Analysis of the Impact of Ear Alignment on Unconstrained Ear Recognition

Elaine Grenot-Castellano(✉), Yoanna Martínez-Díaz, and Francisco José Silva-Mata

Advanced Technologies Application Center, 7th A Avenue #21406 % 214 and 216, Siboney, Playa, 12200 Havana, Cuba
{egrenot,ymartinez,fjsilva}@cenatav.co.cu

Abstract. The use of the ear in biometric recognition has been widely covered in controlled environments. However, the advantages of the ear as a biometric characteristic impose the need to know how it behaves in unconstrained scenarios, where it is common the presence of occlusions, pose variations, illumination changes and different resolutions. According to this challenge and considering the experience in other biometric recognition processes, the alignment has shown to be a key step. In this work, we carry out an exhaustive and detailed study of the impact of the alignment on the performance of several state-of-the-art ear descriptors, when the images are captured in uncontrolled conditions. Our analysis is based on identification experiments against different types of variations in ears image of the challenging UERC dataset. The obtained results corroborate the hypothesis of the alignment also improves the efficacy of the ear recognition process and show how this improvement behaves for various factors such as head rotation, occlusions, flipping and resolution.

Keywords: Ear alignment · Unconstrained ear recognition · Covariates

1 Introduction

The face, the iris, and the fingerprint are examples of the most popular biometric objects used for person recognition. In recent times, the ear has become important as an identifying part among people. The rich structure of an ear combined with its stability over time is a promising source of data to identify subject since its collection can be done in a noninvasive way, has a high degree of permanence, distinctiveness and universality [5]. However, some factors such as partial o full occlusions, pose variations and the presence of ear accessories can be affect sensitively the ear recognition performance.

A typical fully automatic ear recognition system follows a traditional pipeline of detection, alignment, feature extraction and classification. Many approaches have been proposed attempting to improve ear recognition capabilities for reliable deployment in surveillance and commercial applications [1,11]. Most of these

I. Nyström et al. (Eds.): CIARP 2019, LNCS 11896, pp. 283–293, 2019.
https://doi.org/10.1007/978-3-030-33904-3_26

works rely on develop feature descriptors that can be resilient to variability found in unconstrained conditions. Depending on the type of feature extraction technique used, ear recognition approaches can be grouped into hand-crafted [5] and deep-learning descriptors [2,8].

As in other modalities such as the face and iris, image alignment plays a crucial role in a recognition system, since most approaches are very sensitive to the pose and scale variations. Even the best performing state-of-the-art descriptors require that images are aligned as good as possible in order to achieve better results. In the case of ear, several methods [12,14,15] have been develop for aligning images but it is still not completely clear how these methods are able to improve the recognition performance in the presence of factors found in unconstrained settings such as head rotation, occlusions or image resolution.

The main contribution of this work is a comprehensive experimental evaluation of several state-of-the-art ear recognition techniques on the challenging UERC dataset with the aim of studying the effect of alignment on uncontrolled conditions. Specifically, we perform a comparative assessment of recent hand-crafted and deep-learning descriptors using both aligned and no aligned images and investigate their robustness in front to unseen data characteristics such rotations, occlusions and image resolution. As result, we present an extensive experimental analysis in terms of recognition rates which contributes to a better understanding of the behavior of the alignment on the evaluated methods, showing its importance on unconstrained ear recognition.

The remainder of this paper is organized as follows. Section 2 describes the existing works related to the ear alignment topic. In Sect. 3 we present the ear alignment method and recognition techniques considered in this work. The experimental setup and the results obtained are provided in Sect. 4. Finally, conclusion and future work are given in Sect. 5.

2 Alignment Methods

Different from other biometric features such as the iris (radial symmetry and approximately circular shape) or the face (it is possible to determine an axis of approximate symmetry that divides the face into two similar parts), the ear lacks symmetrical properties. Therefore, the attempts to align the ear images have depended to a large extent on defining certain axes or parts of its that serve to be taken as reference for the alignment [12,14,15].

Some authors have been used the helix (outer edge of the ear) as reference to align ear images [14,15]. The main difficulty of this approach is that it must use precise methods of edge detection, in order to determine the reference axes or landmark. The elliptical shape of the ear has also exploited by using a cascaded pose regression [12]. This method fits the ear outer rim with an abstract elliptical model and then, transforms it to its normal position given by the main axes of the ellipse. In [13] the Random sample consensus (RANSAC) [7] is used over SIFT descriptors to estimate the transformation in the plane of each image to an average image, which is then applied together with ear mask. Various statistical

deformable models with different features descriptors were evaluated in [17] for ear landmark localization on images taken from uncontrolled environments. As result, their best combination was achieved by using a holistic Active Appearance Model based on SIFT features.

