Neural Computing and Applications (2020) 32:15785–15800
https://doi.org/10.1007/s00521-018-3530-1

S.I.: ADVANCES IN BIO-INSPIRED INTELLIGENT SYSTEMS

Evaluation and analysis of ear recognition models: performance, complexity and resource requirements

Žiga Emeršič[1] · Blaž Meden[1] · Peter Peer[1] · Vitomir Štruc[2]

Received: 22 December 2017 / Accepted: 4 May 2018 / Published online: 16 May 2018

Abstract
Ear recognition technology has long been dominated by (local) descriptor-based techniques due to their formidable recognition performance and robustness to various sources of image variability. While deep-learning-based techniques have started to appear in this field only recently, they have already shown potential for further boosting the performance of ear recognition technology and dethroning descriptor-based methods as the current state of the art. However, while recognition performance is often the key factor when selecting recognition models for biometric technology, it is equally important that the behavior of the models is understood and their sensitivity to different covariates is known and well explored. Other factors, such as the train- and test-time complexity or resource requirements, are also paramount and need to be consider when designing recognition systems. To explore these issues, we present in this paper a comprehensive analysis of several descriptor- and deep-learning-based techniques for ear recognition. Our goal is to discover weak points of contemporary techniques, study the characteristics of the existing technology and identify open problems worth exploring in the future. We conduct our analysis through identification experiments on the challenging Annotated Web Ears (AWE) dataset and report our findings. The results of our analysis show that the presence of accessories and high degrees of head movement significantly impacts the identification performance of all types of recognition models, whereas mild degrees of the listed factors and other covariates such as gender and ethnicity impact the identification performance only to a limited extent. From a test-time-complexity point of view, the results suggest that lightweight deep models can be equally fast as descriptor-based methods given appropriate computing hardware, but require significantly more resources during training, where descriptor-based methods have a clear advantage. As an additional contribution, we also introduce a novel dataset of ear images, called AWE Extended (AWEx), which we collected from the web for the training of the deep models used in our experiments. AWEx contains 4104 images of 346 subjects and represents one of the largest and most challenging (publicly available) datasets of unconstrained ear images at the disposal of the research community.

Keywords Ear recognition · Covariate analysis · Convolutional neural networks · Feature extraction

1 Introduction

Automatic ear recognition represents a subproblem of biometrics with important applications in security, surveillance and forensics. Many techniques have been proposed in the literature for ear recognition systems ranging from geometric and holistic techniques [2, 4, 11, 54, 56] to more recent descriptor- [6, 8, 33, 36, 42] and deep-learning-based [16–18, 23, 55] methods. While descriptor-based methods have dominated the field over the last years, research is moving away from these methods and is now focusing increasingly on deep-

✉ Žiga Emeršič
ziga.emersic@fri.uni-lj.si

Blaž Meden
blaz.meden@fri.uni-lj.si

Peter Peer
peter.peer@fri.uni-lj.si

Vitomir Štruc
vitomir.struc@fe.uni-lj.si

[1] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

[2] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia

learning-based models, which recently brought about considerable advancements in various areas of computer vision and beyond.

When developing ear recognition technology, it is of paramount importance to understand how different recognition models behave when applied on challenging data captured in unconstrained environments where the technology is ultimately deployed. Variations across pose, gender, ethnicity, occlusions and alike are common in these environments and may influence the choice of recognition approach for a particular application. While one approach may result is lower performance on established ear recognition benchmarks, it may still exhibit robustness to certain covariates and, hence, be favored over another in specific circumstances. Similarly, if computing resources are limited or the run-time complexity is important, one recognition model may be preferred over the other even at the cost of somewhat lower performance. To facilitate informed choices during the R&D work, it is therefore crucial to have empirical evidence about the properties of the available recognition approaches and have insight into their characteristics.

While most of the research on ear recognition is focused on new detection and enrollment techniques, more elaborate recognition models and more discriminative representations of ear images, studies focusing on the strengths and weaknesses of existing techniques are largely missing from the literature. In this paper, we try to fill this gap and present a comprehensive analysis of several ear recognition techniques considered state-of-the-art today. Specifically, we experiment with eight (dense) descriptor-based ear recognition techniques and three recent deep-learning-based recognition models and analyze their characteristics through comprehensive experiments on a challenging dataset of ear images, captured in unconstrained settings (a.k.a in the wild). We aim at identifying which factors (or covariates) influence the recognition techniques the most and, hence, contribute the greatest to recognition errors, what kind of computational complexity is induced by the recognition techniques during train and test time and what resources are required to run the selected techniques. A detailed understanding of these factors is extremely important not only because it allows us to devise more effective recognition techniques, but because it helps to identify future research trends in this area as well.

1.1 State-of-the-art recognition models

Most of the existing surveys on ear recognition, e.g., [1, 24, 41] identify descriptor-based recognition techniques as the state of the art in this field. However, as shown by more recent group evaluations and challenges, e.g, [23] deep-learning methods, and in particular convolutional neural networks (CNNs), are starting to dominate the field due to their excellent performance and ability to learn image representations (descriptors) directly from the training data. These two groups of techniques approach the ear recognition in fundamentally different ways.

Descriptor-based techniques, for example, extract identity cues from local image areas and use the extracted information for identity inference. As emphasized by Emeršič et al. [24], two groups of techniques can in general be considered descriptor based: (1) techniques that first detect interest points in the image and then compute descriptors for the detected interest points and (2) techniques that compute descriptors densely over the entire images based on a sliding-window approach (with or without overlap). Examples of techniques from the first group include [3, 9] or more recently [46]. A common characteristic of these techniques is the description of the interest points independently one from the other, which makes it possible to design matching techniques with robustness to partial occlusions of the ear area. Examples of techniques from the second group include [5, 10, 31, 52]. These techniques also capture the global properties of the ear in addition to the local characteristics which commonly results in a higher recognition performance, but the dense descriptor computation procedure comes at the expense of the robustness to partial occlusions. Nonetheless, recent trends in ear recognition favor dense descriptor-based techniques primarily due to their computational simplicity and high recognition performance.

Deep-learning-based methods, on the other had, typically process the input images in a holistic manner and learn image representations (features, descriptor) directly from the training data by minimizing some suitable loss at the output of the recognition model. The most popular deep-learning models, CNNs, commonly process the data through a hierarchy of convolutional and pooling layers that can be seen as stacked feature extractors and once fully trained can be used to derive highly discriminative data representations from the input images that can be exploited for identity inference. While these representations commonly ensure formidable recognition performance, the CNN-training procedure typically requires a large amount of training data, which may not always be available and is not needed with descriptor-based methods. In the field of ear recognition, deep-learning-based methods appeared only recently [16–18, 23, 55], but were already shown to outperform local descriptor-based methods (see [23]).

We contribute to a better understanding of these methods in this work by conducting an analysis of some of the key characteristics of both groups of techniques.

1.2 Contributions and paper organization

We make several important contributions in this paper. A short list with brief summaries is given below:

- *Experimental evaluation* We conduct a comprehensive experimental evaluation of several state-of-the-art ear recognition techniques on a challenging dataset of ear images gathered from web with the goal of studying unconstrained ear recognition. As part of the evaluation, we perform a comparative assessment of recent descriptor- and deep-learning-based ear recognition techniques and investigate their robustness by studying the impact of various covariates, such as ethnicity, head rotation (in terms of yaw, roll and tilt angles), gender and presence of occlusions and accessories.
- *Dataset* To be able to train the deep-learning-based recognition models, we gather a new dataset of unconstrained ear images from the web and make it publicly available to the research community. The new dataset, named Annotated Web Ears Extended (AWEx), contains 4104 images of 346 subjects and to the best of our knowledge represents one of the most challenging and largest datasets of this kind available for research purposes. The dataset can be downloaded from: http://awe.fri.uni-lj.si/.
- *Analysis* We present an extensive analysis of the evaluated recognition techniques in terms of recognition performance, computationally complexity and resource requirements and thus contribute to new knowledge in the field and a better understanding of their characteristics.

The rest of the paper is structured as follows. In Sect. 2, we review existing work related to our paper and further motivate our analysis. We describe our experimental setup and the ear recognition techniques considered in this work in Sect. 3 and introduce the experimental dataset and protocol in Sect. 4. We present the results of our analysis and discuss its implications in Sect. 5. We conclude the paper with some final comments and directions for future work in Sect. 6.

2 Motivation and related work

Understanding the characteristics of biometric recognition technology is of considerable importance to the advancement of the field and key for researchers in this area. What properties of the input data make the recognition process difficult? What properties make it is easy? Are certain techniques better suited for specific data characteristics than others? What is the computational complexity of the recognition techniques? What kind of resources are required? Answers to questions like these make it possible to target weak points of the existing technology and provide directions for future research.

