The Unconstrained Ear Recognition Challenge

Žiga Emeršič, Dejan Štepec, Vitomir Štruc, Peter Peer
University of Ljubljana
Ljubljana, Slovenia
ziga.emersic@fri.uni-lj.si

Anjith George
IIT Kharagpur
Kharagpur, India

Adil Ahmad, Elshibani Omar, Terrance E. Boult
University of Colorado Colorado Springs
Colorado Springs, CO, USA

Reza Safdari
Islamic Azad University
Qazvin, Iran

Yuxiang Zhou, Stefanos Zafeiriou
Imperial College London
London, UK

Dogucan Yaman, Fevziye I. Eyiokur, Hazim K. Ekenel
ITU Department of Computer Engineering
Istanbul, Turkey

◆

Abstract—In this paper we present the results of the Unconstrained Ear Recognition Challenge (UERC), a group benchmarking effort centered around the problem of person recognition from ear images captured in uncontrolled conditions. The goal of the challenge was to assess the performance of existing ear recognition techniques on a challenging large-scale dataset and identify open problems that need to be addressed in the future. Five groups from three continents participated in the challenge and contributed six ear recognition techniques for the evaluation, while multiple baselines were made available for the challenge by the UERC organizers. A comprehensive analysis was conducted with all participating approaches addressing essential research questions pertaining to the sensitivity of the technology to head rotation, flipping, gallery size, large-scale recognition and others. The top performer of the UERC was found to ensure robust performance on a smaller part of the dataset (with 180 subjects) regardless of image characteristics, but still exhibited a significant performance drop when the entire dataset comprising $3,704$ subjects was used for testing.

Fig. 1. Sample images from the UERC dataset. The images are tightly cropped and belong to left and right ears. The variability of the dataset is across ear orientation, occlusion, gender, ethnicity and resolution. Images in one row belong to the same subject.

1 INTRODUCTION

Recognizing people from ear images with automatic machine-learning techniques represents a challenging problem that is of interest to numerous application domains. Past research in this area has mostly been focused on images captured in controlled conditions and near perfect recognition performance has already been reported on many of the available (laboratory-like) ear datasets, e.g., [6], [18], [21], [29], [40]. The literature on unconstrained ear recognition, on the other hand, is relatively modest

and the performance of existing ear recognition techniques on so-called data captured *in the wild* is not well explored. It is not completely clear how existing techniques are able to cope with ear-image variability encountered in unconstrained settings and how the recognition performance is affected by factors such as head rotation, image resolution, occlusion or gallery size.

To address this gap and study the problem of ear recognition from unconstrained ear images, the Unconstrained Ear Recognition Challenge (UERC), a first of its kind group benchmarking effort, was organized in the scope of the 3rd International Joint Conference on Biometrics (IJCB 2017). The goal of the challenge was to exploit the combined expertise of multiple research groups for a comprehensive evaluation of ear recognition technology,

to consolidate research and identify open challenges through experiments on a common dataset and a well defined experimental protocol. As illustrated in Fig. 1, a large and challenging dataset of ear images captured in unconstrained settings was collected specifically for the UERC and made available to the participants for algorithm development and testing. Five groups (from the US, UK, Turkey, Iran and India) took part in the challenge and submitted results for a total of six recognition approaches. A detailed analysis was then conducted investigating various aspects of the submitted techniques, such as sensitivity to head rotation (separately for pitch, roll and yaw angles), impact of gallery size and the ability of the techniques to scale with larger probe and gallery sets.

The research and development efforts of the UERC organizers and participating research groups resulted in the following contributions that are presented in this paper:

- The first group benchmarking effort of ear recognition technology on a large datasets of unconstrained ear images corresponding to several thousands of subjects.
- A new and challenging dataset of ear images collected from the internet, which (with a total of $11,804$ images of $3,706$ identities) is among the largest ear datasets publicly available to the research community.
- An analysis of the characteristics of different ear recognition techniques and identification of the main factors affecting performance.

2 RELATED WORK

Challenges and group evaluations have a long history in the field of computer vision and biometrics in particular. The goal of these events is to benchmark existing work in a certain problem domain, provide a snapshot of the current state-of-technology and point to open issues that need to be addressed in the future. As a result of these challenges, significant advancements have been made over the years that pushed the capabilities of computer-vision technology.

Examples of recent challenges that had a profound impact on the field of computer vision are the ImageNet Large Scale Visual Recognition Challenges (ILSVRC) [11], [43], which focus on image classification and object localization problems, the Visual Object Tracking (VOT) [25], [26], [27] challenges that aim at evaluating various solutions to object tracking in videos, and the ChaLearn Looking at People [16], [17], [51] series of challenges, where human-centric vision problems are at the center of attention.

Among past biometrics-oriented challenges, events focusing on fingerprints [31], [32], [33] and facial images [3], [22], [41], [44] have likely been the most visible. Recent years have seen challenges and competitions on other modalities and biometric sub-problems ranging from iris [4], speaker [20], [34], sclera [9], finger-vein [52], [55], or keystroke dynamics [36] recognition to spoof detection [5], [7], [49], liveness detection [19], [35], [53], segmentation [9], [10] and others.

Here, we add to the outlined body of work and present the results of the (first of its kind) ear recognition challenge. The challenge focuses on the under-explored problem of ear recognition from images captured in unconstrained environments and aims at presenting a comprehensive analysis of the existing ear recognition technology and at providing the community with a new dataset for research in this area.

3 METHODOLOGY

In this section we describe the methodology used for the UERC. We first introduce the experimental dataset and then present the experimental protocol and performance metrics used for the evaluation.

3.1 The UERC dataset

The data used for the UERC is a blend of two existing and one newly collected dataset and contains a total of $11,804$ ear images of $3,706$ subjects. The core part of the data comes from the Annotated Web Ears (AWE) [15] dataset and features $3,300$ ear images of 330 subjects. Images from this part contain various annotations, such as the extent of head rotation (in terms of pitch, yaw and roll angles), gender, level of occlusion, ethnicity and others. A detailed description of the AWE dataset is available from [15]. The second part of the UERC data, i.e., 804 images of 16 subjects, comes from the auxiliary AWE dataset[1], while the majority of images, i.e., $9,500$ images of $3,540$ subjects, was gathered exclusively for the UERC.