Recently, deep convolutional neuronal networks (CNNs) have also been used to detect landmarks in different areas of the ear in order to carry out their alignment [8,16]. A cascading convolution neural network was proposed in [16] to detect six landmark points. These points were defined in accordance with the morphological and geometric characteristics of the ear; three of them were located in its internal region and three in the external contour of the ear. In [8] the authors introduced a two-stage landmark detector based on Convolutional Neural Networks to locate a set of 55 landmarks which are then employed to translate, rotate and scale the input ear image.

Although several methods have been proposed for ear alignment, few works have investigated its role in unconstrained ear recognition. In [13] the authors evaluate the influence of RANSAC method but only in a subset of images of AWE dataset, according to the severity of pose variations. Hansley et al. [8] analyze the benefits of their alignment method by checking the difference in the recognition performance with and without alignment; but only for hand-crafted descriptors. In [17] aligned versus non-aligned ears were compared in ear verification and close identification experiments. All these works demonstrate that in general, the alignment consistently improves the ear recognition performance. However, they do not provided detailed information about its importance in front different covariates present in uncontrolled conditions.

In the present work we evaluate the impact of alignment on ear recognition under novel aspects that were not considered before such as occlusions and image resolution. In addition, we perform an extensive experimental analysis for both hand-crafted and deep-learning descriptors on the challenging UERC database.

3 Baselines

In order to align the ear images we select the method proposed in [8], since unlike others methods it directly attacks one of the more difficult problem: the pose variations. Their solution relies on a two-stage landmark detector based on CNNs to locate a set of 55 landmarks. The first network is used to create an easier landmark detector by reducing scale and translation variations. The coordinates obtained by this network are used to refine the center and orientation of an ear image and then, the rectified image is used as input of the second network to fine-tune small variations. After landmark detection, the ears are normalized by applying PCA on the retrieved landmarks. Finally, in order to diminish the effects of poses variations, different sampling rates are used in such a way the width and the height of the normalized ear are approximately the same. In addition, before the automatic aligning process, the authors use a simple side classifier to detect explicitly whether they are processing left or right ears and then flip the images to a common reference, so that all ears would have the same orientation.

3.1 Ear Recognition Techniques

With the aim of evaluating the benefits of the previous alignment method, we considered seven hand-crafted and two deep-learning descriptors based on their good performance reported for ear recognition [3,5].

Specifically, we used Local Binary Patterns (LBP), Gabor wavelets, Binarized Statistical Image Features (BSIF), Local Phase Quantization Features (LPQ), Rotation Invariant LPQ (RILPQ), Patterns of Oriented Edge Magnitudes (POEM) and Histograms of Oriented Gradients (HOG) as hand-crafted descriptors [5]. In the case of deep-learning descriptors, we selected the MobileNet [10] and the ResNet-18 [9] networks, which cover some of the most popular architectures for ear recognition.

4 Experimental Setup

In this section, we asses the performance of hand-crafted and deep-learning techniques with and without alignment. First, we describe the recognition dataset used and give some implementations details. Then, we present the recognition results obtained, taking into account different covariates.

4.1 Ear Recognition Dataset and Protocols

For the experimental evaluation, we use the UERC 2019 dataset [6], that consists of 11 000 ear images collected from the web of 3 690 subjects, making it the largest publicly available dataset of unconstrained ear images.

The main part of this dataset was taken from the Extended Annotated Web Ears (AWEx) dataset [5] and comprised 3 300 ear images of 330 subjects. Images from this part are annotated with different covariates hence, it was used as the basis for our analysis. The rest of the data was taken from the UERC 2017 dataset [4], which presents characteristics and variability similar to the AWEx images, but with greater variations in the size of the images. Sample images of the UERC dataset are illustrated in Fig. 1.

Fig. 1. Examples of images from the UERC 2019 dataset.

In order to develop and test models, the public UERC 2019 dataset was partitioned into disjoint training and testing sets. The training set consists of 2 304 images of 166 subjects from the AWEx dataset, whereas the testing set contains the remaining AWEx data and the rest of the images from UERC 2017.

