USTB-Helloear: A Large Database of Ear Images Photographed Under Uncontrolled Conditions

Yi Zhang[1], Zhichun Mu[1(✉)], Li Yuan[1], Chen Yu[2], and Qing Liu[1]

[1] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
mu@ies.ustb.edu.cn
[2] Xi'an Musheng Electronic Technology Co., LTD., Xi'an 710061, China

Abstract. The capabilities of biometric systems, such as face or fingerprint recognition systems, have recently made extraordinary leaps by the emergence of deep learning. However, due to the lack of enough training data, the applications of deep neural network in the ear recognition filed have run into the bottleneck. Therefore, the motivation of this paper is to present a new large database that contains more than 610,000 profile images from 1570 subjects. The main distinguishing feature of the images in this USTB-Helloear database is that they were taken under uncontrolled conditions with illumination and pose variation. In addition, all of individuals were required to not particularly care about ear occlusions. Therefore, 30% of subjects had the additional control groups with different level of ear occlusions. The ear images can be utilized to train a deep learning model of ear detection and recognition; moreover, the database, along with pair-matching tests, provides a benchmark to evaluate the performances of ear recognition and verification systems.

Keywords: Biometrics · Deep learning · Ear recognition
Uncontrolled conditions · Ear database

1 Introduction

Ear based human recognition technology is an important research field in biometric identification. Compared with classical biometric identifiers such as fingerprints, faces, and irises, using an ear has its distinctive advantages. An ear has a stable and rich structure that changes little with age and does not suffer from changes in facial expressions at the same time [1]. Moreover, it is easy and non-intrusive to collect ear images in the application scenarios. As such, ear biometrics has recently received some significant attention.

Researchers have developed several 2D ear recognition approaches in the early years [2–4]. Most of existing ear recognition techniques are based on manually designing features or shallow learning algorithms. However, researchers found that most of these techniques performed poorly when the test images were photographed under uncontrolled conditions. Nevertheless, occlusion, pose and illumination variation are very common in practical application. Therefore, this puts forward a challenging problem, which must be addressed.

Y. Zhao et al. (Eds.): ICIG 2017, Part II, LNCS 10667, pp. 405–416, 2017.
https://doi.org/10.1007/978-3-319-71589-6_35

In the last decade, the algorithms based on deep learning have significantly advanced the performance of state-of-the-art in computer vision. Numbers of vision tasks such as face recognition [5–12], image classification [13–19] and object detection [20–22] have obtained a series of breakthroughs via deep learning models. Face recognition and verification as an example, Facebook trained a deep CNN model utilizing 4.4 million labeled face images. They achieved the best performance on the Labeled Faces in the Wild (LFW) benchmark [23] at the time. [24] proposed a VGG-Face model which was trained on 2.6 million images. Furthermore, the Google FaceNet [12] utilized 200 million labeled face images for its training. Some researchers turned to propose lightened deep models with less labeled data [25] or transfer learning methods [26] to solve the small sample size problem. However, as to the human ear recognition, the existing labeled ear images are so limited that even insufficient for the transfer learning.

To solve this problem, a new large database that contains more than 610,000 profile images from 1570 subjects is present in this paper. These images are extracted from video sequence. The main characteristic of this ear database is all of the images were photographed under uncontrolled conditions with illumination and pose variation. Furthermore, 30% of subjects in this database had an additional control groups with different level of ear occlusions. The ear images can be utilized to train a deep learning model of ear detection or recognition; moreover, the proposed database, along with pair-matching tests, provides a benchmark to evaluate the performances of ear recognition and verification systems.

The rest of this paper is structured as follows: a review of related work is given in Sect. 2, and Sect. 3 overviews the existent databases of ear images. In Sect. 4, a detailed description of the USTB-Helloear database is present. A series of experiments and comparisons can be found in Sect. 5. Finally, Sect. 6 provides the conclusions.

2 Related Works

Current ear recognition approaches exploited 2D images (including range images) or 3D point cloud data. In this section, we discuss some well known or recent 2D ear recognition methods utilizing 2D images or range images.