In the field of biometric ear recognition, some of the questions outlined above are (partially) discussed in recent survey papers, such as [1, 24, 41, 47], where structured comparisons of existing ear recognition techniques are presented. The comparisons in these papers are based on previously reported results and summarize recognition experiments on different datasets with different experimental protocols. While general trends about the advancement of ear recognition technology over the years are presented and some of the strengths and weaknesses are identified, no detailed information about the performance of the existing techniques with respect to different covariates is given.

Similarly to our work, the survey by Emeršič et al. [24] also presents a comparison of some descriptor-based feature extraction techniques from the literature using a challenging dataset and predefined experimental protocol. However, here we focus on the impact of different covariates on the recognition performance of the tested techniques and include analysis of CNN-based approaches as well. Also, in this paper we provide the analysis of important model characteristics such as time and space complexity.

Pflug and Busch [41] compare the performance of various texture and surface descriptors for ear biometrics, but different from our work uses the descriptors in combination with subspace projection techniques. The reported experiments are conducted on a dataset of ear images with laboratory-like quality, but no ablation study is presented.

The study from [44] is likely the closest to our work as far the evaluation of descriptor-based ear recognition techniques with respect to different covariates is concerned. However, the focus here is on only on image characteristics, such as noise and blurring, and not on ear-related covariates and other factors such as in our work. Furthermore, since deep-learning models were not yet applied to ear recognition at the time of writing of [44], the study does not include the most recent and promising ear recognition models.

The analysis presented in this work builds on the preliminary version from [19], but extends the study to deep-learning models, new model characteristics and novel aspects that were not considered before.

3 Methodology

In this section, we present the methodology used for our analysis. We start the section with a description of the experimental setup used and then describe the descriptor-

and deep-learning-based recognition techniques considered.

3.1 Experimental setup

To analyze the characteristics of different recognition models, we use an identification pipeline as illustrated in Fig. 1. Here, the input images are first subjected to a feature extraction technique that converts the input RGB ear images into a discriminative representation by exploiting either a deep CNN model (marked as scenario A in Fig. 1) or a descriptor computation procedure (marked as scenario B in Fig. 1). Once the representation is computed for a given test images, identity inference is conducted based on the cosine similarities with a predefined set of gallery images.

The main difference between the experimental setups (i.e., scenarios A and B) for the deep-learning and descriptor-based pipelines is that for setup A we start by using train and validation sets to learn the parameters of our CNN models, while for setup B no training is needed. If we denote the input images into our pipeline as $\mathbf{x} \in \mathbb{R}^n$, then both setups produce image representations (feature vectors) $\mathbf{y} \in \mathbb{R}^d$ from the input images as:

$$\mathbf{y} = f(\mathbf{x}), \tag{1}$$

where $f(\cdot)$ is a feature extraction function.

For this work, we consider eight descriptor-based recognition approaches and three deep CNN models for our experiments based on either their reported performance for ear recognition [24, 41] or their popularity within the research community [25]. We describe the considered approaches in the following two sections.

3.2 Descriptor-based ear recognition techniques

For the descriptor-based methods, we consider dense descriptor computation and generate the d-dimensional feature vectors needed for recognition from grayscale converted input images (see Fig. 1—setup B). Specifically, we implement methods based on local binary patterns (LBPs [7, 24, 26, 43, 45]), (Rotation Invariant) Local Phase Quantization Features (RILPQ and LPQ [39, 40]), binarized statistical image features (BSIF [24, 30, 43]), Histograms of Oriented Gradients (HOGs, [12, 13, 24, 43]),

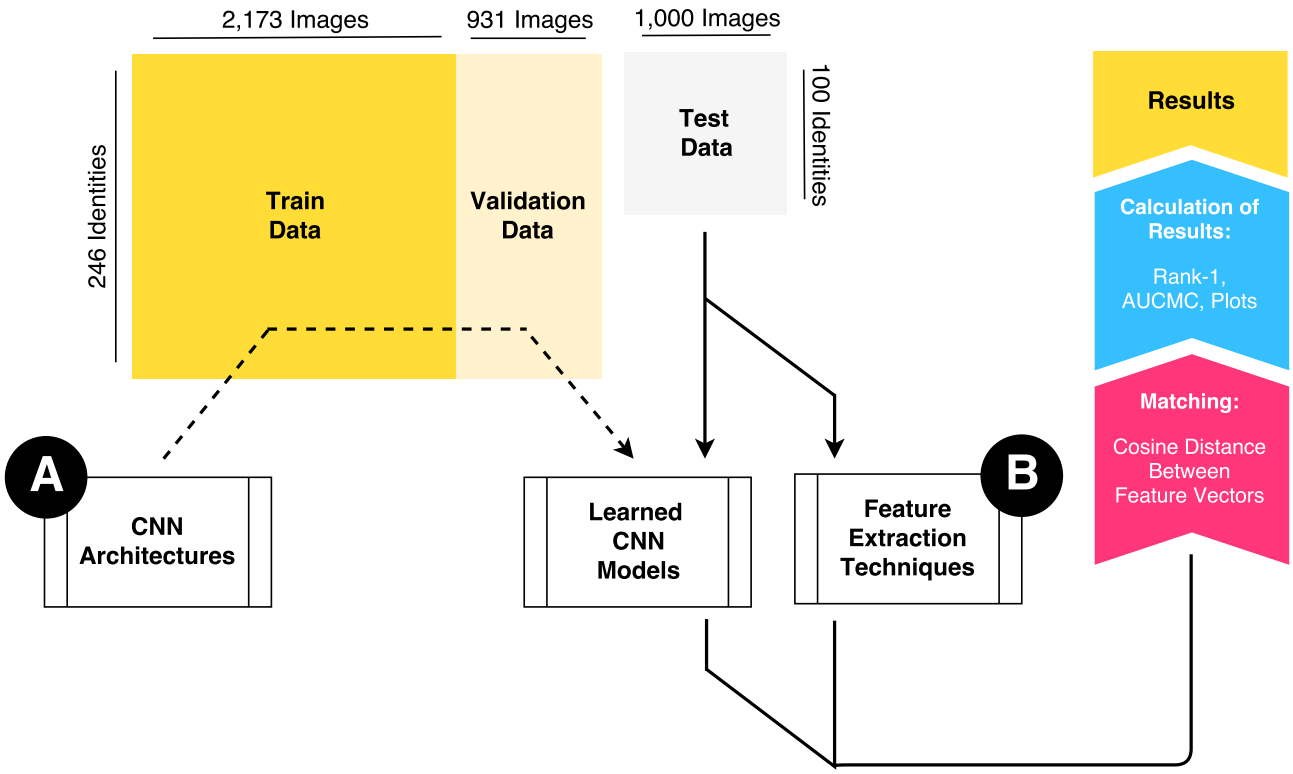

Fig. 1 The identification pipeline used in our experiments. We employ the same pipeline for descriptor-based and deep-learning-based recognition techniques. In both cases, features are extracted from the input images using either dense descriptor computation or CNN models. The main difference between the two approaches is in the experimental setup (A or B), where for the CNN models a training procedure involving train and validation sets is required (case A), whereas for the descriptor-based technique no training is needed (case B). After the feature extraction step, the procedure is the same for both scenarios, (A) and (B). Each extracted test feature vector is matched against all gallery feature vectors using the cosine similarity, and the ID of the most similar gallery vector is returned as the output

the Dense Scale-Invariant Feature Transform (DSIFT, [15, 24, 31]), Gabor wavelets [14, 24, 33, 34, 49–51] and Patterns of Oriented Edge Magnitudes (POEM, [24, 52]) for the analysis.

3.2.1 Local binary patterns

Local binary patterns (LBPs) represent powerful texture descriptors that achieved competitive recognition performance in various areas of computer vision [45]. The use of the LBP descriptor for ear recognition is mainly motivated by its computational simplicity and the fact that the texture of the ear is highly discriminative. Many successful ear recognition techniques have been presented in the literature exploiting LBPs either as stand-alone texture representations or in combination with other techniques, e.g., [7, 26, 43].

LBPs encode the local texture of an image by generating binary strings from circular neighborhoods of points thresholded at the gray-level value of their center pixels. The generated binary strings are interpreted as decimal numbers and assigned to the center pixels of the neighborhoods. The number of sampling points P used to generate the binary strings depends on the radii R of the circular neighborhoods and results in the following encoding [45]:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} 2^p s(g_p - g_c), \tag{2}$$

where $\mathrm{LBP}_{P,R}$ stands for the computed binary pattern of some center pixel, g_c and g_p denote the gray-level values of the center pixel and the pth pixel from the neighborhood, respectively, and the thresholding function $s(\cdot)$ stands for:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In practice, not all binary patterns returned by Eq. (2) are useful for texture representation. Typically, only binary strings with at most two bitwise transitions from 0 to 1 (or vice versa) are considered in the final descriptor. For a 8-pixel neighborhood and a consequent 8-bit binary string, for example, exactly 58 such patterns (called uniform patterns) can be computed. Most methods exploiting LBPs with a 8-pixel neighborhood for texture description, therefore, compute 59-bin histograms from local image blocks and then concatenate the computed histograms over all blocks into a global texture descriptor (our d-dimensional feature vector $\mathbf{x}$) that can be used for recognition. A similar procedure is also used in our experiments in Section 5.