4.2 Implementations Details

In the case of the alignment method [8], we use the demo and the deep models provided by the authors (http://github.com/maups/ear-recognition), where the two-stage landmark detector and the side classifier are available.

For hand-crafted descriptors it was used the implementations provided in the AWE toolbox [5] with their default values. The MobileNet and ResNet-18 models was used with initial parameters learned on the ImageNet dataset and fine-tune certain layers training using aligned and non-aligned ear images, separately. For this, the UERC training set was used and data augmentation was performed with a 50% chance. For both CNNs we set the learning rate to 0.01, the momentum to 0.75 and the weight decay to 0.005. After training, the last fully-connected layers from the networks were used as feature extractors. Once the representations are computed, the cosine similarity is used for the comparison of test images.

4.3 Experimental Results

Several identification experiments were carried out, taking into account different factors that affect the performance of ear recognition process in unconstrained scenarios, such as sensitivity to same-side vs. opposite-side matching, occlusions, rotations (in terms of yaw, pitch and roll angles) and different resolutions.

Table 1. Recognition rates (%) at rank-1 and rank-5 for all evaluated descriptors using aligned and non-aligned images of AWEx and UERC datasets.

	Rank-1 AWEx		Rank-5 AWEx		Rank-1 UERC		Rank-5 UERC	
	No align	Align	No align	Align	No align	Align	No align	Align
LBP	13.17	28.39	26.17	46.39	8.07	15.39	15.76	**25.50**
GABOR	14.28	22.61	28.44	40.94	5.47	9.26	11.33	17.60
BSIF	15.33	28.72	31.06	47.22	8.55	14.93	16.37	24.33
LPQ	14.94	27.28	28.33	45.00	8.75	14.57	17.01	24.59
RILPQ	15.89	27.61	29.11	44.33	7.26	13.33	13.56	21.61
POEM	18.28	31.39	32.72	52.78	8.36	14.70	15.74	25.30
HOG	19.39	**40.50**	36.00	**59.67**	7.93	**17.27**	15.04	**26.31**
ResNet-18	15.72	20.17	36.56	42.33	6.42	7.91	15.32	17.28
MobileNet	**21.44**	25.89	**44.84**	48.06	**8.99**	10.43	**19.58**	20.53

Ear Identification. Table 1 shows the recognition rates at rank-1 and rank-5 for each descriptor using aligned and non-aligned images from the testing sets of AWEx dataset (involving 180 subjects) and UERC 2019 dataset. As it can be seen, in general, all the results are improved when the images are aligned. However, these improvements were noticeably lower for the case of deep

descriptors. We think that this is because these deep models are able to learn in a better way the variations present in the training images. In addition, if these variations are severe it can affect the performance of the alignment method which can introduce some noise information in the learning stage. Consequently, we can said that this kind of descriptors include a partially solution for non-aligned ears. In contrast, the alignment process shows significant improvements for all hand-crafted descriptors, even obtaining better results that deep descriptors.

The best rates at rank-1 for AWEx and UERC datasets using non-aligned images are achieved by the deep descriptor MobileNet with a 21.44% and 8.99%, respectively; while when the images are aligned the best rates are obtained by the hand-crafted HOG descriptor, increasing to 40.50% and 17.27% for AWEx and UERC, respectively. LBP and HOG descriptors are the most benefited when images are aligned, improving their results at least a factor of 1.5x in all cases.

The low recognition rates obtained for the UERC dataset corroborate that all the evaluated descriptors do not present good scalability when the number of subject increases and the ear images have poorer resolution and lower-quality, even using aligned images. The CMC curves shown in Fig. 2 complement the results, being able to observe clearly the increase of the area under the curve when the alignment is used for all the descriptors in AWEx and UERC datasets.