The existing ear recognition methods can be categorized into the holistic methods and the local feature based methods. The holistic methods utilized global features or statistical measures to classify ears. A force field transformation based technique was developed by Hurley et al. [27]. They generated the force field from ear images utilizing the Gaussian function. The directional properties of the force field were exploited to locate potential energy wells, which form the basis of the characteristic vector. Arbab-Zavar and Nixon [28] utilized the log-Gabor filter to exploit the frequency content of the ear boundary curves. A specific aim of this approach was to obtain the information in the ear's outer structures. Abate et al. [29] proposed a rotation invariant descriptor, namely GFD (Generic Fourier Descriptor), to extract features from ear images. This descriptor was robust to both ear rotations and illumination changes.

Researchers also proposed several ear recognition methods utilizing local feature descriptors. Kisku et al. [30] utilized the SIFT feature descriptors for the ear structural representation. The SIFT key points were extracted and an augmented vector of extracted SIFT features were created for matching. In [31], the SURF feature extraction was carried out on ear images to obtain three sets of local features, three nearest neighbor classifiers were trained on these three sets of features. Matching scores generated from all the three classifiers were fused for the final decision. Yuan and Mu [32] proposed a 2D ear recognition approach based on local information fusion to deal with ear recognition under partial occlusion. They separated the whole image to sub-windows and extracted local feature on each sub-windows. Finally, a sub-classifier fusion approach was used for recognition with partially occluded images. Chen and Mu [33] proposed a weighted multi key point descriptor sparse representation-based classification method to use local features of ear images. By adding adaptive weights to all the key points on a query image, the intra-class variations were reduced.

It is worth noting that most of the mentioned ear recognition works were tested on images that were photographed under controlled conditions. The recognition rates may have sharply dropped when those systems were applied in a realistic scenario, which contains occlusion, illumination variation, scaling, and rotation.

3 Related Databases

Most widely-used standard image databases for ear recognition systems are described in brief below.

3.1 USTB Databases

The USTB ear database contains 4 subsets which were collected by University of Science and Technology, Beijing. The USTB database I, II, and III are available under license. All of the databases were collected under controlled condition with single background. The presented USTB-Helloear database in this paper is the fifth database.

USTB database I: There are 180 ear images from 60 subjects in this database. Every volunteer was shot three different images. They are normal frontal image, frontal image with trivial angle rotation and image under different lighting condition.
USTB database II: This collection contained 308 right ear images from 77 volunteers. For each subject, there were 4 images with pose and lighting variation.
USTB database III: In this dataset, 79 subjects were photographed in different poses. There were total 785 images in this dataset, and some of the ears were occluded by hair.
USTB database IV: This database contained 500 subjects. 17 CCD cameras placed round the individual at every 15° and images of the face and ear were captured.

3.2 UND Databases

Those databases were collected by University of Notre Dame. All of the UND databases are available to the public (free of charge). All of the UND databases were collected under controlled condition with a single background.

Collection E: There were 942 profile images of 302 people in 3D and 2D images.
Collection F: This collection total consisted of 464 side face 3D and 2D images of 114 subjects.
Collection G: This collection had 738 side face images of 235 peoples in 3D and corresponding 2D images.
Collection J2: The collection had 1800 profile images from 415 subjects in 3D and corresponding 2D images.

3.3 WPUT Database

This database was collected by The West Pomeranian University of Technology. The database consisted of 2071 images from 501 subjects. For each subject, the database contains 4 to 8 images, which were taken under different lighting conditions. Moreover, there were earrings and hair occlusions in some images.

3.4 UBEAR Database

This database consisted of 9121 profile images from 242 subjects. The images in this database were taken under varying lighting conditions while subjects were moving. In addition, no particular care was required regarding ear occlusions and poses. The ground truth of the ear's location was provided, which made it particularly convenient for researchers to study the accuracy of ear detection.