3.2.2 (Rotation Invariant) Local Phase Quantization

Local Phase Quantization (LPQ) features [39] are very similar in essence to LBPs, as the local image texture is again encoded using binary strings, and histograms are again computed from the binary strings of local image blocks and concatenated into the final representation of the given image. LPQ features are computed from the Fourier phase spectrum of an image and are known to be invariant to blurring under certain conditions. This feature makes LPQs an attractive alternative for ear recognition (see, e.g., [43]), where blurred and low-resolution images represent a problem for the existing technology.

With LPQ, the local neighborhoods of every pixel in the image are first transformed into the frequency domain using a short-term Fourier transform. Local Fourier coefficients are computed at four selected frequency points, and the local phase information contained in these (complex) coefficients is then encoded. Here, a similar quantization scheme is used as in iris recognition systems, where every complex Fourier coefficient contributes two bits to the final binary string. The result of this coding procedure is a 8-bit binary string for every pixel in the image from which the local 256-bin histograms are computed and later concatenated into a global descriptor of the image.

An extension of this technique to Rotation Invariant Local Phase Quantization (RILPQ) features was presented in [40]. The idea here is similar to the original LPQ technique with the difference that a characteristic orientation is first estimated for the given local neighborhood, and then, this orientation is used to compute a directed version of the binary descriptor. The binary descriptor is computed with the same procedure as the original LPQ, but every local neighborhood is first rotated in accordance with its characteristic orientation. RILPQ descriptors are not only blur invariant, but also exhibit a certain degree of robustness toward image rotation.

3.2.3 Binarized statistical image features

Binarized statistical image features (BSIF) [30] represent a more recent tool for texture description. Here, binary strings (encoding texture information) are again constructed for each pixel in the image, but this time by projecting image patches onto a subspace, whose basis vectors are learned from natural images. The subspace coefficients are then binarized using simple thresholding. This procedure is equivalent to filtering the input image with a number of pre-learned filters and binarizing the filter responses at each pixel location. Each filter contributes 1 bit to the binary string of a pixel making the length of the binary string dependent on the number of filter used. Similar to LBP and LPQ, the binary string of each pixel is

interpreted in decimal form and a global histogram-based representation (our d-dimensional feature vector $\mathbf{x}$) is constructed for the given images by concatenating histograms constructed from smaller image blocks.

The main characteristic that makes BSIF features so appealing is the fact that the binary strings are not constructed based on heuristic operations, but on the basis of statistics of natural images. The idea behind BSIF-based texture description is in line with recent feature learning approaches, which produced competitive results for many computer vision problems in recent years. The use of BSIF features for ear recognition was advocated by Pflug et al. [43], where excellent performance was reported.

3.2.4 Histograms of Oriented Gradients

Descriptors exploiting Histograms of Oriented Gradients (HOGs) were originally introduced for the problem of human detection by Dalal and Triggs [12], but have since been successfully applied to various fields of computer vision, including ear recognition [13, 43]. HOG descriptors have excellent texture description properties and are considered robust toward moderate illumination changes. This fact makes them highly suitable for problems, such as ear recognition, where illumination-induced variability is one of the major problems.

HOGs are computed based on a simple procedure. The computation starts by calculating the gradient of the image using 1-dimensional convolutional masks, i.e., $[-1, 0, 1]$ and $[-1, 0, 1]^{\mathrm{T}}$. In the next step, the image is divided into a number of cells and compact histograms of quantized gradient orientations are computed for each cell. Here, a voting procedure is used during histogram construction, so that pixels with higher gradient magnitudes contribute more to the histogram bins than pixels with lower magnitudes. Neighboring cells are then grouped into larger blocks and normalized to account for potential changes in contrast and illumination. This normalization procedure is applied in a sliding-window manner over the entire image with some overlap between neighboring blocks. Ultimately, all normalized histograms are concatenated into the final HOG descriptor (our feature vector $\mathbf{x}$) that can be used for matching and recognition.

3.2.5 Dense Scale-Invariant Feature Transform

The Scale-Invariant Feature Transform (SIFT), introduced by Lowe in [35], represents one of the most successful techniques for image description in computer vision. The original approach to SIFT calculation includes both a keypoint detector, capable of finding points of interest in an image, as well as a local descriptor that can effectively represent the local neighborhood around the detected keypoints. As indicated in the introductory section, early techniques to ear recognition relied on the SIFT keypoint detector as well as the SIFT descriptor, e.g., [3, 15] and [9], and, therefore, demonstrated a high degree of robustness toward partial occlusions.

More recent techniques, on the other hand, compute dense SIFT (DSIFT) representations from the images and do not rely on the keypoint detector. Here, the keypoints are simply arranged uniformly into a grid that is placed over the image. Techniques based on DSIFT (e.g., [31, 37]) have reported excellent recognition performance as well as robustness to partial occlusions similar to techniques based on the original SIFT formulation. We evaluate a DSIFT-based technique in the experimental section and, thus, discuss here only the SIFT keypoint descriptor. The reader is referred to [35] for a detailed description of the keypoint detection procedure.

The SIFT descriptor shares similarities with the HOG descriptor. For every point of interest, SIFT considers a local neighborhood of 16×16 pixels. This neighborhood is partitioned into subregions of 4×4 pixels, and for each subregion, an 8-bin histogram is computed based on the orientations and magnitudes of the image gradient in that subregion. The gradients are also weighted by a Gaussian function to give more importance to image gradients closer to the point of interest and normalized by the dominant gradient orientation to achieve rotation invariance. The final dimensionality of the SIFT descriptor is 128 for a single keypoint, so care needs to be taken when computing DSIFT representations from the image. The dimensionality of final feature vector can easily become computationally prohibitive if too many grid points are chosen for DSIFT calculation.

3.2.6 Gabor wavelets

2D Gabor wavelets were originally introduced by Daugman [14] for the problem of iris coding, but due to their ability to analyze images at multiple scales and orientations, they have been successfully employed in other problem areas as well. In the spatial domain, Gabor wavelets are defined with the following expression[49, 50]:

$$\psi_{u,v}(x,y) = \frac{f_u^2}{\pi\gamma\eta} \mathrm{e}^{-\left(\frac{f_u^2}{\gamma^2}x'^2 + \frac{f_u^2}{\eta^2}y'^2\right)} \mathrm{e}^{j2\pi f_u x'}, \tag{4}$$

where

$$\begin{aligned} x' &= x\cos\theta_v + y\sin\theta_v, \\ y' &= -x\sin\theta_v + y\cos\theta_v \end{aligned} \tag{5}$$

and the parameters f_u and θ_v represent the center frequency and orientation of the complex sinusoidal from Eq. (4),

respectively. γ and η define the ratio between the center frequency and the size of the Gaussian and ensure that all generated wavelets share some specific properties [51]. For feature extraction, a family of wavelets is typically created and used to extract features from the processed image. This family commonly consist of wavelets of 5 scales and 8 orientations, i.e., $f_0, f_1, ..., f_7$ and $\theta_0, \theta_1, ..., \theta_4$.

To extract Gabor features from an image, the image is convolved with the entire family of Gabor wavelets (filters), and the magnitude responses of the convolution outputs are retained (the phase responses are discarded), down-sampled and concatenated into a global feature vector encoding multi-resolution, orientation-dependent texture information of the input image.

Techniques based on the outlined procedure and its modifications (e.g., using log-Gabor wavelets) are among the most popular techniques for ear recognition [33, 34, 36, 38, 53]. Their advantages lie in their excellent discriminative properties; however, Gabor features are computational relatively complex to compute, as the input image needs to be filtered with an entire family of filters.

3.2.7 Patterns of Oriented Edge Magnitudes

Patterns of Oriented Edge Magnitudes (POEM) [52] represent another popular approach to texture description that combines ideas from LBP and HOG descriptors as well as Gabor wavelets.

The POEM construction procedure starts by computing the gradient of the input image and building magnitude-weighted histograms of gradient orientations for every pixel in the image. This histogram is computed from local pixel neighborhoods referred to by the authors as cells. In this regard, POEM shares similarities with the HOG descriptor, which also relies on gradient directions to encode an image, but different from HOG, POEM computes the histograms densely in a sliding-window manner over the entire image. After this step, every pixel in the image is represented by a local histogram of quantized gradient orientations, or in other words, the image is decomposed into m oriented gradient images, where m is the number of discrete orientations of the local histograms. Each of these images is then encoded using the LBP operator, and a global image descriptor is constructed by concatenating all block histograms computed from the oriented gradient images.