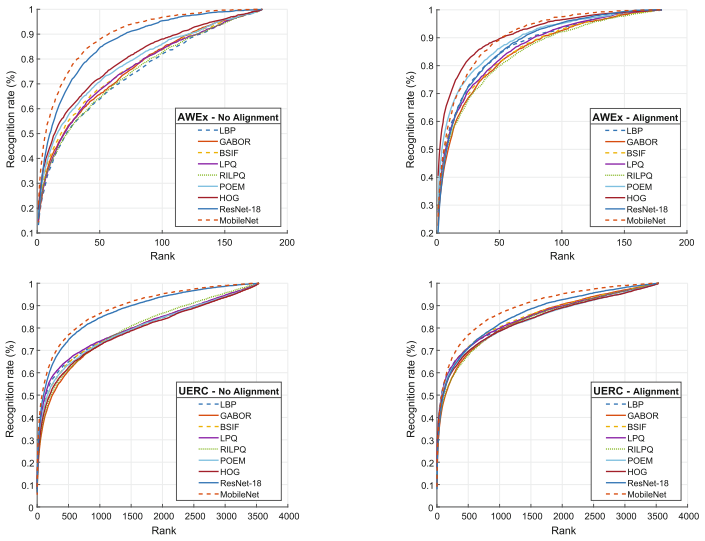

Fig. 2. CMC curves from the AWEx and UERC datasets.

Same-Side vs. Opposite-Side Matching. This experiment leads to know the impact of alignment when we match ear images from the same side (e.g., right-to-right), and from opposite sides of the head. Table 2 presents the recognition rates

at rank-1 when the original images, the side classifier (Flip) and the alignment method (Align) are used. The results evidence that determining which side of the head the ear images came from (Flip), improves the recognition especially when the ear images are from opposite sides, but the difference in performance is less than when alignment is applied. When the images are from the same side, the best results are obtained by the HOG descriptor, achieving the highest recognition rates by using aligned images. In case of ear images from opposite sides, deep-learning descriptors outperform hand-crafted descriptors when the images are not aligned, especially, when original images are used. This is because the deep models are trained to learn features that are mostly not affected by the corresponding ear side, although if the ear is misaligned by some error of the alignment method, this stage may work against it. However, using aligned images again HOG descriptor reaches the highest scores.

On the other hand, we can see that sometimes, when we match ears images from the same side, the results of using flipped ears are a little worse than those using the original version. This is due to errors of the classifier used to determine automatically the side of a given ear.

Table 2. Recognition rates (%) at rank-1 for same-side vs. opposite-side matching using original, flipped and aligned images from the AWEx dataset.

	Right-Right			Right-Left			Left-Left			Left-Right		
	Orig.	Flip	Align	Orig.	Flip	Align	Orig.	Flip	Align	Orig.	Flip	Align
LBP	11.42	11.19	21.28	0.89	8.17	17.02	14.88	14.33	22.16	1.10	11.36	19.40
GABOR	13.21	12.77	17.81	1.23	7.84	13.33	15.66	15.10	19.96	1.21	9.37	13.23
BSIF	13.10	12.99	22.84	1.23	11.42	18.93	17.86	17.53	21.61	0.99	12.89	17.86
LPQ	14.56	14.22	22.28	0.78	10.86	16.69	15.66	15.22	21.17	0.88	12.35	18.08
RILPQ	15.23	14.78	21.61	0.89	9.97	14.11	16.65	15.33	22.82	1.43	11.69	16.20
POEM	17.36	16.79	25.87	0.56	11.98	19.26	19.74	18.52	26.79	0.55	12.79	20.62
HOG	**18.25**	**17.58**	**30.35**	0.78	11.86	**23.29**	**20.84**	**21.06**	**35.94**	1.43	12.89	**23.59**
ResNet-18	12.54	12.65	16.46	9.52	11.31	13.10	15.44	17.64	16.54	9.04	10.58	13.89
MobileNet	16.46	17.36	20.60	**12.43**	**13.77**	17.58	20.07	18.85	21.50	**14.11**	**14.77**	17.75

Occlusion. Figure 3 shows the recognition rates at rank-1 for different occlusion levels. We can see that while the performance of evaluated descriptors for minor and mild occlusions are considerably improved by using the alignment method, this improvement is not the same when severe occlusions are present. The hand-crafted descriptors from aligned images are seem to ensure better performance across all levels. However, for deep-learning techniques using aligned images causes performance degradation in front severe occlusions. One possible hypothesis of this is that, in case of severe occlusions, as the problem consists of a great absence of information in the images to match, the alignment process

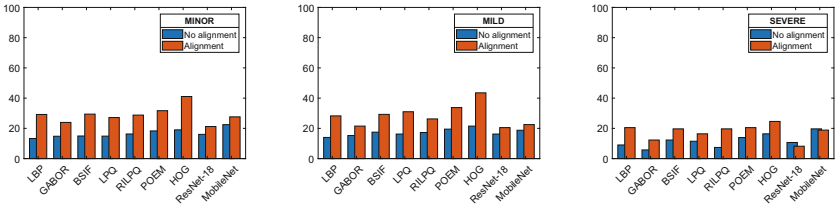

Fig. 3. Recognition rates (%) at rank-1 for different levels of occlusion using aligned and non-aligned images of AWEx dataset.

is not able to solve it. In contrast, the deep learning process is more capable of lead with this phenomenon thanks to the training process, where ear images with several occlusions were included.

