of the model was adapted from Ref. [46] and utilises the PyTorch [47] deep learning framework for CNNs. The output of the network is a tensor $\hat{\mathbf{Y}}$ with $n = 55$ heatmaps, each encoding one landmark location on the ear through the maximum response in the heatmap. A few illustrative examples of such heatmaps computed for a test image are shown in Figure 4. It can be seen how the heatmaps corresponding to landmark locations on distinct ear structures of the ear are well localised, while the heatmaps on the ear outline (i.e. on the helix) are spread out over a larger area, as expected.

It needs to be noted that we initially tried to solve the problem of 2D ear landmark detection with the use of a single HGNet module, which, however, proved to be less capable in detecting landmark locations in terms of the localisation accuracy. Stacking two HGNets, on the other hand, drastically improved the accuracy of landmark detection, while still being able to perform in real time. By stacking the HGNets we enable the 2-SHGNet to gradually fine-tune the landmark

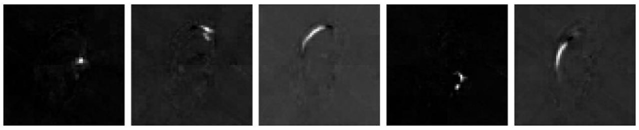


FIGURE 4 Illustrative examples of the heatmaps generated by the 2-SHGNet model for five randomly selected landmarks estimated for a test image. Note how the heatmaps are well localised for landmarks located at distinct ear structures, while they are spread out over a wider area for less distinct positions/structures, such as the helix.

positions through stacks, since the information and the context constraints about ear landmarks are enhanced as the number of hourglass modules is increased to two [48].

3.3 | Ear normalisation

Assume that the n landmark locations $\hat{\mathbf{x}}$ have been predicted for a given input ear image X. In the last step of the proposed framework, we align the image X with a pre-defined target shape x_S, as shown in the right part of Figure 2. To this end, we estimate the parameters of an affine geometric transform $\mathbf{T} \in \mathbb{R}^{3\times 3}$ and apply it to the input image X:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \mathbf{T} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{4}$$

where (x, y) and (x', y') are the coordinates of the initial input X and aligned ear image X′, respectively.

Because not all landmark locations may be perfectly accurate, we aim to improve the robustness of the alignment step and utilise the Random Sample Consensus algorithm (RANSAC) [49] when estimating the parameters of T. RANSAC represents the de facto standard for estimating the parameters of geometric transforms in the presence of outliers and is therefore also used in this work.

4 | EXPERIMENTAL SETUP

In this section, we present the experimental setup used to evaluate the 2-SHGNet model and overall alignment framework. Specifically, we discuss the datasets selected for the experiments, the pre-processing procedure, training details and the performance measures utilised for scoring.

4.1 | Datasets and experimental protocol

For the experimental evaluation, two diverse datasets are used, that is, ITWE [4] and AWEx [6]. The selected datasets contain images gathered from the Internet and feature ears captured in completely unconstrained environments, as seen from the examples in Figure 5. The ITWE dataset is, to the best of our knowledge, the only publicly available ear dataset annotated with landmark annotations and is, therefore, used for training of the 2-SHGNet model and evaluation of the landmark-fitting performance. The AWEx dataset, on the other hand, is used in recognition experiments to evaluate the impact of ear alignment on recognition performance. Details on both datasets are given below.

4.1.1 | The ITWE dataset

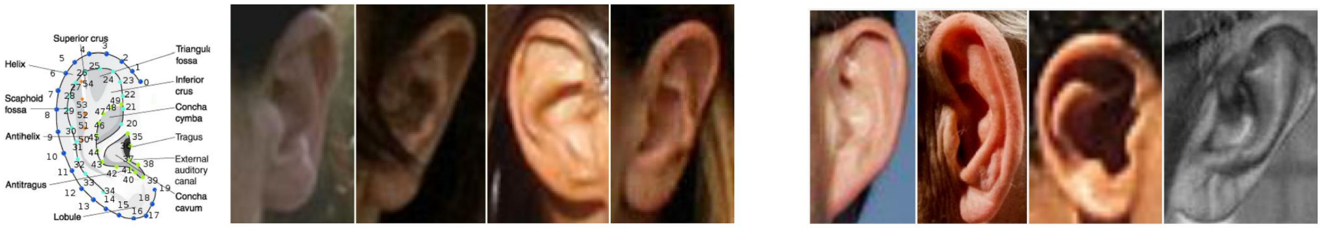
The ITWE dataset [4] (Collection A) consists of 605 ear images captured *in-the-wild* with each image representing one subject. All images are annotated with 55 landmarks in accordance with the markup scheme on the left side of Figure 5a. As can be seen, the 55 fiducial points are arranged along the key morphological structures of the ear. The images in ITWE dataset were collected from the web by querying Google's image search service with ear-related tags. For the experiments, we split the dataset into a training and testing part, where 500 images with augmentation degree of 35 are used for learning the landmarking model and 105 images of equal augmentation degree are used for performance assessment—following Ref. [4].

4.1.2 | The AWEx dataset

The AWEx dataset [2, 6] consists of 4104 images of 346 and, similarly to ITWE, was also collected from the web. The dataset exhibits considerable variability across viewing angles, occlusions, acquisition conditions (indoor and outdoor), image quality and resolution. Because the AWEx dataset does not come with annotated landmarks, we measure the impact of the landmarking procedure through recognition experiments, where (following the setup from Ref. [6]) 3104 images of 246 subjects are used for training and the remaining (disjoint) set of 1000 images of 100 subjects for testing.