The data-collection procedure for the newly gathered images was semi-automatic. Similarly to other datasets gathered *in the wild* [22], [24], a list of celebrities was first generated and web crawlers were used to pull candidate face images from the internet. An automatic ear segmentation procedure based on convolutional encoder-decoder networks [12] was then applied to the candidate images to identify potential ear regions. Finally, the segmented ears were manually inspected and miss detections and partial segmentations were discarded. The final images included in the UERC dataset were tightly cropped, but were not normalized to a common side. Thus, original images of left and right ears are featured in the dataset. A major characteristic of the UERC images is the large variability in size, since the smallest images containing only a few hundred pixels, whereas the largest contain close to $400k$ pixels. The average pixel count per image is $3,682$.

Some examples of the UERC ear images are shown in Fig. 1. As can be seen, the images were not captured in controlled (laboratory-like) environments, the variability of the images is, therefore, substantial. The UERC dataset represents, to the best of our knowledge, the first attempt to gather data for ear-recognition research from the web, and as shown in [15] is significantly more challenging than competing ear datasets where performance is mostly saturated.

3.2 Protocol and performance metrics

For the evaluation, the UERC ear images were split into two disjoint groups, one for training and one for testing. The training part contained images from the auxiliary AWE dataset and around half of the original AWE data, whereas the testing part comprised the other half of the AWE data and all newly collected images. A summary of the data split is given in Table 1.

The training part of the UERC data was used to train or fine-tune potential models (e.g., deep models, classifiers, etc.), while the testing part was used exclusively for the performance evaluation. Using any images from the testing part for training was not allowed. The UERC participants were asked to submit a similarity matrix of size $7,742 \times 9,500$ to the organizers, which served as the basis for the performance assessment. The similarity matrix was produced by matching the $7,742$ probe images (of

---

1. Available from http://awe.fri.uni-lj.si/download.

TABLE 1
Overview of the UERC data partitioning. The annotated part of the UERC data, i.e., the AWE dataset, was split between the training and testing parts. The partitioning was disjoint in terms of subjects.

UERC data partition	Image origin	# Images	# Subjects	# Images per subject	Total # Images (# Subjects)
Training part	Auxiliary AWE dataset	804	16	Variable	2, 304 (166)
	Part of AWE dataset	1, 500	150	10	
Testing part	Part of AWE dataset	1, 800	180	10	9, 500 (3, 540)
	Newly collected	7, 700	3, 360	Variable	

$1,482$ subjects) to all $9,500$ gallery images (belonging to 3540 subjects). The gallery featured all images from the testing part of the UERC data, while the $7,742$ probe images represented a subset of these $9,500$ galleries from subjects having at least 2 images in the dataset. Each similarity score had to be generated based solely on the comparison of two ear images and no information about other subjects in the testing part of the UERC data was allowed to be used for the score generation. A sample script that implemented the protocol was distributed among the UERC participants to ensure that all submitted similarity matrices were generated consistently with the same training and testing data.

The performance of the submitted approaches was measured through recognition (identification) experiments and Cumulative Match Score (CMC) curves were used to visualize the results. Additionally, performance was measured with the following quantitative metrics: *i)* the recognition rate at rank one (rank-1), which corresponds to the percentage of probe images, for which an image of the correct identity was retrieved from the gallery as the top match, *ii)* the recognition rate at rank five (rank-5), which corresponds to the percentage of probe images, for which an image of the correct identity was among the top five matches retrieved from the gallery, and *iii)* the area under the CMC curve (AUC), which analogous to the more widely used Receiver Operating Characteristic (ROC) AUC, measured the overall performance of the recognition approached at different ranks. The latter was computed by normalizing the maximum rank (i.e., the number of distinct gallery identities) of the experiments to one.

To study the performance of the submitted approaches and analyze their characteristics, specific parts corresponding to different data labels (e.g., probe images with specific yaw, roll or pitch angles) were sampled from the similarity matrix and evaluated.

4 PARTICIPATING APPROACHES

In this section we describe all participating approaches of the UERC. Five groups submitted results and several baselines were made available by the challenge organizers.

4.1 Baseline approaches

Eight baseline approaches were provided for the UERC by the organizers from the University of Ljubljana with the goal of ensuring references implementations of the experimental protocol and an initial estimate of the difficulty of the dataset. Two of these baselines were then selected for most of the experimental analysis in Section 5 to keep the results uncluttered. The first was a descriptor-based technique exploiting local binary patterns [42] (LBP-baseline) and the second a deep learning approach based on the 16-layer VGG network architecture [39] (VGG-baseline). The baselines were implemented within the UERC toolkit and

distributed among the participants as a starting point for their research work. A detailed description of the two selected baselines is given below.

LBP-baseline: The first UERC baseline included in the results section follows the usual pipeline used to compute image descriptors based on local binary patterns (LBPs) [42]. Patches of size 16×16 pixels are first sampled from the images using a sliding window approach and a step size of 4 pixels. Each patch is then encoded with uniform LBPs computed with a radius of $R = 2$ and a local neighborhood size of $P = 8$. 59-dimensional histograms are calculated for each patch and the histograms of all image patches are concatenated to form the final ear descriptor. The similarity of two ear images is measured with the cosine similarity between the corresponding LBP descriptors.

VGG-baseline: The second selected UERC baseline is built around a deep learning model, specifically, around a convolution neural network (CNN) based on the 16-layer VGG architecture from [46]. The model consists of multiple convolutional layers which, different from competing models, e.g., [28], use small filters (of size 3 pixels) to reduce the number of parameters that need to be learned during training. The convolutional layers are interspersed with max-pooling layers that reduce the size of the intermediate image representations and followed by two fully-connected layer. The output of second fully-connected layer is used as a $4,096$-dimensioanl ear descriptor. The VGG-baseline model is trained from scratch using the training data from the UERC dataset and aggressive data augmentation. A detailed description of the training procedure is presented in [14]. Two ear images are matched by computing the cosine distance between the corresponding image descriptors.

4.2 University of Colorado Colorado Springs

The group from the University of Colorado Colorado Springs (UCCS) approached the challenge with a novel ear descriptor based on *Chainlets* and utilized a recent deep-learning-based approach for contour detection [45] to facilitate the descriptor computation procedure.