Ear Rotation. Figure 4 illustrates the recognition rates at rank-1 under different rotation variations (roll, yaw, pitch), where for most of the descriptors a remarkable improvement is obtained through alignment step in almost all the cases and angles, being the HOG the best overall descriptor.

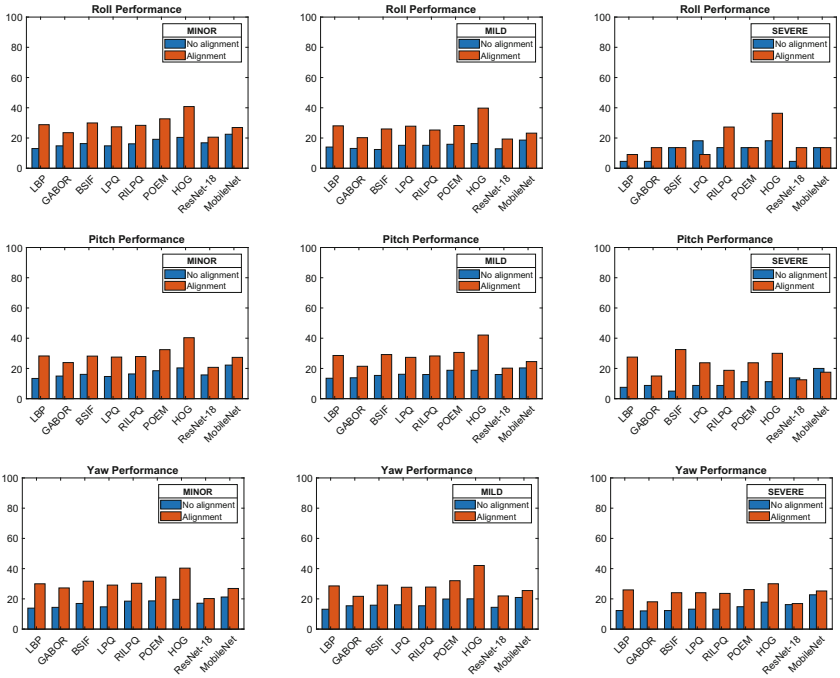

Fig. 4. Recognition rates at rank-1 across the pitch, roll and yaw rotation angles.

It can be seen that, severe roll and pitch angles impact more negatively on the recognition performance. For example, in case of roll rotations, the alignment shows it worse behavior due to this is one of the greatest challenges for any alignment method. However, in the case of pitch angles, the recognition rates of hand-crafted descriptors increase considerably by using aligned images. Similar to previous experiments, the improvements for hand-crafted descriptors are higher than for deep descriptors, especially for extreme variations.

Image Resolution. In this experiment we asses the impact of alignment in front different resolutions on UERC test dataset. In Table 3 presents the results in terms of recognition rates at rank-1 for all the descriptors. As can be seen, aligning images helps to increase the recognition performance in all cases for all tested resolutions. Smallest images with less than 1k pixels inevitably lead to performance degradation, hence, the improvements are smaller. However, as the resolution increases, they become more significant, especially for the hand-crafted descriptors. This fact is due to resolution significantly contributes to represent by the descriptors the details of the ear images.

Results achieved with resolution images between 5K and 10K pixels are similar to the results with images having more than 10K pixels, which suggests that images of at least 5K pixels are needed to obtain an adequate recognition performance. In these cases, HOG descriptor obtains the best results by using aligned images, while MobileNet is the best one when alignment is not applied.

Table 3. Recognition rates at rank-1 for different resolutions on UERC database.