4 USTB-Helloear Database

In cooperation with Xi'an Musheng Electronic Technology Co., LTD., we present a large scale collection of ear images along with labels and pair-matching tests. In this section, a detailed description of the USTB-Helloear database is provided. The images in this database are extracted from video sequence. The entire database is divided into two subsets. There are 336,572 profile images from 1104 subjects in subset A and 275,909 profile images from 466 subjects in subset B. The more detailed description is provided in Tables 1 and 2.

Table 1. Detailed description of video and images in USTB-Helloear database.

Video acquisition parameters	
Camera	Iphone 6s
Focal length	29 mm
Aperture	f/2.2
Video resolution	1980 * 1080 pixels
Frames per second	30
Videos codec	MOV
Details of the images	
Image resolution	1980 * 1080 pixels
Image codec	JPG

Table 2. Overview of the ear database.

Attribute	Range
No. of subjects	1570 (34.7% female, 65.3% male)
No. of photos	612,661
Age of subjects	11–18 (32.9%)
	19–21 (46.8%)
	22–26 (13.5%)
	27 and above (4.8%)
Occlusions in subset B	Minor (37.4%)
	Normal (42.9%)
	Major (19.7%)
Type of occlusions in subset B	Earphones (13.7%)
	Hair (86.3%)

Ear images in subset A only contain pose variations. For every subject in subset A, about 150 images on average per one ear are extracted from a 10 s video. Both left and right ears of every subject are photographed so that there are about 300 images from one person on average in subset A. As shown in Fig. 1, for each 10 s video, the camera moves around the ear to get ear pictures from different views. In the first 5 s, the camera moves from the front to back (Fig. 1(a)); In the rest 5 s, the camera moves from the top to bottom (Fig. 1(b)). In this 10 s video, We utilize the relative movement between the camera and ear to simulate the pose variations of the human ear under uncontrolled conditions. Therefore, the profile images in subset A extracted from the videos can be utilized to evaluate the performance of ear detection and recognition systems with pose variations. Some examples in subset A are illustrated in Fig. 2(a).

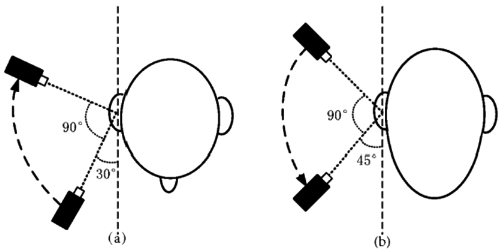

Fig. 1. The camera shoots the ear from different views. (a) The camera moves from the front to back. (b) The camera moves from the top to bottom.

Among all the 1570 volunteers, 30% of them had different level of ear occlusions. As mentioned above, all of individuals were required to not particularly care about ear occlusions. Therefore, we collected subset B of USTB-Helloear database which contained 466 subjects with pose variations and ear occlusions.

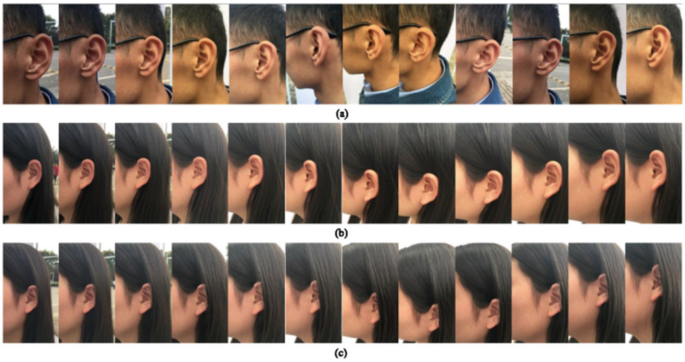

Fig. 2. Some example of images in the USTB-Helloear database. (a) Images in subset A. (b) Images without ear occlusion in subset B. (c) Images with ear occlusions in subset B.