The UCCS approach starts with a preprocessing procedure (illustrated in Fig. 2) aimed at detecting the ear region in the image. With the preprocessing procedure the input image is first converted to gray-scale and resized to 100×100 pixels. Next, contrast limited adaptive histogram equalization (CLAHE [57]) is applied for contrast enhancement and a binary version of the image is produced through intensity thresholding. Morphological operations such as dilation and opening are employed to remove noise and to help accentuate the structural information of the ear. The processed binary image is then analyzed and the largest connected region is selected as the ear mask and used to exclude all background pixels from the image that could adversely affect

Fig. 2. Illustration of the UCCS preprocessing procedure (from left to right): the input image, the computed binary ear mask, the masked input image, the input image without accessories and occlusions. The preprocessing aims at removing non-ear regions that could adversely affect the descriptor computation process.

the descriptor computation procedure. To also remove potential earrings and other accessories left in the image, color segmentation focusing on skin-tone values in the HSV color model is used. The result of this procedure is a clean region-of-interest as shown in the right most image of Fig. 2.

Once the the input image is segmented and the region-of-interest is detected, the UCCS approach proceeds with the descriptor calculation step and computes a chainlet-based image descriptor from the cleaned ear area. The main idea here is that the appearance and shape of an object can be well described by the density of Relative Chain Codes, which encode rotation-invariant edge directions. The descriptor is similar in essence to HOG [8] but relies on longer connected edges and provides a richer and rotation invariant description of edge orientation.

To compute chainlets, image pixels are first grouped into cells, and the direction of each pixel is computed through a Relative Chain Code [2]. In the case of ear recognition, cells of size 8×8 pixels are selected. Also, computation of a Relative Chain Code Histogram (CCH) is done over the code directions within each cell. To ensure invariance to contrast changes, neighboring cells are grouped into "blocks" and the CCHs of all cells from a given block are jointly normalized. The normalization is performed for all image blocks in a sliding window (or better said sliding block) manner. The chainlets descriptor is finally formed by concatenating the normalized CCHs from all (overlapping) image blocks. To measure the similarity of two ear images required for the UERC evaluation procedure, the chi-square distance between the corresponding chainlets descriptors is calculated. A detailed description of the UCCS approach is available from [2].

4.3 Islamic Azad University

The group from the Islamic Azad University (IAU) participated in the challenge with an ear recognition approach exploiting the 16-layer VGG network from [46] and transfer learning. The approach is similar to the UERC VGG-baseline but relies on weights learned from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2014). The main idea here is to keep part of the pretrained VGG model in tact, while retraining other parts that are relevant for the new problem domain, i.e., ear recognition.

The amount of training data available in the UERC dataset is relatively modest and differs from ImageNet both in terms of image content as well the number of classes. It might, hence, not be optimal to train a classifier on top of the network, as image representations generated by the higher network layers are commonly problem- and dataset-specific. A potentially better approach is to rely on activations from earlier network layers and use these to train a linear classifier for the new problem domain. The IAU approach, therefore, adds two fully-connected layers on top of the

7th layer of the pretrained VGG model. The pretrained weights of the early layers are frozen and kept unchanged, while the newly added fully-connected layers are trained from scratch with the UERC training data using stochastic gradient descend (SGD) and the softmax loss function. In order to prevent overfitting, a dropout rate of 0.7 and L_2 weight decay regularization are applied on the fully-connected layers [47]. Additionally, data augmentation [28] is used to increase the amount of the data available for training. After convergence, the last fully-connected layer (i.e, the classifier) is removed and the output of the penultimate layer is used an image descriptor.

Once the model is fully trained, it is used as a "black-box" feature extractor. A given ear image is simply rescaled to a size of 64×64 pixels, mean centered and passed through the network. The output of the model is a $512-$dimensional ear descriptor that can be matched against other ear descriptors using the cosine similarity measure.

4.4 Imperial College London

The group from the Imperial College London (ICL) participated in the UERC with an approach build around Statistical Deformable Models (SDMs) [56] and Inception-ResNets [48]. The SDM was used for dense ear alignment and the Inception-ResNet for descriptor computation.

For the ear-alignment model the ICL group used an in-house dataset of 605 ear images annotated with 55 landmarks[2]. A specially designed SDM, capable of handling training data with inconsistent annotations, was then trained with the annotated ears and used to densely align all images from the UERC dataset. During the alignment step, the ear images were flipped and the deformable model was fitted to the original as well as flipped images. The image that resulted in a lower loss during model fitting was chosen as the basis for descriptor computation.

In the feature-learning part of the ICL approach, an Inception-ResNet [48] was trained from scratch using aligned ear images from the training part of the UERC dataset. The Inception-ResNet architecture was chosen for this part because of its competitive performance and the ability to be trained fast due to the residual connections of the model. The network was trained with a marginal- and softmax-cross-entropy loss for 80 epochs. After the training, the classification layer was removed and the output of the model was used as a 512-dimensional ear descriptor.

In the evaluation stage, the probe images were first aligned with the trained SDM and ear descriptors were computed with the Inception-ResNet model. A similarity score for two images was produced based on the L_2-norm between the corresponding ear descriptors.

4.5 Indian Institute of Technology Kharagpur

The group from Indian Institute of Technology Kharagpur (IITK) approached the UERC with a two-step technique that first detected whether the input images belong to the left or the right ear and then computed ear descriptors for matching from the side-normalized images using a pretrained 16-layer VGG model.

The IITK group trained a simple SVM model over HOG descriptors to classify whether the given image corresponds to the left or the right ear. For this step, the training images were resized to 30×60 pixels prior to the extraction of the HOG features.

2. https://ibug.doc.ic.ac.uk/resources/ibug-ears/

TABLE 2
Comparative overview of the participating approaches. The table provides a short description of each approach, information on whether ear alignment and flipping was performed and the model size (if any). The model size is approximate and given in MB.