The POEM descriptor has demonstrated impressive performance for face recognition [52] and exhibits desirable properties, such as orientational selectivity, robustness to moderate illumination changes and low-computational complexity, which make it appealing for image representation in ear recognition systems.

3.3 Deep-learning-based ear recognition models

We use the deep-learning-based recognition models as black-box feature extractors in this work and exploit the image representations produced by one of the final layers (i.e., one of the layers before the softmax) of the models as the d-dimensional feature vectors (descriptors) of the input images needed for recognition (see Fig. 1—setup A). We consider three different CNN models for our analysis, which cover some of the most popular architectures for recognition networks from the literature, i.e., ResNet [27], SqueezeNet [29] and the VGG network from [48].

3.3.1 VGG network

The VGG network (or model) [48] is a representative of so-called very deep CNN models and in the most common configuration comprises a total of 16 network layers (VGG-16). The VGG model has been successfully applied to numerous recognition problems, including ear recognition (see, e.g., [17, 18, 22, 23]), and has been shown to ensure state-of-the-art performance of challenging ear datasets.

The main characteristic of the model is the use of several consecutive convolutional layers with small 3×3 filters. The consecutive stacks of 3×3 convolutional layers are able to capture the same information as the larger filters used in older model architectures, such as AlexNet [32], but require significantly less parameters that need to be estimated during training. The 3×3 filter stacks are interspersed with max-pooling layers which reduce the dimensionality of the activation maps produced by the model layers. The convolutional part of the VGG model is followed by three fully connected layers with 4096, 4096 and 1000 channels, respectively. Finally, a softmax layer is used at the top of the model to facilitate training. For our experiments, we perform network surgery on the VGG model and use the $d = 4096$ dimensional output of the penultimate fully connected layer ($fc7$) as the feature vector of the VGG model.

3.3.2 SqueezeNet

SqueezeNet (SNet in the experimental section) represents a recent CNN model which was shown to ensure AlexNet-level accuracy with $50\times$ fewer parameters [29]. The model was first introduced to the field of ear recognition in [22] with highly competitive results.

SqueezeNet builds on ideas from residual networks [27, 28], but additionally introduces so-called squeeze layers (i.e., convolutional layers with 1×1 convolutions) that serve as bottlenecks of the CNN architecture and aim at further reducing the parameter space of the

overall model. The network exploits a few additional design principles: (1) replacement of part of the 3×3 filters in the convolutional layers with 1×1 filters, (2) postponing the down-sampling steps to later stages in the network so that convolutional layers have large activation maps [29] and (3) network pruning. The result of these design choices is a model that has significantly less parameters to tune than competing models and can be trained on relatively small amounts of data. This means that the model has a small data footprint and that the training is much faster in comparison with the original AlexNet implementation. For our experiments, we use the output of the model layer preceding softmax as the $d = 86,527$ dimensional feature vector of the SqueezeNet model.

3.3.3 Residual network: ResNet50

Residual networks (ResNets or RNets, [27]) belong to a recent class of deep models that introduced shortcut (or skip) connections into CNN models. These connections represent identity shortcuts that bypass some of the convolutional layers and forward information from the lower to the higher model layers. This ensures that no information is lost along the network, but also facilitates training of deeper models, i.e., models with a larger number of layers. The added value of the shortcut connections during model training is that they also serve as shortcuts for the back-propagation algorithm used to learn the model parameters and hence make sure that gradients do not vanish down the network. From an architectural point of view, the ResNets are similar to the VGG model and exploit small 3×3 filter stacks in their convolutional layers to keep the number of parameters that need to be learned during training low. In general, residual networks may feature several hundreds of model layers; however, for this work we use the standard ResNet50 architecture.

Residual networks have, to the best of our knowledge, not been used before in the field of ear recognition, but have due to their performance in other areas been selected for our analysis as well. Similar to VGG and SqueezeNet, we use the output of the last layer before the softmax as the $d = 2048$ dimensional feature vector of the ResNet model.

4 Experimental dataset and protocol

For our analysis, we conduct identification experiments on the introduced Annotated Web Ears Extended (AWEx) dataset in accordance with the pipeline described in Sect. 3.1 (see Fig. 1 for details). Our experimental dataset contains 1000 ear images of 100 subjects (with 10 images per subject) and was gathered from the web with a semiautomatic two-step procedure [20, 21, 24]. In the first step, candidate images for the dataset were collected from the web using web crawlers that looked for appropriately tagged imagery on Flickr and Google's image search. The candidate images were then manually screened and curated in the second step to ensure that ears were indeed present in all images. This approach ensured that the appearance variability of the images was not artificially reduced through automatic ear detection techniques and resulted in a challenging dataset of ear images captured in unconstrained settings [24]. This real-life variability is also, to the best of our knowledge, the biggest advantage compared to the other ear datasets available. This ensures that the results of the experiments based on this dataset are to the large extent applicable to real-life scenarios.

The images of the AWE dataset contain ground truth annotations in terms of gender, extent of head pitch, roll and yaw rotations, ethnicity and presence of occlusions and thus provide a perfect starting point for our analysis. The labels/annotations were assigned to the images by a trained annotator and validated by the authors of the dataset. Because the image acquisition procedure was not controlled, each image from the dataset typically exhibits variations across several attributes (e.g., large pitch, roll and yaw angles at the same time) and is annotated with multiple labels, so attribute cross-talk effects need to be taken into account when interpreting the results presented in the results section. The distribution of the individual label categories is presented in Fig. 2.

To assess the impact of the different covariates, we conduct identification experiments with the AWE dataset and report the rank 1 recognition rate (Rank-1) and the (normalized) area under the cumulative match score curves (AUC) when presenting results. For each of the experiments, the probes consist of all images with a specific label (e.g., severe head yaw), while the galleries represent all images from the AWE dataset. With this setup, the gallery size is fixed for all experiments, while the number of probes (and consequently number of conducted identification experiments) depends on the label distribution (shown in Fig. 2) and differs from experiment to experiment. Related labels are merged for the experiments to ensure sufficient numbers of samples for each experiment, e.g., mild head yaw from both left and right, are merged into one group of mild yaw, the same for the severe yaw rotation and the other head rotations (roll and pitch).

For the descriptor-based feature extraction methods, we use the implementations that ship with AWE toolbox and make no change to the default parameters, which are described in detail in [24].

Since the considered deep models need to be trained before they can be exploited for feature extraction, we collect a novel dataset of ear images from the web, using

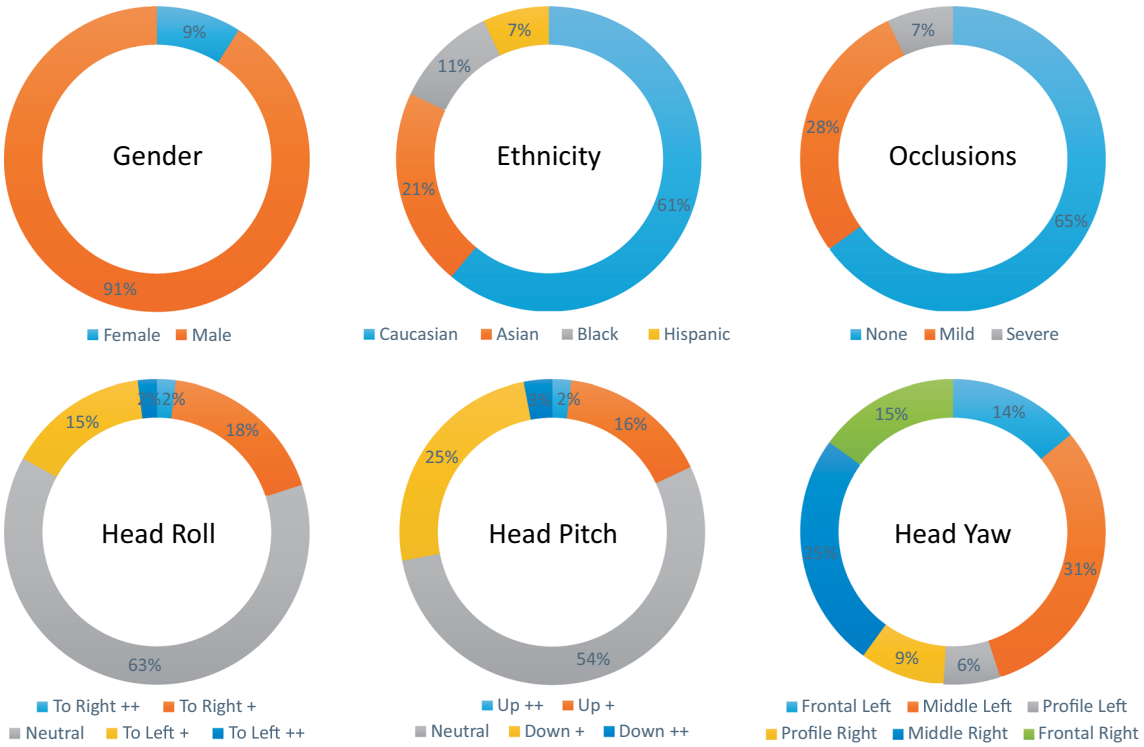

Fig. 2 The graphs show the distribution of covariates (labels) of the images of the AWE dataset. The dataset contains 1000 images of 100 subjects. Gender and ethnicity are labeled on a *per subject* basis, whereas occlusions, head roll, head pitch and head yaw vary for each image in the dataset. Accessories are not shown explicitly here, but from the 1000 AWE images, 91% have no accessories, 8% have some accessories, and 1% (or 9 images) has a significant amount of accessories