4.2 | Data pre-processing

To ensure a common starting point for model training and testing, all images are subjected to a pre-processing procedure prior to the experiments. The first step of this pre-processing procedure crops the ear images based on the landmark-defined bounding box and then randomly pads the cropped region in all four image directions with the goal of introducing translational variability. In the second step, the images are resized to a fixed (and pre-defined) size of 256 × 256 pixels. The ground truth landmark coordinates are also modified to account for the cropping, padding and scale changes. Because the two experimental datasets were gathered from the Internet, the ear images exhibit very different characteristics. A normalisation procedure is therefore used next, which (i) first equalises the histograms of the ear images, (ii) then rescales the pixel values to the range of $0 - 1$, and (iii) finally subtracts the dataset's mean value from each sample image to centre the data. Such



(a) ITWE [4]

(b) AWEx [6]

FIGURE 5 Example images from the two datasets used in this work. The ITWE datasets comes with landmark annotations (see the left part of (a)) and is used for training and testing of the landmarking models. The AWEx dataset has multiple images per subjects and is used to measure the impact of ear alignment on recognition performance. Both datasets were captured in unconstrained settings and therefore exhibit considerable appearance (and pose) variability.

range normalisation is performed on the heatmaps as well, similar to Ref. [4].

4.3 | Performance measures

Following established evaluation methodology, for example, Ref. [4], we use the point-to-point error normalised by the diagonal of the ground truth ear bounding box w to evaluate the performance of the landmarking model. The normalised point-to-point (PPE) error is formally defined as follows:

$$\mathrm{PPE} = \frac{1}{n}\frac{\sum_{i=0}^{n}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{w}, \quad (5)$$

where n is again the number of estimated landmarks. The error takes a value of 0 in the ideal case when the predicted landmarks equal the annotated ground truth for a given test image. In general, lower values of PPE imply better performance. In addition to the normalised point-to-point error, we also report Cumulative Error Distribution (CED) curves, again in accordance with standard evaluation methodology [4, 50].

For the ear recognition experiments, we use the evaluation protocol proposed by Emeršič et al. [2] and report results in terms of the Rank-1 and Rank-5 recognition performance defined in Equations (6) and (7), respectively. The Rank-1 recognition rate measures the proportion of predictions, which match the ground truth label and the Rank-5 recognition rate the proportion of top five predictions that match the ground truth label. The two recognition rates are defined as follows:

$$\text{Rank - 1} = \frac{\#Correct\ predictions}{\#Predictions\ made}, \quad (6)$$

$$\text{Rank - 5} = \frac{\#Correct\ predictions\ among\ top\ five}{\#Predictions\ made}. \quad (7)$$

We also compute complete Cumulative Match Score Characteristic (CMC) curves and report the Area Under the normalised CMC curves (AUCMC) to measure the ranking capabilities of the tested ear recognition models.

4.4 | Implementation details and model complexity

To train the proposed model, the Adaptive Moment Estimation (Adam) [51] optimisation algorithm was used, which is suggested as a default optimisation technique for training deep models as it achieves good results fast with little to no tuning of parameters [52]. Adam is able to perform well on problems with sparse gradients and on noisy data, which provided large datasets and large models in terms of trainable parameters [51]. A small learning rate ($3e-6$) was selected to ensure slow learning and a weight decay of $1e-5$ to penalise large weights and to improve the network's performance. A summary of the Adam configuration parameters is given in Table 1.

A data augmentation process was used to avoid overfitting. This included random scaling between 80% and 120% of the original image height and width, rotations between −45 and 45° and a shearing factor in the range of −16 to 16, and filling newly created pixels with edge values, so as not to introduce new structures into the image. With this procedure, we generated a fixed-size training dataset of 18,000 ear images.

The models were trained on a GeForce RTX 2070 SUPER GPU using the CUDA Toolkit 10.0 with 8 GiB GDDR6 memory. We employed GPU-based training since it is most suitable and optimised for training deep model architectures with large amounts of data. In the final implementation, the 2-SHGNet model has 9,757,068 trainable parameters. The model is able to perform in real time, needing on average 0.45 s to perform landmark prediction on a single image using the above specified hardware.

5 | EXPERIMENTS AND RESULTS

In this section, we present experimental results with the aim to (i) benchmark the performance of the 2-SHGNet model against state-of-the-art competitors and demonstrate the impact of some design choices made, (ii) analyse the capabilities of the model in a qualitative manner, (iii) explore the limitations of 2-SHGNet, (iv) study the feasibility of the geometric alignment/normalisation step, and (v) explore the impact of ear alignment of ear recognition performance.

5.1 | Evaluation of landmarking performance

We first evaluate 2-SHGNet in landmarking experiments on the test part of the ITWE dataset, but use augmentation techniques (similarly to the procedure described above for the training data) to generate a larger test set consisting of 3,780 test images. To put the performance of the model into perspective, we also implement a number of competing techniques that can broadly be grouped into two categories, that is:

- *Model-validation Techniques*. The first group of techniques is meant to validate the characteristics of 2-SHGNet. We implement three landmarking techniques for this purpose, that is:
    - *Baseline*: We use the average of all ground truth annotations over the training images of ITWE as the prediction of the landmarks on the test images. This approach provides an estimate of the baseline landmarking performance that is achievable without using a landmarking model.
    - *2-SUNet*: The second technique represents a stack of two U-Net models [53] and is included in the experiments to demonstrate the benefit of using the hour-glass models within the landmarking procedure instead of other alternatives.
    - *3-SUNet*: The third techniques is a stack of three sequential U-Net models and demonstrates the performance of a similar but more complex (3-stack) model design.
- *State-of-the-art Techniques*. The second group of techniques are state-of-the-art competitors for ear-landmark detection. Here, we follow the work from Ref. [4] and compare against the following landmarking approaches:
    - *ZZ Init*: Similar to the *Baseline* solution above, ZZ Init uses a fixed shape with pre-defined landmarks computed from the training data to predict the landmark locations in the test images. The approach is based on the initialisation procedure used by Zhou and Zaferiou in Ref. [4].
    - *AAMs*: We implement a number of Active Appearance Models (AAMs) [11, 54] using dense SIFT [55], HOG [56] and DCNN features [57]. We denote these techniques as SIFT + AAM, HOG + AAM and DCNN + AMM, respectively.
    - *PAAM*: The last baseline from this group is a Patch-based Active Appearance Model (PAAM) [58] implemented with dense SIFT features, that is, SIFT + PAAM. This model has been shown to yield highly competitive results for face alignment and is therefore also considered here.

We present the cumulative error distributions generated based on the normalised point-to-point errors in Figure 6 and report quantitative performance scores in Table 2. Note that the fraction of test images that achieve a normalised point-to-point error of less or equal to 0.10 is reported separately, similarly to established literature. As can be seen, the proposed 2-SHGNet model achieves an average PPE score of 0.0519 and outperforms all other tested methods. Additionally, it also exhibits the most consistent performance across the test images, as evidenced by the standard deviation of 0.0173, which again is the lowest among all evaluated landmarking models. When looking at the fraction of images with a PPE below 0.10, we observe that 2-SHGNet ensures that 99% of all test images have a lower PPE score than the threshold value, the highest percentage among all competitors.

When looking at the comparison with the model-validation techniques, we observe that the proposed 2-SHGNet significantly outperforms the Baseline solution, suggesting that considerable performance gains can be expected when using a trained landmarking approach, as opposed to a pre-defined ear shape. Additionally, we see that the hourglass backbone consistently results in better performance than the U-Net models, regardless of whether two or three such models are stacked one after the other. The comparison in the lower part of Table 2 shows that all considered state-of-the-art models

TABLE 1 Adam configuration parameter values and descriptions.

Configuration parameters	Value	Description
Learning rate	3*e* − 8	Speed of learning
β_1	0.9	Exponential decay rate for the first moment estimates.
β_2	0.999	Exponential decay rate for the second moment estimates.
ϵ	1*e* − 8	Small constant to prevent division by zero.
Weight decay	1*e* − 5	Learning rate decay or L2 penalty.

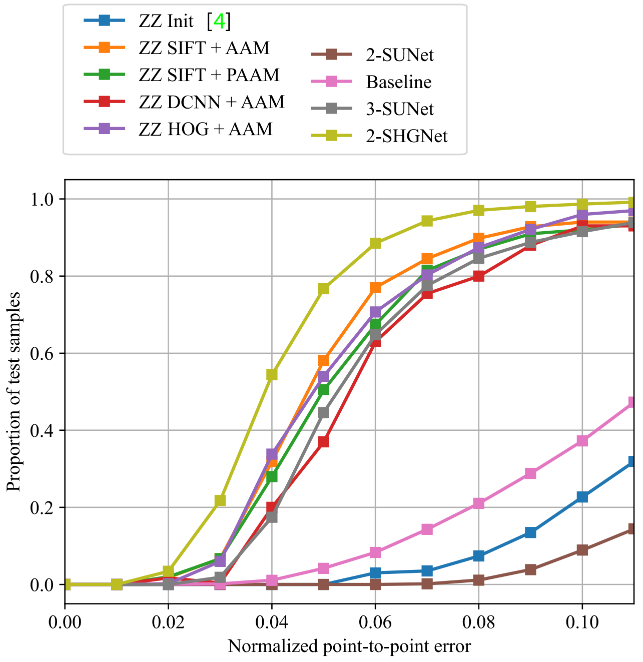


FIGURE 6 Cumulative (point-to-point) error distributions for the proposed 2-SHGNet model and the considered competitors. The curves were generated on the test part of the ITWE dataset.

yield very competitive results. Nonetheless, 2-SHGNet still outperforms the best performing competitor (HOG + AAM) by around 2% in terms of the fraction of images below the PPE threshold value of 0.10 and yields an average PPE error that is 3.85% lower than the average error of the HOG + AAM runner-up.

5.2 | Qualitative landmark detection results

Next, we present qualitative examples of the landmarking performance of the 2-SHGNet model. To provide an idea of how specific PPE scores (from different bins of the CED graphs) translate into (visual) landmark fitting quality, we first show in Figure 7 a few illustrative landmarking results with different PPE values. Here, the annotated reference shape is shown in red and the predicted landmarks are shown in blue. Observe the obvious differences in the accuracy of the landmark alignment between the presented examples.

To further analyse the performance of the 2-SHGNet model, we provide a cross-section of results generated with images from the ITWE and AWEx datasets in Figures 8 and 9, respectively. The figures again show the predicted landmarks (blue) versus the ground truth landmarks (red). It can be seen that the (qualitative) accuracy of predictions on the AWE dataset is comparable to accuracy on the ITWE dataset. The model has the ability to reliably predict landmark locations on colour images of unseen ears in various poses. Moreover, the model is able to predict landmarks on greyscale images, low quality noisy images and images with smaller occlusions of the ears. We observe that the performance of 2-SHGNet is not affected by whether colour information is present or not, as the predicted landmark locations accurately capture the ear and its shape. Moreover, the model also ensures precise landmark prediction and description of with very low quality and noisy images, pointing to its ability to reliably consolidate information from different scales and to learn useful features for ear description. Small partial occlusions of the ear are easily handled by the 2-SHGNet model. However, in cases where earrings or a considerable amount of hair is occluding the ear shape, the model is unable to predict the obscured landmarks with high accuracy. Nevertheless, the landmarking performance on the visible ear structures is not unaffected by the erroneous landmarks detected in the occluded areas.