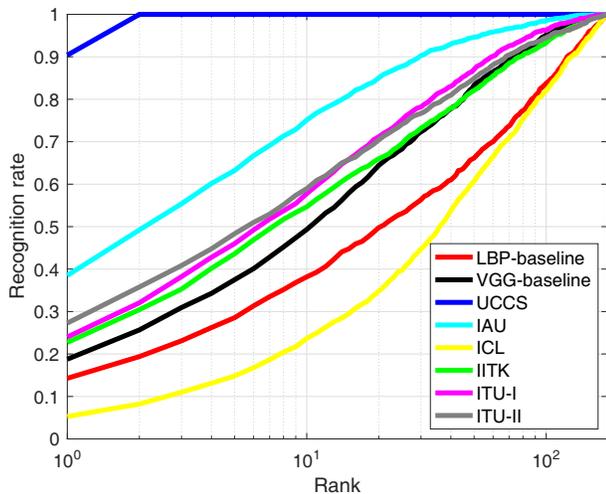| Approach	Description	Descriptor type	Ear alignment	Flipping	Model size
UCCS	Descriptor-based (chainlets)	Hand-crafted	No	No	no model used
IAU	VGG network (trained on ImageNet) and transfer learning	Learned	No	No	$\sim$ 100 MB
ICL	Deformable model and Inception-ResNet	Learned	Yes	Yes	$\sim$ 100 MB
IITK	VGG network (trained on the VGG face dataset)	Learned	No	Yes	$\sim$ 100 MB
ITU I	VGG network (trained on ImageNet) and transfer learning	Learned	No	Yes	$\sim$ 100 MB
ITU II	Ensemble method (VGG-network + LBP)	Learned + Hand-crafted	No	Yes	$\sim$ 100 MB
LBP-baseline	Descriptor-based (uniform LBPs)	Hand-crafted	No	No	no model used
VGG-baseline	VGG network trained solely on the UERC training data	Learned	No	No	$\sim$ 500 MB

Fig. 3. CMC curves of the comparative assessment on the images from the AWE dataset. The rank of the experiments is plotted on logarithmic scale to better highlight the performance at the lower ranks. The figure is best viewed in color.

TABLE 3
Comparative evaluation on probe images originating from the AWE dataset. The table shows Rank-1, Rank-5 and AUC values for all tested techniques.

Approach	Rank-1 (in %)	Rank-5 (in %)	AUC
UCCS	90.4	100	0.994
ICL	5.3	14.8	0.717
IAU	38.5	63.2	0.940
IITK	22.7	43.6	0.861
ITU-I	24.0	46.0	0.890
ITU-II	27.3	48.3	0.877
LBP-baseline	14.3	28.6	0.759
VGG-baseline	18.8	37.5	0.861

generated from the initial set of $2,304$ ear images. For data augmentation rotation, translation, flipping, intensity changes, cropping, scaling, and sharpening were used. Once the model was fine-tuned, the output of the first fully connected layer ($FC6$) of the VGG model was selected as the ear descriptor.

During run-time, each probe image and its flipped version were processed by the fine-tuned VGG model and the corresponding ear descriptors were matched against the given gallery descriptor with the chi-square distance. Then z-score normalization was performed. This procedure resulted in two scores that were summed up to produce the final similarity for the probe-to-gallery comparison.

ITU-II: The second approach of the ITU group exploited learned as well as hand-crafted ear descriptors and a fusion procedure applied at the matching score level. Specifically, the ITU-II approach relied on the VGG-based model described above (i.e., see ITU-I) and the LBP-baseline provided by the UERC organizers. Descriptors computed with the two techniques were matched separately against the corresponding gallery descriptors using the chi-square distance, followed by z-score normalization, and resulted in two match scores. The matching procedure was then repeated with flipped probe images and produced another pair of scores. The final probe-to-gallery similarity was ultimately generated by summing up the four scores produced during matching.

4.7 Summary of participating approaches

A high-level overview of all participating approaches is given in Table 2. The majority of participating groups approached the challenge with deep learning techniques despite the availability of a relatively small amount of training data. Moreover, among the five deep-learning approaches, four relied on the VGG model

The output of the classifier was then used to flip the images to a common reference. The side-normalized images were resized to a fixed size of 224×224 pixels and fed to the pretrained VGG face network [39]. Because the last few fully connected layers of the VGG model were tuned specifically for face recognition, only the feature maps from the convolutional layers were considered and the pooled output of these feature maps was used as the ear descriptor. For similarity score calculation, the cosine distance between two ear descriptors was adopted.

4.6 Istanbul Technical University

The participants from the Istanbul Technical University (ITU) submitted two approaches to the UERC. The first was a deep learning approach based on the VGG network [46] architecture (ITU-I hereafter) and the second an ensemble approach combining the VGG model with hand-crafted LBP descriptors (ITU-II from hereon) [42]. A brief summary of both approaches is given below.

ITU-I: The ITU-I approach used a 16-layer VGG model pretrained on the ImageNet dataset and fine-tuned on the UERC training images. Since CNN-based models, such as VGG, need significantly more training data than is available with the UERC dataset to ensure competitive recognition performance, data augmentation was performed and a total of $250,000$ ears were
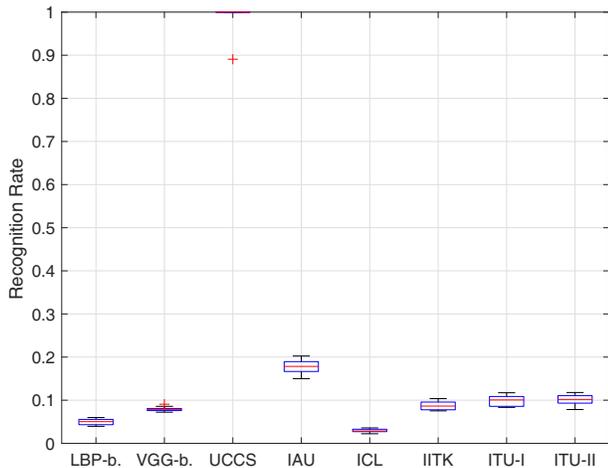
Fig. 4. Recognition performance with a single gallery image per subject. The box plots show the distribution of the rank-1 recognition rates computed over 10 experimental runs.

architecture but used different preprocessing (e.g., alignment, flipping) and descriptor extraction strategies (e.g., output layer selection). A single approach (ICL) used another architecture, i.e., Inception-ResNet. Hand-crafted descriptors were less popular. Only the UCCS group submitted an approach based on chainlets, while a second one was provided by the challenge organizers in the form of the LBP-baseline.

5 EXPERIMENTS AND RESULTS

In this section we present a comprehensive analysis of all participating approaches and the results of the challenge. For potential updates and additions to the results, the reader is referred to the UERC website[3].