the same procedure as used for the AWE dataset. In total, we gather 3104 additional images belonging to 246 subjects that were not present in the original AWE data. We split these images into training (2173 images) and validation sets (931 images) and use the data splits to train our deep models. The models are trained on a desktop PC with Intel Core i7-6700K CPU, 32 GB of RAM and Nvidia Titan Xp until convergence. We set the learning rate to 0.001 and the weights decay rate to 0.001 divided by the number of epochs. To avoid over-fitting, we use a high dropout rate of 0.1 and introduce random perturbations of 100-fold the training data (each image resulted in new 100 perturbated images), where the data transformations are performed (or not) with a 50% chance. Below is a list of the perturbations.

- horizontal flipping,
- trimming 0–10% of images on each side,
- Gaussian blurring with σ 0–3.0,
- addition of Gaussian noise with scale 0–0.2,
- brightness reduction/increase of pixel intensities by a value of 10 (over all color channels or over a single channel),
- contrast increase/decrease of up to 50% (over all color channels or over a single channel),
- rotation of up to 20° in both directions,
- scale increase/decrease of up to 30%.

As an additional contribution, we merge the newly collected images with the original AWE dataset of ear images into the Annotated Web Ears Extended (AWEx) dataset, which now contains a total of 4104 images of 346 subjects captured in completely unconstrained environments and makes the dataset publicly available to the research community through: http://awe.fri.uni-lj.si/. Some example images from the new dataset are shown in Fig. 3.

5 Experiments and results

In this section, we present the results of our analysis. We first describe experiments aimed at analyzing the sensitivity of the recognition approaches toward various covariates, then present a comparative assessment of the tested methods and finally explore the time and space complexity of the recognition models.

5.1 Sensitivity analysis

In our first experiments, we evaluate the sensitivity of all 11 recognition approaches to the following covariate factors: gender, presence of accessories, occlusions, ethnicity,

Fig. 3 Sample images from the AWEx dataset

head pitch, head roll and head yaw. The goal of these experiments is to establish how the recognition models behave in general when confronted with data of different characteristics. We are not interested in the performance and sensitivity of individual models, but in general trends that can be seen over all tested techniques, the sensitivity of individual methods will be the focus of the next section. A comparison of the Rank-1 recognition rates for this series of experiments together with the corresponding AUC values is presented in Fig. 4 and quantitatively in Table 1. Note that the AUC values in Fig. 4 are size-encoded, where a bigger circle indicates a higher AUC value.

The results show that severe head rotations, especially roll, negatively impact the identification performance. Large pitch rotations also have a detrimental effect on performance, whereas yaw angles seem to be less crucial for the performance of the tested methods. These results can be explained through the geometrical properties of the ears, which can are mostly flat (with wrinkles). Since all ear images are usually resized to a fixed input size prior to feature extraction, this pre-procedure (partially) compensates for head rotations that only change the viewing angle of the ears (e.g., yaw), but do not result in orientation changes. To compensate for other head rotations (e.g., roll), additional alignment steps would need to be considered in the recognition pipeline.

Among the considered covariates, gender and ethnicity have the smallest impact on identification performance—the results for all subgroups of these covariates are very close, while the minor performance differences are likely a consequence of the different number of probes in each subgroup. Surprisingly, occlusions which consist mostly of hair have a limited impact on performance. The reason for this, we argue, is that the occlusions are more or less consistent throughout all ear images for a selected subject. The presence of accessories, on the other hand, has a considerable (negative) effect on the recognition performance of all techniques, which again is reasonable, as this type of occlusion varies significantly from image to image.

Although the impact of accessories requires a more in-depth analysis, we presume that the performance drop can be attributed to the fact that samples that fall into this category contain large hearing aids, headphones or some large ornaments, which may not be present in the gallery images. The Rank-1 recognition rates of 0, 4, 11.1 and 22.2% for DSIFT, HOG, LPQ, RILPQ, Gabor wavelets, LBPs, POEM, VGG and RNet need to be interpreted with reservation since only 8 samples were available for this experiment. Nevertheless, we believe that the low performance still shows to a trend with respect to the performance of ear recognition models in the presence of large accessories.

5.2 Comparative evaluation

In the second series of experiments, we compare all considered techniques and explore how different covariates effect the performance of individual recognition models. In Fig. 5, a comparison of the Rank-1 recognition rates is presented for all assessed techniques and all considered covariates in the form of radar graphs. Here a larger area covered by the graphs suggests a better performance.

The graphs show that among the descriptor-based methods DSIF, LPQ and RILPQ are overall slightly inferior to the other methods, which all perform similarly. In terms of robustness, the POEM-based approach seems to be a little more stable than the other techniques as the outline of the graph is most stable with this approach. All in all, the descriptor-based methods exhibit similar behavior when confronted with different covariates pointing to the need to improve the recognition technology in specific areas (e.g., in the presence of large accessories, under severe head rotations).

Among the deep-learning-based model, SNet is clearly the top performer and also the most robust among the CNN-based approaches. The worst-performing deep-learning model is RNet, which is outperformed by the all other tested models. These results can be attributed to the number of parameters that need to learned for the deep models and the still relatively modest (from the perspective of deep learning) amount of ear images available for training. Here, SNet has the clear advantage, as it contains the lowest number of open parameters among the deep models. The CNN models show similar sensitivity