For every subject in subset B, two 10 s videos were shot per one ear. Firstly, in the first video, the natural occluded ear pictures were photographed with different poses. Then we take the second video of the same ear without occlusion from different views. Finally, 150 images are extracted from each video sequence. It's worth noting that, the way of camera move around the ear in subset A and subset B are all the same (as shown in Fig. 1). Therefore, each of the ears in subset B consists of two sets of images, one set with pose variations, and another set with both pose variations and hair or earphone occlusions. This subset of USTB-Helloear database can be utilized for training and evaluating ear detection and recognition models. Examples of images in subset B are illustrated in Fig. 2(b) and (c).

5 Experiments

In this section, we evaluate our database under two scenarios: ear recognition and ear verification. Several deep learning models are trained and tested to evaluate the proposed database. Every subject has left and right ear images in this database. As we know, The left and right ears of a same person are not exactly the same. Therefore, we train and test ear recognition models on left ear images and the matching pairs for ear verification are generated from right ear datasets.

5.1 Ear Recognition

The images in the USTB-Helloear database are 2D profile images. The ear regions have to be detected and extracted from the profile images before recognition procedure. In this paper, the Multiple Scale Faster RCNN algorithm which we proposed in [35] is utilized to detect ears. Examples of extracted ear regions from subset A subset B are shown in Fig. 3.

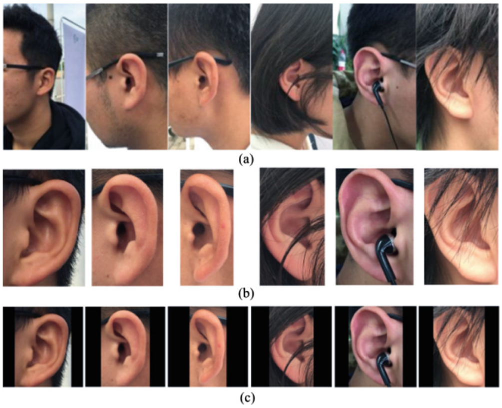

Fig. 3. Examples of extracted ear regions from subset A subset B. (a) Original profile images in USTB-Helloear database. (b) The extracted ear regions from subset A subset B. (c) The normalized ear images.

We train several deep learning models utilizing ear images from the USTB-Helloear database to evaluate the database. As we know, the input size of CNN architecture must be unified. As showed in Fig. 3(b), the image size and the aspect ratio of the extracted ear regions are varied due to the different shape and pose of the ears. Therefore, before we input the images into the CNN, we fill the images to square images and then resize them to 256 * 256 (Fig. 3(c)).

We fine-tune the VGG-Face pre-trained model [24] on subset A and subset B successively. Firstly, the pre-trained model is trained on subset A. We divide all of the images in the subset A into 5 parts, then we train 5 models utilizing the 5-fold cross validation method. During every training process, 4 parts are utilized as training data and the last part is used as test data. After 5 times training, all of the images are trained and tested, then the average recognition rate of 5 models is reported as the final recognition rate. Because the ear images in the presented database are extracted from video sequence, the neighboring images are similar to each other. If we divide all of the images into 5 subsets randomly, the trained model might be over-fitting. Therefore, The 5 subsets are divide in sequence.

The experiments are based on Caffe framework [36] and implemented on a work station with four Titan X GPUs. A batch size of 64 and initial learning rate of 0.001 are used. The last layer is trained from scratch, so that the learning rate of this layer is 0.01. During training, we randomly crop a 224 * 224 pixel square patch and feed it into the network to ameliorate the diversity of training data. The recognition rates are given in Fig. 4.

It is shown that, the average recognition rate of 5 models is 98.18%. The trained deep models are robust to ear recognition with pose variation. Therefore, we then fine-tune this pre-trained model on subset B to get deep models which are both robust to pose variation and occlusions. One of the trained models with the highest recognition rate is utilized to be fine-tuned on subset B. The parameters of this network are the same as previous networks. As mentioned above, different from the images in subset A,

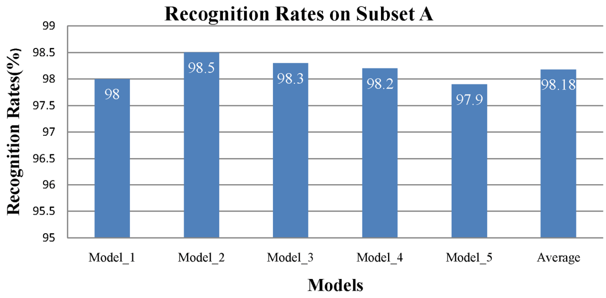

Fig. 4. The recognition rates on subset A.