Overall performance comparison: We first investigate how the submitted algorithms perform on ear images captured *in the wild* and how they compare among each other. For this part of the evaluation, we use only $1,800$ probe images originating from the AWE dataset and compute our results on a similarity matrix of size $1,800 \times 1,800$ in an all-vs-all experimental setup. Each of the 180 subjects involved in this experiments is represented in the gallery with 10 ear images and retrieving any of these 10 images based on the given probe is counted as correct recognition attempt.

The results of this experiment are presented in the form of CMC curves in Fig. 3 and with different performance metrics in Table 3. Overall the UCCS approach results in the best performance with a recognition rate of 90.4% at a tank of one (100% at a rank of 2), followed in order by the IAU, ITU-1, ITU-II, IITK and ICL approaches. A similar ranking can be established if the AUC values are considered instead of the rank-1 recognition rates. The LBP and VGG baselines achieve a rank-1 recognition rate of 14.3% and 18.8%, respectively, with the deep learning baseline having a slight advantage over the hand-crafted LBP descriptor.

Number of galleries per subject: We next study the effect of reducing the number of gallery images for each subject from 10 to 1. This represents a significantly harder problem than in the first experiment, as only a single comparison is available per subject to make an identification decision. Because the 10 gallery images

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

that are available in the AWE dataset for each subject are divided among left and right ears, comparisons in this experiment may include comparisons of ears from the opposite sides of the head. We perform the experiment 10-times, so that each of the gallery images that available in the dataset per subject is used once. The probe set consists of all $1,800$ AWE images.

As we can see from the box plots in Fig. 4, the performance for most of the approaches is halved (on average), except for the UCCS approach, which achieves a rank-1 recognition rate of 1 (100%) in 9 out of the 10 experimental runs. These results suggest that having multiple images (of left and right ears) per subject is detrimental for the recognition performance of most techniques. Even if techniques detect whether the probe and gallery images are from the same side of the head, it is not necessary possible to match the right ear to the left and vice versa. As pointed out by previous studies, e.g., [1], [54] ears are not always bilateral symmetric, though this is true for most subjects.

Head rotation: Images from the AWE dataset contain annotations with respect to pitch, roll and yaw angles that can be used to explore the impact of head rotation on ear recognition performance. The available annotations (see [15] for details) are grouped into three categories, i.e., *None*, *Moderate* and *Severe* according to the extent of the head-rotation. Recognition experiments are then performed with probe images of a single category at the time, while the gallery is kept unchanged (i.e., all $1,800$ gallery images are used). As we can see from Fig. 5, where the AUC values obtained during the experiments are shown, both pitch and roll angles have an adverse affect of most of the techniques, while differences in yaw angles affect the performance to a lesser extent. This results indicates that resampling the ear images to a common size already compensates for the difference in yaw angles, while roll and pitch angles would need to be compensated for explicitly. The UCCS approach is the most robust and is not affected by head rotation.

Same-side vs. opposite-side matching: As already indicated above, the images of all subject in the AWE dataset are split between right and left ears (i.e., images are not side-normalized). When matching probes to galleries, some of the participating approaches try to detect explicitly whether they are processing left or right ears and then flip the images to a common reference. To evaluate how this process effects performance, we conduct two types of experiments: *i)* experiments with ear images from the same side (e.g., left-to-left), and *ii)* experiments with ear images from opposite sides of the head (e.g., right-to-left).

The results in Fig. 6 show that the SDM fitting approach from the ICL group is the most successful at determining which side of the head the ear images came from, as the performance change in the two experiments is marginal for the ICL approach. The strategies from ITU and IITK are less successful and still result in observable performance drops, but the difference in performance is less than with the baseline techniques, where no effort is made to distinguish left from right ears. The UCCS approach also shows a high level of robustness as a consequence of the exploited chainlet-based descriptor.

Learned vs. hand-crafted descriptors: In our next experiment we compare deep-learning approaches that try to learn ear descriptors from data to techniques that exploit so-called hand-crafted descriptors. Since only UCCS submitted a technique based on hand-crafted descriptors, we also report results for the remaining UERC baselines that are based on HOG [8], [40], LBP [42], BSIF [23], DSIFT [30], LPQ [37], RILPQ [38] and POEM [50]
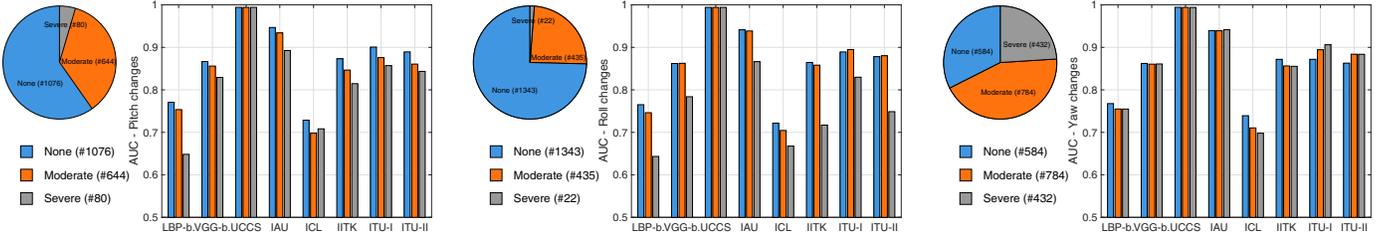
Fig. 5. Effect of head rotation on ear recognition performance. The left graph shows the change in the AUC with respect to increasing pitch angles, the middle graph shows the AUC change with respect to increasing roll angles and the right graph shows the change in the AUC with respect to increasing yaw angles. The pie charts show the distribution of the images across the different rotation-angle labels. Pitch and roll angles have an adverse effect on the recognition performance of most techniques, whereas yaw angles affect the performance to a lesser extent.
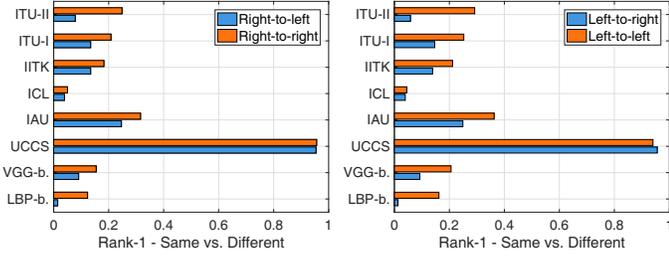