TABLE 2 Landmarking performance on the ITWE dataset.

Method	Average PPE ± std	≤0.10 (%)
2-SHGNet (ours)	0.0519 ± 0.0173	99
3-SUNet	0.0708 ± 0.0313	92
2-SUNet	0.1837 ± 0.0683	17
Baseline	0.1327 ± 0.0523	47
HOG + AAM	0.0539 ± 0.0248	97
SIFT + AAM	0.0522 ± 0.0246	94
DCNN + AAM	0.0599 ± 0.0272	93
SIFT + PAAM	0.0563 ± 0.0264	93
ZZ init	0.1276 ± 0.0332	23

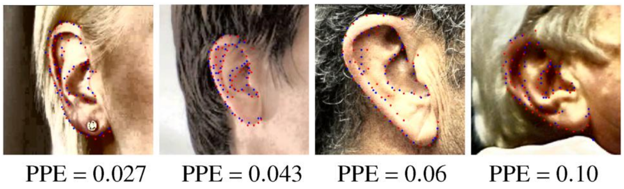


FIGURE 7 Visualisations of the detected landmarks with different PPE values. The reference landmarks are shown in red, the predicted landmarks in blue. Best viewed in colour and electronically.

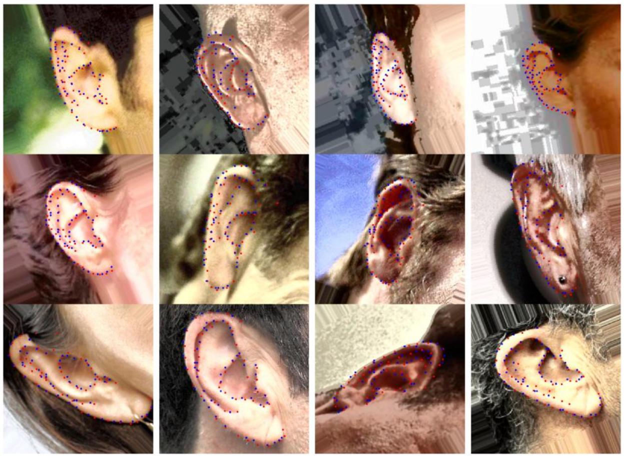


FIGURE 8 Qualitative landmarking examples generated by 2-SHGNet on the ITWE dataset. The predicted landmarks are shown in blue and ground truth in red. The figure is best viewed in colour in zoomed in.

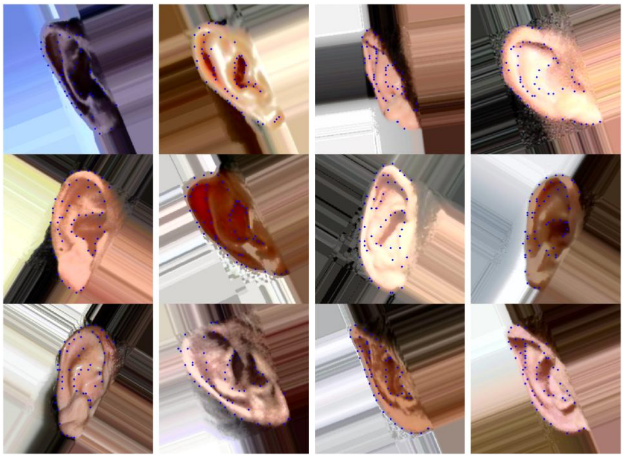


FIGURE 9 Qualitative landmarking examples generated by 2-SHGNet on the AWEx dataset. The predicted landmarks are shown in blue. Note that AWEx does not include ground truth landmark locations. The figure is best viewed in colour in zoomed in.

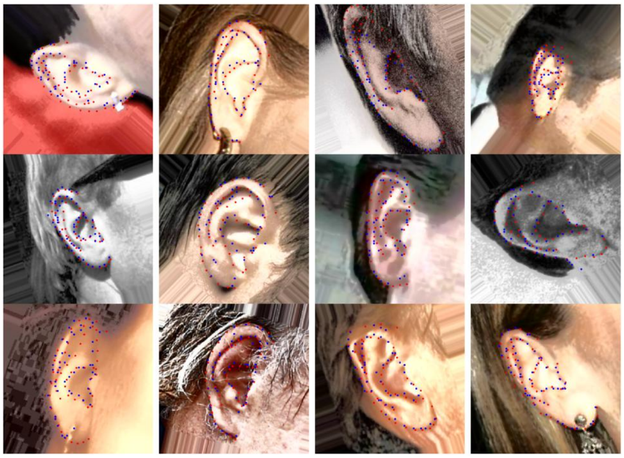


FIGURE 10 Example images from the ITWE dataset, where the proposed 2-SHGNet model produced lower PPE scores. The presented examples provide a cross-section of *failure* cases with the goal of providing insight into the shortcomings of the model. The predicted landmarks are shown in blue, the ground truth in red. The figure is best viewed in colour and zoomed in.