	<1K (#4573)		1K-5K (#3883)		5K-10K (#412)		>10K (#632)	
	No Align	Align	No Align	Align	No Align	Align	No Align	Align
LBP	6.27	**11.27**	8.12	15.58	11.38	28.33	14.58	28.05
GABOR	3.16	5.92	5.27	8.58	13.08	22.28	13.79	21.87
BSIF	6.77	10.94	**8.29**	14.48	15.01	28.33	15.05	29.64
LPQ	**7.19**	11.06	8.09	13.82	13.32	27.12	17.27	28.68
RILPQ	4.76	9.02	6.50	12.93	14.77	26.88	19.33	29.48
POEM	5.59	9.73	7.62	13.69	18.64	31.23	19.97	35.34
HOG	4.28	10.29	7.72	**16.27**	17.92	**42.37**	21.87	**42.95**
ResNet-18	4.23	4.82	5.67	7.66	17.43	17.92	14.58	19.18
MobileNet	5.71	6.71	7.82	9.38	**22.28**	24.21	**23.45**	26.30

5 Conclusion

In this work an exhaustive analysis was carried out to evaluate the impact of ear alignment on the recognition performance of several state-of-the-art techniques on unconstrained conditions, taking into account different covariates. For

this, we conduct several identification experiments for hand-crafted and deep-learning descriptors on the challenge UERC dataset by using both aligned and non-aligned images. As result, we evidence that the alignment is an important step in the ear recognition process to achieve better results. Specifically, we found that for the hand-crafted descriptors, the alignment has a greater impact than for the deep models trained with aligned images. It can be said that when the images are not aligned, the deep-learning descriptors achieve a more discriminative description of ears with severe covariates, although in most cases an improvement is obtained. We argue that deep models are more capable of learning extreme variations, especially when these affect the performance of the alignment method. Among the tested hand-crafted descriptors, the HOG was the most benefited with the use of alignment, obtaining the highest recognition rates in almost all cases. However, it was evidenced that to achieve high recognition rates, not only alignment is sufficient, robust descriptors are also necessary.

With this work we provide a better understanding about the impact of ear alignment step on unconstrained ear recognition and identify the most challenges covariates which affect it. Note that, in cases where deviations are minimal or medium and hand-crafted descriptors are used, we can expect great improvements in recognition by alignment, contrary to severe variations, where there is still work to be done especially for those cases where the images contain high roll rotations, severe occlusions and low resolutions.

References

1. Abaza, A., Harrison, M.A.F.: Ear recognition: a complete system. In: Biometric and Surveillance Technology for Human and Activity Identification X, vol. 8712, p. 87120N (2013)
2. Dodge, S., Mounsef, J., Karam, L.: Unconstrained ear recognition using deep neural networks. IET Biom. **7**(3), 207–214 (2018)
3. Emeršič, Ž., Križaj, J., Štruc, V., Peer, P.: Deep ear recognition pipeline. In: Hassaballah, M., Hosny, K.M. (eds.) Recent Advances in Computer Vision. SCI, vol. 804, pp. 333–362. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03000-1_14
4. Emeršič, Ž., et al.: The unconstrained ear recognition challenge. In: IEEE IJCB, pp. 715–724 (2017)
5. Emeršič, Ž., Štruc, V., Peer, P.: Ear recognition: more than a survey. Neurocomputing **255**, 26–39 (2017)
6. Emeršič, Ž., et al.: The unconstrained ear recognition challenge 2019-arxiv version with appendix. arXiv preprint arXiv:1903.04143 (2019)
7. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
8. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and hand-crafted features for unconstrained ear recognition. IET Biom. **7**(3), 215–223 (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

10. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Oravec, M., et al.: Mobile ear recognition application, pp. 1–4 (2016)
12. Pflug, A., Busch, C.: Segmentation and normalization of human ears using cascaded pose regression. In: Bernsmed, K., Fischer-Hübner, S. (eds.) NordSec 2014. LNCS, vol. 8788, pp. 261–272. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11599-3_16
13. Ribič, M., Emeršič, Ž., Štruc, V., et al.: Influence of alignment on ear recognition: case study on awe dataset. In: International Electrotechnical and Computer Science Conference (2016)
14. Shu-zhong, W.: An improved normalization method for ear feature extraction. IJSIP **6**(5), 49–56 (2013)
15. Yazdanpanah, A.P., Faez, K.: Normalizing human ear in proportion to size and rotation. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS, vol. 5754, pp. 37–45. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04070-2_5
16. Yuan, L., Zhao, H., Zhang, Y., Wu, Z.: Ear alignment based on convolutional neural network. In: Zhou, J., et al. (eds.) CCBR 2018. LNCS, vol. 10996, pp. 562–571. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97909-0_60
17. Zhou, Y., Zaferiou, S.: Deformable models of ears in-the-wild for alignment and recognition. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 626–633 (2017)