Fig. 6. Effect of same-side vs. opposite-side matching. The right graph shows the rank-1 recognition rates when matching right ear probes to either right or left ear galleries. The left graph shows results for the same experiment, but for left-ear probes.
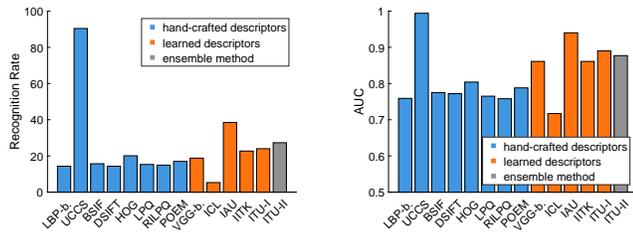

Fig. 7. Comparison of learned and hand-crafted descriptors. The left graph shows the rank-1 recognition rate and the right graph shows the AUC of the experiments with 1800 probe and gallery images of the AWE dataset.
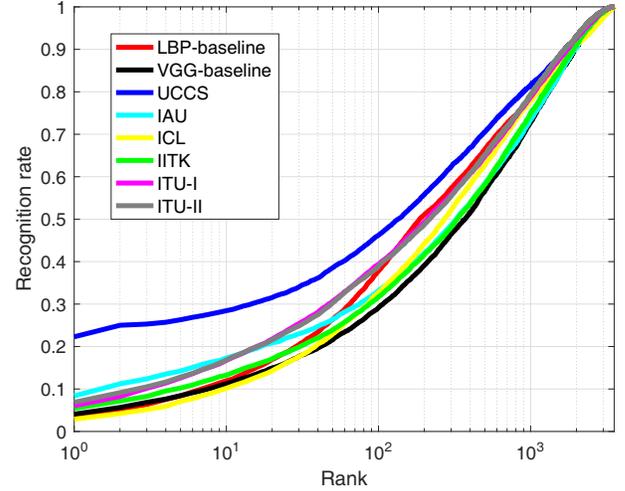

Fig. 8. CMC curves of the scalability experiments on the entire UERC dataset. The rank of the experiments is plotted on a logarithmic scale to better highlight the performance at the lower ranks. The figure is best viewed in color.

descriptors. For a detailed description of these techniques please see [13].

From the results in Fig. 7 we see that most of the techniques build around hand-crafted descriptors result in similar performance with a rank-1 recognition rate between 14.3% and 20.1%, except for the UCCS approach which achieves a recognition rate of 90.4% at rank one. The deep learning approaches, on the other hand, vary significantly in performance despite the fact that 5 of them use the deep model. We observe rank-1 rates between 5.3% and 38.5%. This suggests that the training strategy is of paramount importance when the available training data is limited and has a larger impact on the recognition performance than the model architecture. The only ensemble method (ITU-II) benefits from two sources of information and improves upon the performance of the individual techniques (VGG and LBP), from which it was built.

Scalability: In our last experiment we evaluate how the recognition techniques scale with larger probe and gallery sets. We show CMC curves generated based on 7,442 probe images belonging to 1,482 subjects and 9,500 gallery images of 3,540 subjects in Fig. 8. Note that the gallery also contains identities

TABLE 4
Key performance metrics of the scalability experiments. The table shows Rank-1, Rank-5 and AUC values for all tested techniques.

Approach	Rank-1 (in %)	Rank-5 (in %)	AUC
UCCS	22.3	26.3	0.858
ICL	2.9	7.08	0.823
IAU	8.4	14.2	0.810
IITK	5.4	10.2	0.814
ITU-I	6.1	12.6	0.842
ITU-II	6.9	12.8	0.844
LBP-baseline	3.75	8.38	0.835
VGG-baseline	4.07	8.43	0.804

that are not in the probe set and act as distractors (to use the terminology from [24]) for the recognition techniques. Quantitative performance metrics of the experiments are presented in Table 4. All techniques deteriorate significantly in performance with scale. The best performer is again the UCCS approach with a recognition rate of 22.3% at a rank of 1, followed in order by the IAU, ITU-II, ITU-I, IITK and ICL approaches. When the AUC values are considered the ITU approaches become competitive and get close to the performance of the UCCS technique due to a better performance at the higher ranks.

6 CONCLUSION

We have presented the results of the first Unconstrained Ear Recognition Challenge (UERC) that aimed at evaluating the current state of technology in the field of ear recognition from images captured in unconstrained environments. While additional experiments are presented in the Appendix, open questions still remain for future challenges.

Several important findings were made in this paper, e.g.: *i)* significant performance improvements are needed before the ear-recognition technology is suitable for deployment in unconstrained environments at scale, *ii)* existing approaches are mostly sensitive to specific head rotations (in pitch and roll directions, but not yaw), and *iii)* it is detrimental to have multiple images of both ears in the gallery, as identity inference based on a single-image-per-subject basis leads to poor recognition performance.

REFERENCES

[1] A. Abaza and A. Ross. Towards understanding the symmetry of human ears: A biometric perspective. In *International Conference on Biometrics: Theory applications and systems*, pages 1–7. IEEE, 2010.

[2] A. Ahmad, E. Omar, and T. E. Boult. Chainlets: Towards ear recognition. In *International Joint Conference on Biometrics, under review*. IEEE/IAPR, 2017.

[3] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Lee, V. E. Liong, J. Lu, M. de Assis Angeloni, T. de Freitas Pereira, H. Li, G. Hua, et al. The IJCB 2014 PaSC video face and person recognition competition. In *International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[4] K. W. Bowyer. The results of the NICE.II Iris biometrics competition. *Pattern Recognition Letters*, 33(8):965–969, 2012.

[5] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, et al. Competition on counter measures to 2-D facial spoofing attacks. In *International Joint Conference on Biometrics*, pages 1–6. IEEE, 2011.

[6] T.-S. Chan and A. Kumar. Reliable ear identification using 2-D quadrature filters. *Pattern Recognition Letters*, 33(14):1870–1881, 2012.

[7] I. Chingovska, J. Yang, Z. Lei, D. Yi, S. Z. Li, O. Kahm, C. Glaser, N. Damer, A. Kuijper, A. Nouak, et al. The 2nd competition on counter measures to 2D face spoofing attacks. In *International Conference on Biometrics*, pages 1–6. IEEE, 2013.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[9] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein. Ssrbc 2016: sclera segmentation and recognition benchmarking competition. In *International Conference on Biometrics*, pages 1–6. IEEE, 2016.