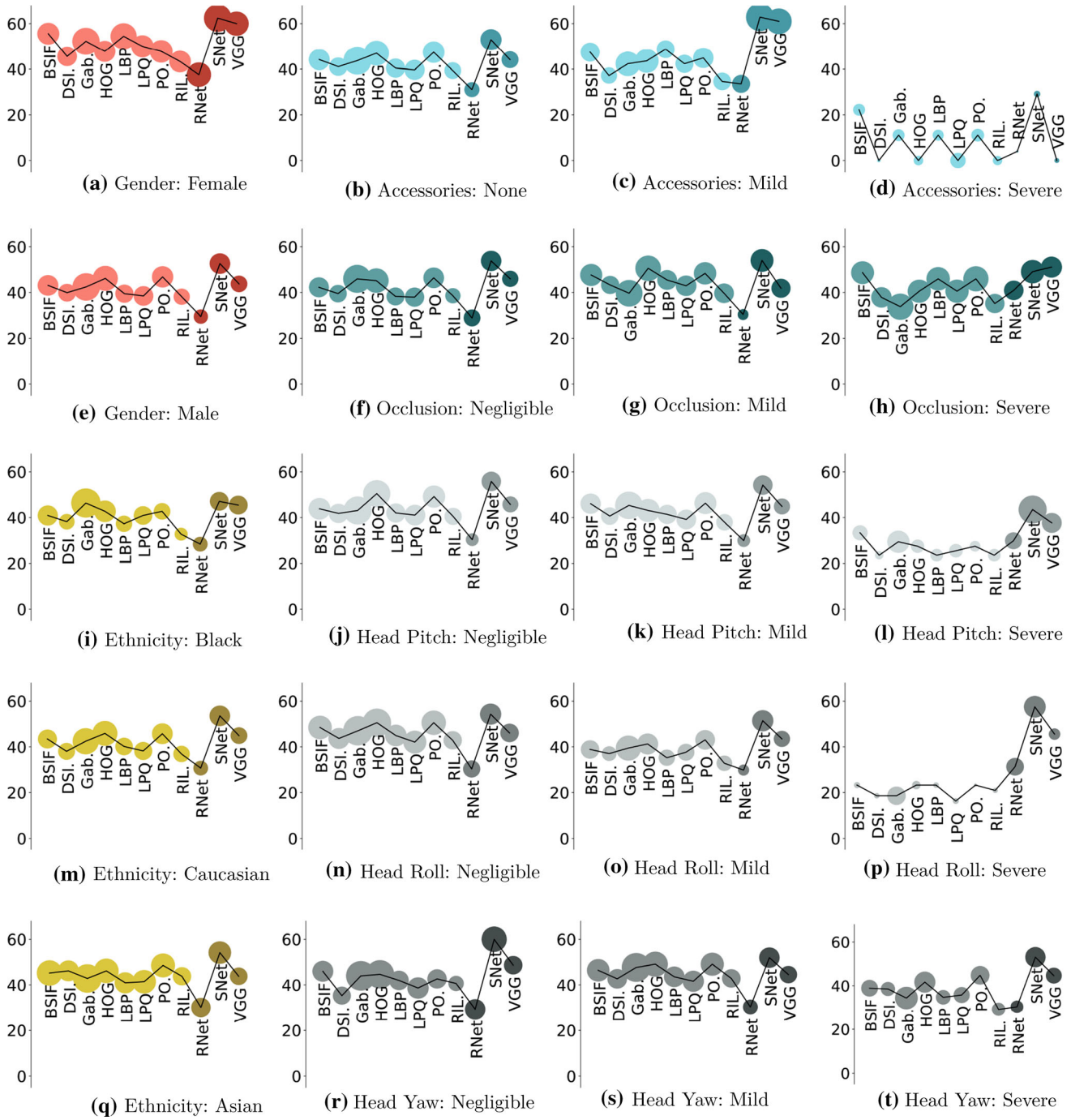

Fig. 4 The performance plots show a comparison of eleven state-of-the-art ear recognition techniques with respect to different covariates. The plots show Rank-1 recognition rates in percent [%] (*y*-axis) and relative AUC values (circle size: where the smallest AUC value among the AUC values was set as the smallest, still visible dot, and the largest value was set as the largest, visually acceptable circle). Due to the small number of subjects for the Hispanic ethnicity subgroup, the results for this subgroup were omitted from the comparison. The small number of images also contributes to the higher Rank-1 recognition rates for women compared to men. Large pitch, roll and yaw (head) rotations show a negative impact on the performance of all assessed techniques. The biggest impact is observed with large accessories, but the results for this test are generated with a small number of probe images. The results are best viewed in color; the darker tones in each plot denote deep-learning-based approaches

characteristics as the descriptor-based methods, but the sensitivity is less pronounced. For example, if we look at the performance differences with respect to head rotations, we can see that small differences can be observed, but these are minimal when compared to the local techniques.

Table 1 Comparative assessment of the eight descriptor-based and three deep-learning-based techniques considered in this work

Perf. metric	Rank-1 (%)											AUC (%)										
Method	BSIF	DSI	Gab	HOG	LBP	LPQ	PO	RIL	RNet	SNet	VGG	BSIF	DSI	Gab	HOG	LBP	LPQ	PO	RIL	RNet	SNet	VGG
Female	55.6	45.6	52.2	47.8	54.4	50.0	47.8	43.3	37.6	62.4	60.0	90.3	88.8	93.4	89.8	92.8	90.6	90.8	90.1	91.8	93.6	91.8
Male	43.1	39.9	42.4	46.2	39.5	38.5	46.8	38.1	29.5	52.6	43.7	89.5	87.5	93.6	92.2	87.8	88.8	89.9	86.3	84.8	89.4	86.7
Asian	45.2	46.2	42.9	46.2	41.0	41.4	48.6	43.8	30.1	54.2	43.8	92.4	89.6	94.1	91.9	89.8	91.3	91.5	87.9	88.6	90.7	87.4
Caucasian	43.4	38.0	42.5	45.9	40.2	38.2	45.7	36.9	30.7	53.6	44.9	88.4	86.9	92.9	92.1	87.4	87.7	89.5	86.5	85.2	89.4	86.7
Black	40.9	38.2	46.4	42.7	37.3	40.9	42.7	32.7	28.4	47.1	45.5	89.1	85.9	94.5	90.1	86.5	88.1	86.7	83.7	85.2	88.2	88.2
Accessories /	44.1	41.1	43.7	47.0	40.4	39.7	47.4	39.3	31.0	52.9	44.1	89.7	87.9	93.8	92.1	88.4	89.1	90.1	86.7	85.5	89.8	86.9
Accessories +	47.4	37.2	42.3	43.6	48.7	42.3	44.9	34.6	33.5	62.9	61.0	88.4	86.3	91.8	91.4	87.6	87.7	89.3	87.4	87.7	93.8	92.6
Accessories ++	22.2	0.0	11.1	0.0	11.1	0.0	11.1	0.0	4.0	29.1	0.0	82.9	72.3	82.8	80.5	82.0	85.7	83.6	80.3	69.7	76.9	74.8
Pitch /	43.8	41.8	43.1	50.5	42.0	41.1	49.2	40.5	30.3	55.8	45.8	90.2	88.6	93.8	93.4	88.5	89.5	90.5	87.0	83.6	88.7	86.2
Pitch +	46.1	40.6	45.3	43.1	41.4	39.2	46.3	37.9	29.9	54.2	44.9	89.2	87.3	93.6	91.1	88.5	88.8	90.3	86.6	83.6	88.7	86.3
Pitch ++	33.3	23.5	29.4	27.5	23.5	25.5	27.5	23.5	29.8	43.5	37.7	85.5	79.3	90.5	84.0	83.5	84.3	81.7	83.4	86.9	93.9	89.0
Roll /	48.5	43.7	47.0	50.6	45.0	42.1	50.6	42.9	30.3	54.3	46.1	91.3	89.7	94.6	94.0	90.1	90.9	91.8	87.9	87.1	89.8	88.0
Roll +	38.9	37.1	39.5	41.3	35.2	37.7	43.1	32.8	29.8	51.4	43.5	88.0	85.2	92.2	89.7	86.4	86.9	88.8	85.8	82.1	89.8	86.4
Roll ++	23.3	18.6	18.6	23.3	23.3	16.3	23.3	20.9	31.3	57.5	45.6	76.5	75.6	88.0	79.4	75.8	76.3	72.0	75.2	87.4	90.2	82.5
Yaw /	46.0	35.3	44.0	44.7	42.0	38.7	42.7	40.7	29.3	60.0	48.7	89.6	87.7	94.7	94.1	88.7	90.0	88.6	85.7	89.3	92.2	88.0
Yaw +	46.4	42.7	47.7	49.1	43.6	41.6	49.1	42.9	30.4	52.0	44.6	90.7	88.8	94.4	92.3	89.5	89.8	91.0	88.0	85.3	89.3	87.3
Yaw ++	38.9	38.5	34.4	41.7	34.7	35.8	44.8	29.2	30.3	53.1	44.7	87.3	85.2	91.3	90.3	85.6	86.8	88.7	84.5	84.0	89.8	86.7
Occlusion /	42.2	39.5	45.9	45.2	38.3	37.9	46.4	38.6	28.9	53.8	46.0	89.1	87.3	93.8	91.9	87.6	88.5	89.5	85.6	86.7	89.3	86.4
Occlusion +	47.6	43.3	39.6	50.6	45.5	42.9	48.4	39.6	30.4	54.0	41.8	90.4	87.8	93.1	92.5	88.9	89.4	90.6	88.8	81.7	90.9	88.4
Occlusion ++	48.7	37.8	33.8	40.5	46.0	40.5	46.0	35.1	40.9	49.1	51.1	90.9	89.3	92.8	91.1	91.2	91.1	92.3	88.6	88.7	91.2	89.7

Results were generated on the whole AWE dataset of 1000 images of 100 subjects in the form of Rank-1 recognition rates, and the area under the cumulative match score curve (AUC). All results are given in percentages. *RIL.*, *PO.*, *DSI.*, *Gab.*, *RNet* and *SNet* denote RILPQ, POEM, DSIFT, Gabor, ResNet and SqueezeNet, respectively