5.3 | Model limitations

In Figure 10 we analyse the shortcomings of 2-SHGNet through a few example images, on which the model yielded less convincing results. As can be seen, problems appear when the model fails to capture the whole ear shape, resulting in spurious landmarks. Significant occlusion of the ear structures adversely impact the model, producing erroneous landmarks—see, for example, the example in the top right corner of Figure 10. Moreover, landmarks are sometimes also detected in incorrect locations with low quality and noisy images. In cases where the colour of the background and ear are similar, the landmarks do not get accurately detected on the outer ear. Significant rotation in roll directions in combination with noisy data produces missing landmarks as well. Despite these issues, the presented results still show that even in challenging settings, 2-SHGNet still generates reasonable landmark predictions that can be used for ear alignment (or geometric normalisation), as also demonstrated in the next section.

5.4 | Geometric normalisation

The detected landmark locations are used to align the ear images with a pre-defined reference shape (i.e. a template with fixed landmark locations). In Figures 11 and 12, we show a few qualitative examples of ears from the ITWE and AWEx datasets before and after geometric normalisation. Note that despite the presence of spurious landmarks, the 2-SHGNet predictions are sufficiently accurate for alignment of ear images from the two datasets.

To provide a reference frame for the proposed alignment approach, we present in Figure 13, a qualitative comparison with two competing alignment techniques on the AWEx dataset. The first is based on the Cascaded Pose Regression (CPR) framework from Refs. [1, 23] and the second is a combination of the SIFT keypoint detector and RANSAC-based alignment, originally presented in Ref. [7]. As can be seen, the CPR-based method ensures only approximate alignment that mainly normalises for rotation and the rough size of the ears. The SIFT + RANSAC combination is conceptually similar to the alignment approach proposed in this paper, but due to the generality of the keypoint detector often leads to suboptimal results. The proposed approach, on the other hand, provides the most convincing results due to semantic relevance of the detected landmarks, which makes RANSAC-based alignment straight forward. We also need to note that the CPR[1] and SIFT + RANSAC[2] implementations used for the

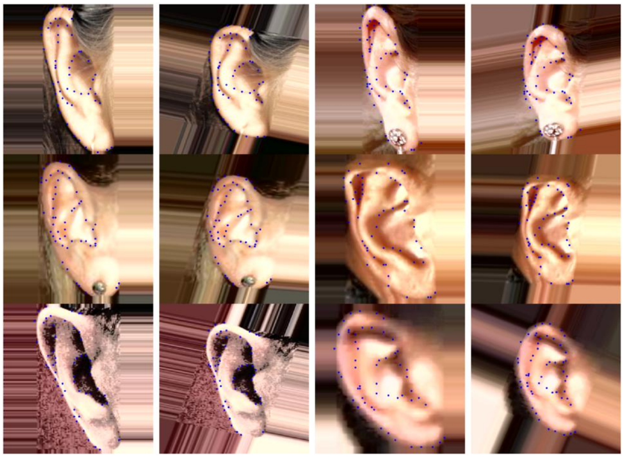


FIGURE 11 Normalisation results on the ITWE dataset. In each pair of images, the left part shows the unaligned ear and the right one shows the aligned version. The landmarks before and after geometric normalisation are marked blue.

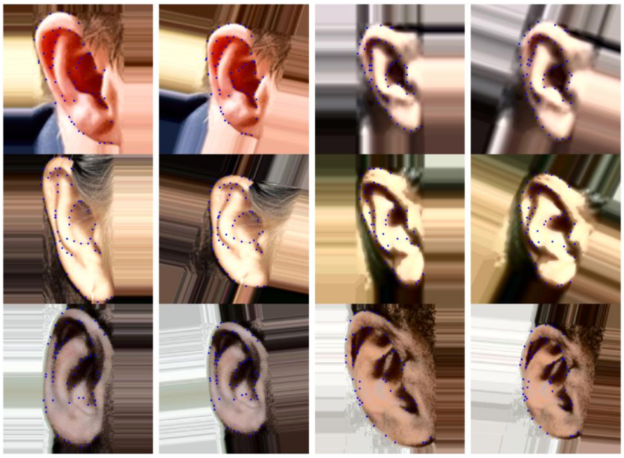


FIGURE 12 Normalisation results on the AWEx dataset. For each pair of images, the left part shows the unaligned ear and the right one shows the aligned version. The landmarks before and after geometric normalisation are marked blue.

[1] https://github.com/metodribic/ear-alignment-cpr

[2] https://github.com/metodribic/ear-alignment-ransac

experiments generated a large number of failure cases and were not able to properly align a significant number of images. The examples in Figure 13 represent some of the successfully aligned images.

To further validate, the performance of the implemented normalisation method, we computed the average ears of unaligned and aligned ITWE and AWEx images. The average ears of both datasets are shown in Figure 14. The produced average ears, in case of the unaligned images, are very blurry with a loosely defined ear shape. The average ears of aligned images, on the other hand, produce a clearer average ear shape with well-defined ear structures, pointing to the feasibility of our approach. For comparison purposes, the average ears computed from the (successfully aligned) AWEx image are presented in Figure 15 for the CPR-based and SIFT + RANSAC alignment approaches. While these aligned ears were computed from a much lesser number of images (due to a high failure rate) than the ones in Figure 14, they are somewhat crisper than the average unaligned image in the third column of Figure 14, suggesting that the images are now better aligned overall. However, compared to the result produced by the proposed methods, the average ears of the two competing methods are still much blurrier, suggesting weaker alignment capabilities.