[10] A. Das, U. Palb, M. A. Ferrerc, and M. Blumensteina. Ssbc 2015: Sclera segmentation benchmarking competition. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–6. IEEE, 2015.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[12] Z. Emersic, L. L. Gabriel, V. Struc, and P. Peer. Pixel-wise ear detection with convolutional encoder-decoder networks. *arXiv preprint arXiv:1702.00307*, 2017.

[13] Z. Emersic, B. Meden, V. Struc, and P. Peer. Covariate analysis of descriptor-based ear recognition techniques. In *International Work Conference on Bioinspired Intelligence*, pages 1–9, 2017.

[14] Z. Emersic, D. Stepec, V. Struc, and P. Peer. Training convolutional neural networks with limited training data for ear recognition in the wild. In *International Conference on Automatic Face and Gesture Recognition – Workshop on Biometrics in the Wild*. IEEE, 2017.

[15] Z. Emersic, V. Struc, and P. Peer. Ear recognition: More than a survey. *Neurocomputing*, in press:1–22, 2017.

[16] S. Escalera, X. Baró, H. J. Escalante, and I. Guyon. Chalearn looking at people: A review of events and resources. *arXiv preprint arXiv:1701.02664*, 2017.

[17] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Conference on Computer Vision and Pattern Recognition – Workshops*, pages 1–8, 2016.

[18] A. Fabate, M. Nappi, D. Riccio, and S. Ricciardi. Ear recognition by means of a rotation invariant descriptor. In *International Conference on Pattern Recognition*, volume 4, pages 437–440. IEEE, 2006.

[19] L. Ghiani, D. Yambay, V. Mura, S. Tocco, G. L. Marcialis, F. Roli, and S. Schuckcrs. Livdet 2013 fingerprint liveness detection competition 2013. In *International Conference on Biometrics*, pages 1–6. IEEE, 2013.

[20] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds. The NIST 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 224–230, 2014.

[21] Y. Guo and Z. Xu. Ear recognition using a new local matching approach. In *International Conference on Image Processing*, pages 289–292. IEEE, 2008.

[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[23] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *International Conference on Pattern Recognition*, pages 1363–1366. IEEE, 2012.

[24] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[25] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, T. Vojir, G. Häger, A. Lukežič, and G. Fernandez. The visual object tracking VOT2016 challenge results. Springer, Oct 2016.

[26] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder. The visual object tracking VOT2015 challenge results. In *International Conference on Computer Vision Workshops*, pages 1–23, 2015.

[27] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[29] A. Kumar and T.-S. T. Chan. Robust ear identification using sparse representation of local texture descriptors. *Pattern recognition*, 46(1):73–85, 2013.

[30] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[31] D. Maio, D. Maltoni, R. Cappelli, J. Wayman, and A. Jain. Fvc2004: Third fingerprint verification competition. *Biometric Authentication*, pages 31–35, 2004.

[32] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2000: Fingerprint verification competition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):402–412, 2002.

[33] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *International Conference on Pattern Recognition*, volume 3, pages 811–814. IEEE, 2002.

[34] S. Marcel, C. McCool, P. Matějka, T. Ahonen, J. Černockỳ, S. Chakraborty, V. Balasubramanian, S. Panchanathan, C. H. Chan, J. Kittler, et al. On the results of the first mobile biometry (mobio) face and speaker verification evaluation. In *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 210–225. Springer, 2010.

[35] G. Marcialis, A. Lewicke, B. Tan, P. Coli, D. Grimberg, A. Congiu, A. Tidu, F. Roli, and S. Schuckers. First international fingerprint liveness detection competitionlivdet 2009. *Image Analysis and Processing*, pages 12–23, 2009.

[36] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, and S. Marcel. Keystroke biometrics ongoing competition. *IEEE Access*, 4:7736–7746, 2016.

[37] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, pages 236–243. Springer, 2008.

[38] V. Ojansivu, E. Rahtu, and J. Heikkilä. Rotation invariant local phase quantization for blur insensitive texture analysis. In *International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[40] A. Pflug, P. N. Paul, and C. Busch. A comparative study on texture and surface descriptors for ear biometrics. In *International Carnahan Conference on Security Technology*, pages 1–6. IEEE, 2014.

[41] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954. IEEE, 2005.

[42] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Local binary patterns for still images*. Springer, 2011.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[44] W. J. Scheirer, P. J. Flynn, C. Ding, G. Guo, V. Struc, M. Al Jazaery, K. Grm, S. Dobrisek, D. Tao, Y. Zhu, et al. Report on the BTAS 2016 video person recognition evaluation. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–8. IEEE, 2016.

[45] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Conference on Computer Vision and Pattern Recognition*, page 39823991, 2015.

[46] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[48] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[49] P. Tome, R. Raghavendra, C. Busch, S. Tirunagari, N. Poh, B. Shekar, D. Gragnaniello, C. Sansone, L. Verdoliva, and S. Marcel. The 1st competition on counter measures to finger vein spoofing attacks. In *International Conference on Biometrics*, pages 513–518. IEEE, 2015.

[50] N.-S. Vu and A. Caplier. Face recognition with patterns of oriented edge magnitudes. *Computer Vision*, pages 313–326, 2010.

[51] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Conference on Computer Vision and Pattern Recognition – Workshops*, pages 56–64, 2016.

[52] R. Xian, L. Ni, and W. Li. The ICB-2015 Competition on Finger Vein Recognition. In *International Conference on Biometrics*, pages 85–89. IEEE, 2015.

[53] D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers. Livdet-iris 2013-iris liveness detection competition 2013. In *International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[54] P. Yan and K. Bowyer. Empirical evaluation of advanced ear biometrics. In *Conference on Computer Vision and Pattern Recognition – Workshops*, pages 41–41. IEEE, 2005.

[55] Y. Ye, L. Ni, H. Zheng, S. Liu, Y. Zhu, D. Zhang, W. Xiang, and W. Li. Fvrc2016: The 2nd finger vein recognition competition. In *International Conference on Biometrics*, pages 1–6. IEEE, 2016.

[56] Y. Zhou, E. Antonakos, J. Alabort-i Medina, A. Roussos, and S. Zafeiriou. Estimating correspondences of deformable objects. In *Conference on Computer Vision and Pattern Recognition*, pages 5791–5801, 2016.

[57] K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.

APPENDIX

In this appendix we present additional results from the UERC. The results relate to the impact of occlusions, gender and image size on performance. Qualitative results are also presented.