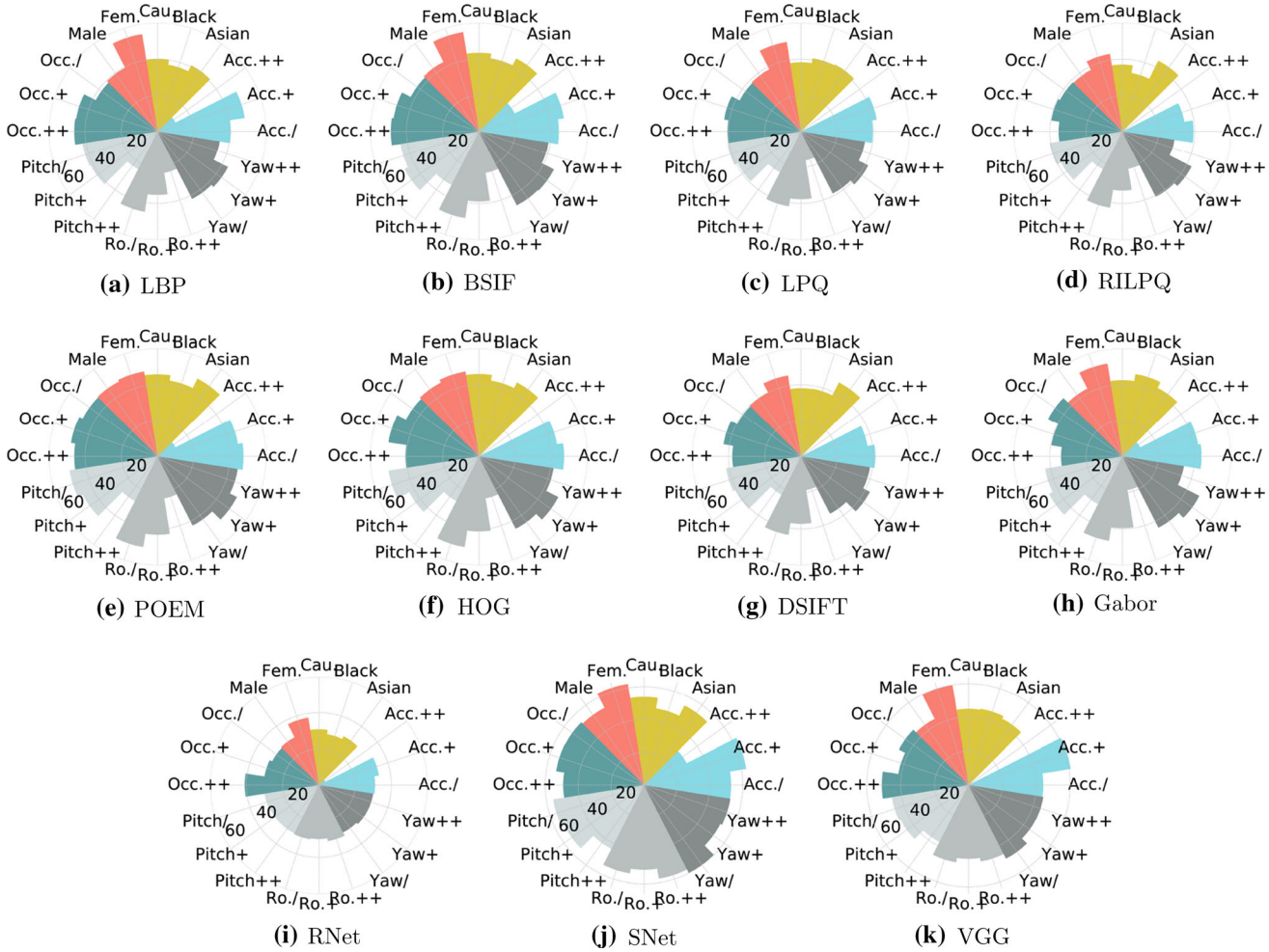

Fig. 5 The radar graphs show a comparison of the Rank-1 recognition rates of the evaluated recognition methods with respect to different covariates. The axes show values from 0 to 60%. *Acc.*, *Ro.*, *Occ.*, *Fem.* and *Cau.* denote accuracy, roll, occlusion, female and Caucasian, respectively. For a description of the labels, please refer to Fig. 2 and the description in [24]. The graphs are best viewed in color

When comparing descriptor-based methods to deep-learning-based models, we can see the overall winner of the comparison i SNet, but the worst-performing model is gain a CNN suggesting that deep models are competitive but need sufficient data to be trained effectively or feature a sufficiently small number of parameters that need to be learned.

5.3 Time and space complexity

Last but not least, we compare and analyze the time and space complexity of the tested recognition techniques in Fig. 6 and Table 2. Here, Fig. 6 (left) shows a comparison of the average time needed to process one image vs. the achieved recognition performance. The average time was computed over the entire test set of 1000 AWE images. For the descriptor-based methods, the processing time was computed by running the experiments on the CPU of our experimental hardware, whereas for the deep-learning-based models the experiments were conducted on the GPU. Ideally, a fast and efficient method should be as close as possible to the *x*-axis and as far away from the *y*-axis as possible. As we can see, the average time for all methods is similar; a clear outlier here is the Gabor wavelet technique, which takes significantly more time on average than the competing methods. SNet is again the best approach in this comparison, as it is reasonably fast, but with the highest performance.

In Fig. 6, we see a similar comparison, but here the feature vector length is plotted against the Rank-1 recognition performance. As we can see, the clear outlier this time is SNet, which generates a significantly larger feature vector than the remaining techniques. Thus, when this is

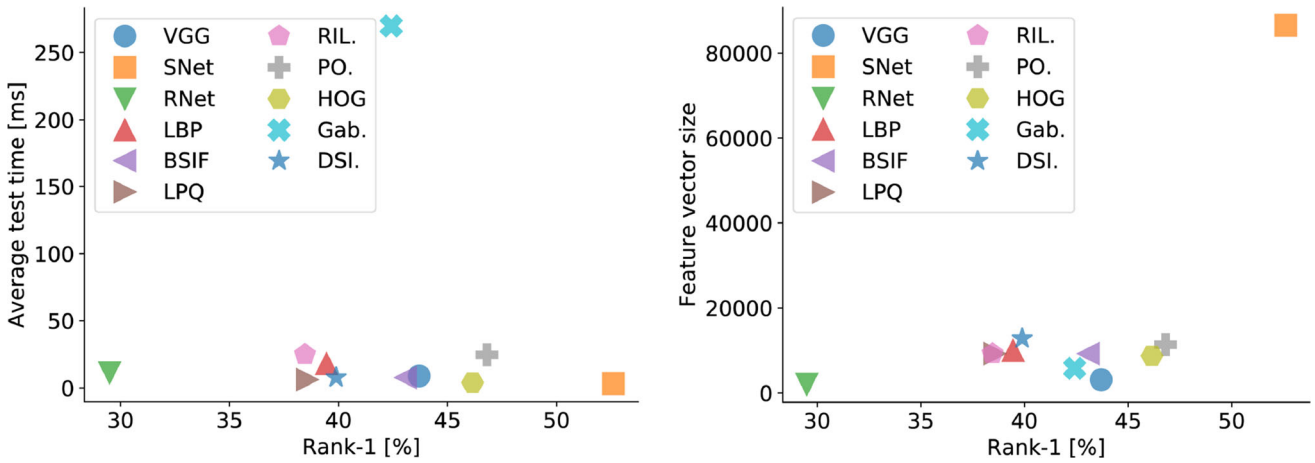

Fig. 6 Time and space complexity vs. the recognition performance. Left: The average test time is plotted versus the Rank-1 recognition rate. The closer a methods is located to the lower right corner, the better the characteristics with respect to the time complexity. Right: The feature vector size is plotted against the Rank-1 recognition rate. The closer a methods is located to the lower right corner, the better the characteristics with respect to the space complexity. Best viewed in color

Table 2 Time and space complexity

Method	Model size (in MB)	# Parameters to train	Feature vector size	Training time (in h)	Average test time—per image (in ms)
BSIF	0	0	9216	0	8
DSIFT	0	0	12,800	0	8
Gabor	0	0	5760	0	270
HOG	0	0	8712	0	4
LBP	0	0	9971	0	18
LPQ	0	0	9216	0	6
POEM	0	0	11,328	0	25
RILPQ	0	0	9216	0	25
RNet	96.8	25,636,712	2047	~ 18	11
SNet	3.5	1,235,496	86,527	~ 8	3
VGG	541.1	117,479,232	4095	~ 12	9

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

problematic due to the limited availability of resources, alternative models would need to be sought, despite the best overall performance of SNet so far.

When look at the information presented in Table 2, we notice a striking difference between descriptor-based and deep-learning-based methods, i.e., the model size that needs to be stored in RAM is in the range of MBs, with the biggest model, VGG, requiring a total 541.1 MB just to load the model. The cost for the descriptor-based methods, on the other hand, is 0 as far as memory requirements is concerned. Descriptor-based methods also do not require any training procedure (their training time is 0) and can be applied to ear recognition problems without the need for enormous amounts of training data. Hence, if training data are scarce, descriptor-based methods may still have an edge over deep-learning-based models, where typically millions of parameters need to be learned during training (that takes several hours or even days depending on the hardware and amount of training data). As we already pointed out above, the feature vector size is comparable among all methods (except for SNet), but serves here only for illustrative purposes to show the approximate magnitude of the vector sizes. In general, the sizes can vary depending on the choice of open parameters of the recognition techniques considered.

6 Conclusion

We have evaluated eight popular dense descriptor-based feature extraction methods for ear recognition and three popular approaches based on convolutional neural networks and analyzed their performance with respect to different covariates. The results show that gender and ethnicity with some exceptions do not impact identification performance significantly. However, severe head rotations and severe use of accessories all negatively impact recognition performance. Furthermore, we showed that hair occlusions negatively impact performance to a much more limited extent than other factors. The reason for this, we argue, is that hair that belongs to a specific person is similar throughout all (or most) ear images. We found that the tested methods differ significantly in terms of time and space complexity and that in situations where resources are scarce, descriptor-based methods may be favored over CNN models, despite their slightly inferior performance.