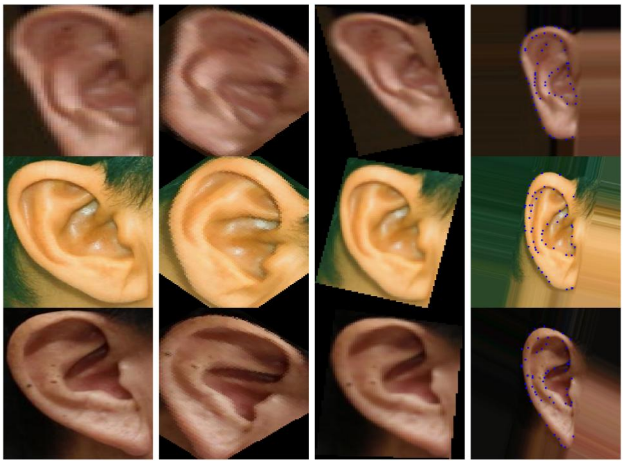


FIGURE 13 Normalisation results on the AWEx dataset. For each row, the first (far left) image shows the unaligned ear cropped from the input images after ear detection, the second one shows the aligned CPR-based [1] version, the third shows the image aligned with SIFT and RANSAC [7] and the fourth (far right) shows the proposed alignment.

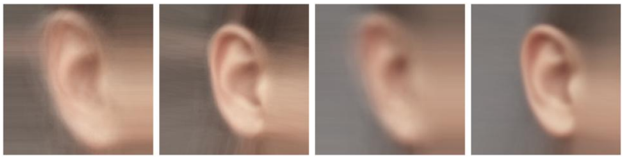


FIGURE 14 Average ears computed from unaligned and aligned ear images of the ITWE (left) and AWE (right) datasets. Note that the proposed geometric normalisation contributes towards better alignment of specific ear structures, which is reflected in the more crisp average appearance.

5.5 | Ear recognition results

In the final experimental series, we investigate the impact of alignment on ear recognition performance. To this end, we perform recognition experiments on aligned and unaligned images from the AWEx dataset [6], and evaluate the quality of our normalisation process using a number of recognition models [41]. Specifically, we train several deep learning models with a residual network topology [59], that is, ResNet-18, ResNet-50, ResNet-101, ResNet-152. Additionally, we report results for three different MobileNet versions [60]. MobileNet is a light-weight CNN architecture, developed for mobile and embedded computer vision applications. The architecture uses a hyperparameter to balance the size of the model, with higher values (the highest is 1) resulting in heavier models. For our evaluation, we train MobileNet (0.25), MobileNet (0.5) and MobileNet (1) following Ref. [41].

We report the results of the recognition experiments in terms of the Rank-1, Rank-5 and AUCMC scores in Table 3. As can be seen, ear alignment significantly improves recognition performance for all tested models and across all performance scores. We observe the highest performance increase with the ResNet-50 model, where the Rank-1 score is improved by 50% due to the alignment and the Rank-5 score is increased by 27%. If we focus only on the Rank-1 recognition rates, we can see relative improvements of 41% for ResNet-18, 23% ResNet-101, 24% for ResNet-152, 14% for MobileNet (0.25), 31% for MobileNet (0.5) and 17% for MobileNet (1). Similar improvements can also be observed for the other two performance indicators. These results clearly show that the normalisation procedure improves the performance of all tested ear recognition models.

6 | CONCLUSION

Unconstrained ear recognition remains a challenging task, especially when considering large ear pose variations. To address this issue, we studied the ear-landmark detection and alignment tasks in this work and developed a framework capable of locating a large number of ear landmarks in input images with diverse characteristics and aligning the ears with a pre-defined shape template. We formulated ear-landmark

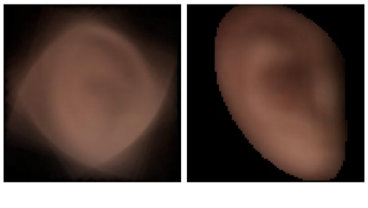


FIGURE 15 Average ears computed from the aligned ear images of the AWEx dataset using CPR-based [1] and SIFT + RANSAC [7] alignment. Compared to the average of the unaligned ears (third image in Figure 14), the average CPR (left) and SIFT + RANSAC (right) aligned ear are less crisp. Also, we note that these two average ears were computed only from a subset of images, on which the alignment procedures produced useful results, otherwise this would yield much blurrier results.

TABLE 3 Recognition results on the Extended AWE (AWEx) dataset. The performance is reported for aligned and unaligned ear images. Note that the alignment procedure significantly improves performance of all tested models and across all reported performance indicators.

	Unaligned			Aligned		
Model	**Rank-1 (%)**	**Rank-5 (%)**	**AUCMC (%)**	**Rank-1 (%)**	**Rank-5 (%)**	**AUCMC (%)**
ResNet-18	21.09	41.45	88.31	29.82	54.09	92.63
ResNet-50	20.18	42.14	89.55	30.27	53.55	92.72
ResNet-101	19.95	40.73	89.01	24.64	48.36	92.20
ResNet-152	20.41	42.59	88.98	25.36	48.86	91.85
MobileNet (0.25)	13.55	31.77	85.61	15.55	36.86	87.61
MobileNet (0.5)	15.36	35.45	87.59	20.18	43.09	89.61
MobileNet (1.0)	18.59	39.36	88.33	21.77	44.27	90.22

detection as a heatmap regression problem, leveraging on deep learning techniques as a means to solve it. Specifically, we implemented a stacked hourglass architecture consisting of two hourglass networks or 2D ear landmark detection in unconstrained scenarios, achieving highly competitive performance when compared to the state-of-the-art [4]. Furthermore, experiments on the ITWE and AWE datasets showed that the proposed framework was not only able to successfully align ear images captured in completely unconstrained scenarios but also that the alignment procedure was highly beneficial for the performance of ear recognition models.