Occlusion: We explore the impact of partial ear occlusion on images originating from the AWE dataset. We run three experiments and during each use only probe images labeled with one of the following occlusion labels: *Minor*, *Moderate* and *Severe*. For the gallery we use all $1,800$ AWE images from the UERC test partition. The results for this experiment are presented for the rank-1 recognition rate in Fig. 9 (left) and for the AUC in Fig. 9 (right). As can be seen, minor and moderate occlusions caused, for example, by hair have limited effect on the recognition performance, whereas severe occlusion from scarfs, large accessories and a like have a larger impact on all submitted approaches.

Gender: To study the effect of gender on the recognition performance, we run separate test with probe images from the AWE dataset belonging to male ($1,420$ images) and female subjects (320 images). The gallery again comprises all $1,800$ available images. We can see from the rank-1 recognition rates in Fig. 10 that the performance of all approaches participating in the UERC is more or less unaffected by gender. Both men and women are recognized equally well.

Image size: Next, we explore the impact of image size on the recognition performance. For this experiment we use the entire similarity matrix and select probe images from four categories: *i)* images with a pixel count below $1,000$ ($4,600$ images), *ii)* images of size between $1,000$ and $5,000$ pixels ($3,856$ images), *iii)* images of size between $5,000$ and $10,000$ pixels (413 images), and *iv)* images with more than $10,000$ pixels (631 images). We use all $9,500$ gallery images for the assessment and present the results in the form of rank-1 recognition rates and AUC values in Fig. 11.

As we can see, all approaches are affected by the changes in image size. The biggest performance drop is seen for the UCCS approach, which performs very well on the large images from the AWE dataset, but is less successful on the smaller images of the newly collected part of the UERC data. Also interesting are the results for the ICL approach, which show that the rank-1 recognition rate is close for all image-size categories. The remaining approaches all behave similarly with the performance decreasing with decreasing image size. This observation is important and shows that images of sufficient size need to be available for ear recognition. Alternatively, models robust to image size or supper-resolution techniques may need to be incorporated into the recognition pipelines to make ear recognition work with small images.
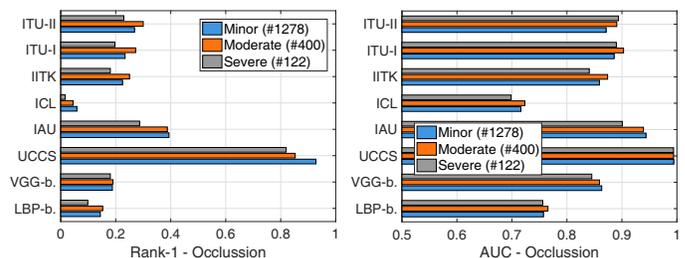

Fig. 9. Impact of occlusions on the recognition performance. The left graph shows the rank-1 recognition rates and the right graph shows AUC values for all tested techniques. The number in the brackets indicate the number of images for each label.

TABLE 5
Qualitative examples of the recognition performance. The table shows a few selected probe images and the best and second best match from the gallery. The correct gallery and information about the rank at which the correct gallery was retrieved is also given. The table shows some of the errors and corrected identification attempts made by the tested approaches.

Approach	Probe	1. match	2. match	Correct	Retrieved at	Probe	1. match	2. match	Correct	Retrieved at
LBP-baseline					rank 39					rank 85
					rank 15					rank 1
VGG-baseline					rank 49					rank 45
					rank 34					rank 1
UCCS					rank 1					rank 1
					rank 1					rank 1
IAU					rank 11					rank 20
					rank 78					rank 1
ICL					rank 6					rank 142
					rank 151					rank 5
IITK					rank 22					rank 151
					rank 78					rank 1
ITU-I					rank 3					rank 25
					rank 45					rank 1
ITU-II					rank 6					rank 41
					rank 21					rank 1

Qualitative evaluation: Last but not least we shows some qualitative examples of the recognition performance of all ap- proaches participating in the UERC in Table 5. For this experiment we again use only images from the AWE dataset and select 1
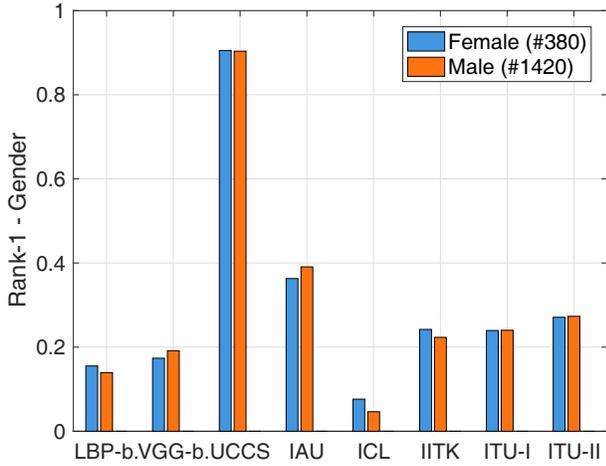
Fig. 10. Impact of gender on the recognition performance. The graph shows the rank-1 recognition rates for all tested techniques. The number in the brackets indicate the number of images for each label.
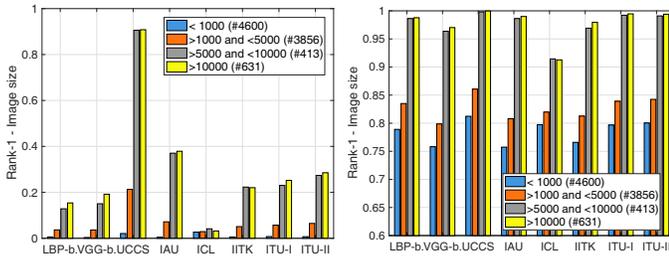

Fig. 11. Impact of image size on the recognition performance. The left graph shows the rank-1 recognition rates and the right graph shows AUC values for all tested techniques. The number in the brackets indicate the number of images for each label.

image per subject for the gallery, so the maximum rank of this experiment is 180. In Table 5 we show selected probe images and for each probe image, the best and second best match that was retrieved from the gallery. The correct galleries and information about the rank at which the correct gallery was retrieved is also given in the table. We can see that the UCCS approach selects the correct gallery for all of the selected examples despite different image characteristics. Other approaches make different types of errors, the most common being selecting a similarly looking image from another subject, but taken at same side of the head.