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

Compliance with ethical standards

Conflict of interest We warrant that the article has not received prior publication, is not under consideration for publication elsewhere and is an original work. On behalf of all co-authors, the corresponding author bears full responsibility for the submission. The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest.

References

1. Abaza A, Ross A, Hebert C, Harrison MAF, Nixon M (2013) A survey on ear biometrics. ACM Comput Surv 45(2):1–22
2. Alaraj M, Hou J, Fukami T (2010) A neural network based human identification framework using ear images. In: Proceedings of the international technical conference of IEEE region 10, pp 1595–1600
3. Arbab-Zavar B, Nixon MS (2008) Robust log-Gabor filter for ear biometrics. In: Proceedings of the international conference on pattern recognition, pp 1–4
4. Baoqing Z, Zhichun M, Chen J, Jiyuan D (2013) A robust algorithm for ear recognition under partial occlusion. In: Proceedings of the Chinese control conference, pp 3800–3804
5. Basit A, Shoaib M (2014) A human ear recognition method using nonlinear curvelet feature subspace. Int J Comput Math 91(3):616–624
6. Benzaoui A, Hezil N, Boukrouche A (2015) Identity recognition based on the external shape of the human ear. In: Proceedings of the international conference on applied research in computer science and engineering, pp 1–5
7. Benzaoui A, Kheider A, Boukrouche A (2015) Ear description and recognition using ELBP and wavelets. In: Proceedings of the international conference on applied research in computer science and engineering, pp 1–6
8. Bourouba H, Doghmane H, Benzaoui A, Boukrouche AH (2015) Ear recognition based on Multi-bags-of-features histogram. In: Proceedings of the international conference on control, engineering information technology, pp 1–6
9. Bustard JD, Nixon MS (2010) Toward unconstrained ear recognition from two-dimensional images. Trans Syst Man Cybern Part A Syst Hum 40(3):486–494
10. Chan T-S, Kumar A (2012) Reliable ear identification using 2-D quadrature filters. Pattern Recogn Lett 33(14):1870–1881
11. Choraś M (2008) Perspective methods of human identification: ear biometrics. Opto-Electron Rev 16(1):85–96
12. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the international conference on computer vision and pattern recognition, pp 886–893
13. Damar N, Fuhrer B (2012) Ear recognition using multi-scale histogram of oriented gradients. In: Proceedings of the conference on intelligent information hiding and multimedia signal processing, pp 21–24
14. Daugman J (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A 2(7):1160–1169
15. Dewi K, Yahagi T (2006) Ear photo recognition using scale invariant keypoints. In: Proceedings of the computational intelligence, pp 253–258
16. Dodge S, Mounsef J, Karam L (2018) Unconstrained ear recognition using deep neural networks. IET Biom 7:207–214
17. Eyiokur FI, Yaman D, Ekenel HK (2018) Domain adaptation for ear recognition using deep convolutional neural networks. IET Biom 7:199–206
18. Earnest H, Segundo P, Sarkar S (2018) Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biom 7:215–223
19. Emeršič Ž, Meden B, Peer P, Štruc V (2017) Covariate analysis of descriptor-based ear recognition techniques. In: 2017 international conference and workshop on bioinspired intelligence (IWOBI), pp 1–9
20. Emeršič Ž, Peer P (2015) Ear biometric database in the wild. In: 2015 4th international work conference on bioinspired intelligence (IWOBI), pp 27–32
21. Emeršič Ž, Peer P (2015) Toolbox for ear biometric recognition evaluation. In: EUROCON 2015—international conference on computer as a tool (EUROCON), IEEE, pp 1–6
22. Emeršič Ž, Štepec D, Štruc V, Peer P (2017) Training convolutional neural networks with limited training data for ear recognition in the wild. In: Proceedings of the 12th IEEE international conference on automatic face and gesture (FG 2017)
23. Emeršič Ž, Štepec D, Štruc V, Peer P, George A, Ahmad A, Omar E, Boult TE, Safdari R, Zhou Y, Zafeiriou S, Yaman D, Eyiokur FI, Ekenel HK (2017) The unconstrained ear recognition challenge. In: International joint conference on biometrics (IJCB)
24. Emeršič Ž, Štruc V, Peer P (2017) Ear recognition: more than a survey. Neurocomputing 255:26–39
25. Grm K, Štruc V, Artiges A, Caron M, Ekenel HK (2017) Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biom 7:81–89
26. Guo Y, Xu Z (2008) Ear recognition using a new local matching approach. In: Proceedings of the international conference on image processing, pp 289–292
27. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
28. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: European conference on computer vision. Springer, Berlin, pp 630–645

29. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. arXiv preprint arXiv:1602.07360,
30. Kannala J, Rahtu E (2012) BSIF: Binarized statistical image features. In: Proceedings of the international conference on pattern recognition, pp 1363–1366
31. Križaj J, Štruc V, Pavešic N (2010) Adaptation of SIFT features for robust face recognition. In: Proceedings of the image analysis and recognition. Springer, New York, pp 394–404
32. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
33. Kumar A, Wu C (2012) Automated human identification using ear imaging. Pattern Recogn 45(3):956–968
34. Kumar A, Zhang D (2007) Ear authentication using log-gabor wavelets. In: Proceedings of the symposium on defense and security. International society for optics and photonics, p 65390A
35. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
36. Meraoumia A, Chitroub S, Bouridane A (2015) An automated ear identification system using Gabor filter responses. In: Proceedings of the international conference on new circuits and systems, pp 1–4
37. Morales A, Ferrer M, Diaz-Cabrera M, Gonzalez E (2013) Analysis of local descriptors features and its robustness applied to ear recognition. In: Proceedings of the international carnahan conference on security technology, pp 1–5
38. Nanni L, Lumini A (2009) Fusion of color spaces for ear authentication. Pattern Recogn 42(9):1906–1913
39. Ojansivu V, Heikkilä J (2008) Blur insensitive texture classification using local phase quantization. In: Image and signal processing, Springer, New York, pp 236–243
40. Ojansivu V, Rahtu E, Heikkilä J (2008) Rotation invariant local phase quantization for blur insensitive texture analysis. In: Proceedings of the international conference on pattern recognition, pp 1–4
41. Pflug A, Busch C (2012) Ear biometrics: a survey of detection, feature extraction and recognition methods. Biometrics 1(2):114–129
42. Pflug A, Busch C, Ross A (2014) 2D ear classification based on unsupervised clustering. In: Proceedings of the international joint conference on biometrics, pp 1–8
43. Pflug A, Paul PN, Busch C (2014) A comparative study on texture and surface descriptors for ear biometrics. In: Proceedings of the international Carnahan conference on security technology, pp 1–6
44. Pflug A, Wagner J, Rathgeb C, Busch C (2014) Impact of severe signal degradation on ear recognition performance. In: 2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp 1342–1347
45. Pietikäinen M, Hadid A, Zhao G, Ahonen T (2011) Computer vision using local binary patterns. Computational imaging and vision. Springer, New York
46. Prakash S, Gupta P (2013) An efficient ear recognition technique invariant to illumination and pose. Telecommun Syst 52(3):1435–1448
47. Purkait R (2015) Role of external ear in establishing personal identity—a short review. Austin J Forensic Sci Criminol 2(2):1–5
48. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
49. Štruc V, Gajšek R, Pavesic N (2009) Principal Gabor filters for face recognition. In: Proceedings of the conference on biometrics: theory, applications and systems, pp 1–6
50. Štruc V, Pavesic N (2009) Gabor-based kernel partial-least-squares discrimination features for face recognition. EURASIP J Adv Signal Process 20(1):115–138
51. Štruc V, Pavešic N (2010) The complete gabor-fisher classifier for robust face recognition. EURASIP J Adv Signal Process 1–26:2010
52. Vu N-S, Caplier A (2010) Face recognition with patterns of oriented edge magnitudes. In: European conference on computer vision, pp 313–326
53. Xiaoyun W, Weiqi Y (2009) Human ear recognition based on block segmentation. In: Proceedings of the international conference on cyber-enabled distributed computing and knowledge discovery, pp 262–266
54. Xie Z, Mu Z (2008) Ear recognition using LLE and IDLLE algorithm. In: Proceedings of the international conference on pattern recognition, pp 1–4
55. Zhang Y, Mu Z, Yuan L, Yu C (2018) Ear verification under uncontrolled conditions with convolutional neural networks. IET Biom 7. https://ieeexplore.ieee.org/abstract/document/8340919/
56. Zhang Z, Liu H (2008) Multi-view ear recognition based on B-Spline pose manifold construction. In: Proceedings of the world congress on intelligent control and automation

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com