As part of our future work, we plan to further refine the landmarking procedure and incorporate stronger (and task specific) shape to be able to robustly locate landmarks occluded by hair and other accessories commonly encountered with ear images. Additionally, we also plan to explore end-to-end models for ear alignment—with and without intermediate landmarking steps.

AUTHOR CONTRIBUTIONS

Anja Hrovatič: Data curation, Methodology, Software, Visualisation, Writing—original draft. **Peter Peer**: Conceptualisation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing—original draft, Writing—review & editing. **Vitomir Štruc**: Conceptualisation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing—original draft, Writing—review & editing. **Žiga Emeršič**: Conceptualisation, Formal analysis, Investigation, Methodology, Software, Supervision, Writing—original draft, Writing—review & editing.

ACKNOWLEDGEMENTS

The research presented in this work was partially supported by the ARRS Research Programme P2-0250(B) ‘Metrology and Biometric Systems’ and the ARRS Research Programme P2-0214 ‘Computer Vision’.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Peter Peer https://orcid.org/0000-0001-9744-4035
Žiga Emeršič https://orcid.org/0000-0002-3726-9404

REFERENCES

1. Pflug, A., Busch, C.: Segmentation and normalization of human ears using cascaded pose regression. In: Nordic Conference on Secure IT Systems, pp. 261–272 (2014)
2. Emeršič, Ž., Štruc, V., Peer, P.: Ear Recognition: More than a survey. Neurocomputing 255, 26–39 (2017). https://doi.org/10.1016/j.neucom.2016.08.139
3. Pflug, A., Busch, C.: Ear biometrics: a survey of detection, feature extraction and recognition methods. IET Biom. 1(2), 114–129 (2012). https://doi.org/10.1049/iet-bmt.2011.0003
4. Zhou, Y., Zaferiou, S.: Deformable models of ears in-the-wild for alignment and recognition. In: International Conference on Automatic Face & Gesture Recognition, pp. 626–633 (2017)
5. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biom. 7(3), 215–223 (2018). https://doi.org/10.1049/iet-bmt.2017.0210
6. Emeršič, Ž., et al.: Evaluation and analysis of ear recognition models: performance, complexity and resource requirements. Neural Comput. Appl. 32(20), 1–16 (2018). https://doi.org/10.1007/s00521-018-3530-1
7. Ribič, M., et al.: Influence of alignment on ear Recognition: case study on AWE dataset. In: International Electrotechnical and Computer Science Conference, vol. 25-B, pp. 131–134 (2016)
8. Hrovatič, A., et al.: Generation of 2D ear dataset with annotated view angles as a basis for angle-aware ear recognition. In: Electrotechnical and Computer Science Conference, pp. 264–267 (2019)
9. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
10. Asthana, A., et al.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3451 (2013)
11. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: EEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23(6), pp. 681–685 (2001)
12. Grenot-Castellano, E., Martínez-Díaz, Y., Silva-Mata, F.J.: Analysis of the impact of ear alignment on unconstrained ear recognition. In: Iberoamerican Congress on Pattern Recognition, pp. 283–293 (2019)

13. Sullivan, E.O., Zafeiriou, S.: 3D landmark localization in point clouds for the human ear. In: International Conference on Automatic Face and Gesture Recognition, pp. 402–406 (2020)
14. Mursalin, M., Islam, S.M.S.: Deep learning for 3D ear detection: A complete pipeline from data generation to segmentation. IEEE Access 9, 164976–164985 (2021)
15. Cihan Camgoz, N., et al.: Facial landmark localization in depth images using supervised ridge descent. In: International Conference on Computer Vision Workshops (ICCVW), pp. 136–141 (2015)
16. Wu, Y., Ji, Q.: Facial landmark detection: a literature survey. Int. J. Comput. Vis. 127(2), 115–142 (2019). https://doi.org/10.1007/s11263-018-1097-z
17. Khabarlak, K., Koriashkina, L.: Fast Facial Landmark Detection and Applications: A Survey (2021). arXiv preprint arXiv:2101.10808
18. Wang, N., et al.: Facial feature point detection: a comprehensive survey. Neurocomputing 275, 50–65 (2018). https://doi.org/10.1016/j.neucom.2017.05.013
19. Križaj, J., et al.: Simultaneous multi-descent regression and feature learning for facial landmarking in depth images. Neural Comput. Appl. 32(24), 17909–17926 (2020). https://doi.org/10.1007/s00521-019-04529-7
20. Chen, L., Su, H., Ji, Q.: Face alignment with kernel density deep neural network. In: International Conference on Computer Vision (ICCV), pp. 6992–7002 (2019)
21. Ye, J., et al.: An improved boundary-aware face alignment using stacked dense U-Nets. Int. J. Adv. Rob. Syst. 17(4), 172988142094090 (2020). https://doi.org/10.1177/1729881420940900
22. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Conference on Computer Vision and Pattern Recognition Workshops, pp. 79–87 (2017)
23. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085 (2010)
24. Wu, W., Cai, Y., Zhou, Q.: Transmarker: a pure vision transformer for facial landmark detection. In: International Conference on Pattern Recognition (ICPR), pp. 3580–3587 (2022)
25. Colaco, S., Yoon, Y.J., Han, D.S.: UIRnet: facial landmarks detection model with symmetric encoder-decoder. In: 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 407–410 (2022)
26. Keong, J., et al.: Multi-spectral facial landmark detection. In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2020)
27. Guo, J., et al.: Stacked Dense U-Nets with Dual Transformers for Robust Face Alignment (2018). arXiv preprint arXiv:1812.01936
28. Zhang, J., Hu, H.: Exemplar-based cascaded stacked auto-encoder networks for robust face alignment. Comput. Vis. Image Understand. 171, 95–103 (2018). https://doi.org/10.1016/j.cviu.2018.05.002
29. Zhang, J., Hu, H., Shen, G.: Joint stacked hourglass network and salient region attention refinement for robust face alignment. ACM Trans. Multimed Comput. Commun. Appl 16(1), 1–18 (2020). https://doi.org/10.1145/3374760
30. Matkowski, W.M., Chai, T., Kong, A.W.K.: Palmprint recognition in uncontrolled and uncooperative environment. IEEE Trans. Inf. Forensics Secur. 15, 1601–1615 (2019). https://doi.org/10.1109/tifs.2019.2945183
31. Yin, Y., et al.: Dual-attention gan for large-pose face frontalization. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 249–256 (2020)
32. Cao, J., et al.: Towards high fidelity face frontalization in the wild. Int. J. Comput. Vis. 128(5), 1485–1504 (2020). https://doi.org/10.1007/s11263-019-01229-6
33. Liu, Y., Chen, J.: Unsupervised face frontalization for pose-invariant face recognition. Image Vis. Comput. 106, 104093 (2021). https://doi.org/10.1016/j.imavis.2020.104093
34. Zhang, Z., et al.: Face frontalization using an appearance-flow-based convolutional neural network. IEEE Trans. Image Process. 28(5), 2187–2199 (2018). https://doi.org/10.1109/tip.2018.2883554
35. Reddy, N., Rattani, A., Derakhshani, R.: Generalizable deep features for ocular biometrics. Image Vis. Comput. 103, 103996 (2020). https://doi.org/10.1016/j.imavis.2020.103996
36. Xin, M., Mo, S., Lin, Y.: EVA-GCN: head pose estimation based on graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1462–1471 (2021)
37. Wood, E., et al.: 3D face reconstruction with dense landmarks. In: European Conference on Computer Vision. (ECCV), pp. 160–177 (2022)
38. Cai, M., et al.: 3d face reconstruction and dense alignment with a new generated dataset. Displays 70, 102094 (2021). https://doi.org/10.1016/j.displa.2021.102094
39. Chang, K., et al.: Comparison and combination of ear and face images in appearance-based biometrics. Transactions on pattern analysis and machine intelligence 25(9), 1160–1165 (2003). https://doi.org/10.1109/tpami.2003.1227990
40. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recogn. 45(3), 956–968 (2012). https://doi.org/10.1016/j.patcog.2011.06.005
41. Emeršič, Ž., et al.: Deep ear recognition pipeline. In: Recent Advances in Computer Vision, pp. 333–362. Theories and Applications (2019)
42. Štepec, D., et al.: Constellation-Based Deep Ear Recognition, pp. 161–190. Deep biometrics (2020)
43. Emeršič, Ž., et al.: ContexedNet: context–aware ear detection in unconstrained settings. IEEE Access 9, 145175–145190 (2021)
44. Biau, G., Devroye, L.: Lectures on the Nearest Neighbor Method, vol. 246 (2015)
45. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: International Conference on Acoustics, Speech and Signal Processing, pp. 8609–8613 (2013)
46. Rockwell, C.: Stacked Hourglass Networks in Pytorch [Online]. (2021). https://github.com/princeton-vl/pytorch_stacked_hourglass. Accessed 2 March 2021
47. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037 (2019)
48. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV), pp. 483–499 (2016)
49. Derpanis, K.G.: Overview of the RANSAC algorithm. Image Rochester NY 4(1), 2–3 (2010)
50. Križaj, J., et al.: Simultaneous multi-descent regression and feature learning for facial landmarking in depth images. Neural Comput. Appl. 32(24), 17909–17926 (2019). https://doi.org/10.1007/s00521-019-04529-7
51. Kingma, D.P., Ba, J., adam: A Method for Stochastic Optimization. ArXiv, 2014. [Online]. arxiv.org/pdf/1603.04467.pdf
52. Ruder, S.: An Overview of Gradient Descent Optimization Algorithms. ArXiv, 2016. [Online]. arxiv.org/pdf/1609.04747.pdf
53. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICAI), pp. 234–241 (2015)
54. Antonakos, E., et al.: Feature-based lucas–kanade and active appearance models. IEEE Trans. Image Process. 24(9), 2617–2632 (2015). https://doi.org/10.1109/tip.2015.2431445
55. Lowe, D.G.: Object recognition from local scale-invariant features. Int. J. Computer Vis. 60(2), 91–110 (2004). https://doi.org/10.1023/b:visi.0000029664.99615.94
56. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)
57. Sermanet, P., et al.: Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks (2013). arXiv preprint arXiv:1312.6229
58. Tzimiropoulos, G., Pantic, M.: Gauss-Newton deformable part models for face alignment in-the-wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1851–1858 (2014)

59. He, K., et al.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

60. Howard, A.G., et al.: Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017). arXiv preprint arXiv:1704.04861

How to cite this article: Hrovatič, A., et al.: Efficient ear alignment using a two-stack hourglass network. IET Biome. 1–14 (2023). https://doi.org/10.1049/bme2.12109