every ear in subset B has 300 images on average. Half of the images are ear images with pose variation, and another half part of the images are control groups with different level of ear occlusions. We divide all of the images in subset B into 5 parts in sequence. In every part, half of the data are ear images with pose variation and another half are control groups. Then we also train 5 models utilizing 5-fold cross validation method. Finally, the average recognition rate of 5 models is reported as the final recognition rate. The recognition rates are 97.9% (Fig. 5). In the next section of this paper, we will discuss the ear feature representation capacity of the models mentioned above.

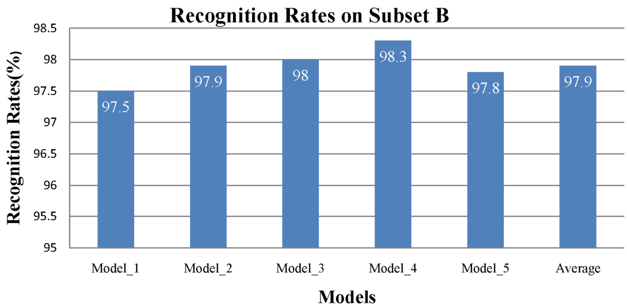

Fig. 5. The recognition rates on subset B.

5.2 Ear Verification

The CNN deep model can be utilized as a feature extractor to get the ear feature representation vector. In [9], Sun et al. used the Joint Bayesian technique for face verification based on the face feature representation vector extracted from a CNN. In this paper, two images are fed in to the trained model and the last layers are extracted as feature vectors. The cosine distance is utilized to measure the similarity of two ear feature vectors.

As a benchmark for comparison, we present pair-matching tests rules, which provide benchmarks to evaluate the performances of ear verification systems.

3000 match pairs and 3000 mismatch pairs are randomly generated from subset A. We also randomly generate 3000 match pairs and 3000 mismatch pairs on subset B which both contain pose variation and occlusions. Lists of pair-matching test will be provided along with the USTB-Helloear database. Researchers can test and compare their ear verification algorithm on those pairs. Some challenging examples are shown in Fig. 6.

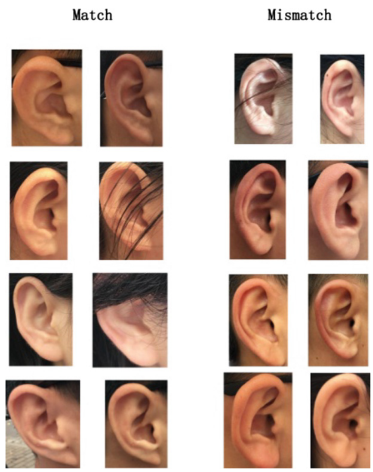

Fig. 6. Some challenging examples in the validation set.

In this section, we evaluate the two trained deep models with highest recognition rate on subset A and subset B. For convenience, we call them *Model_A* and *Model_B* respectively. The lists of pair-matching test are called *Pairs_A* and *Pairs_B* as well. The ear verification results are presented in Table 3. The ROC curves are also showed in Fig. 7.

Table 3. The ear verification results.

	Pairs_A	Pairs_B
Model_A	92.6%	82.83%
Model_B	94.67%	88.5%

It is shown that, The *Model_B* has achieved the best ear verification result both on benchmarks of *Pairs_A* and *Pairs_B*. The stronger capacity of ear feature representation can be obtained via feeding the net with the more challenging training data in subset B. In addition, The fact of the *Model_B* perform better than *Model_A* on the benchmark of *Pairs_A* indicate that the trained deep models don't over-fit the training data. Therefore, we can draw the conclusion that, the ear images in the proposed USTB-Helloear database can satisfy the meet of training and testing ear recognition systems.

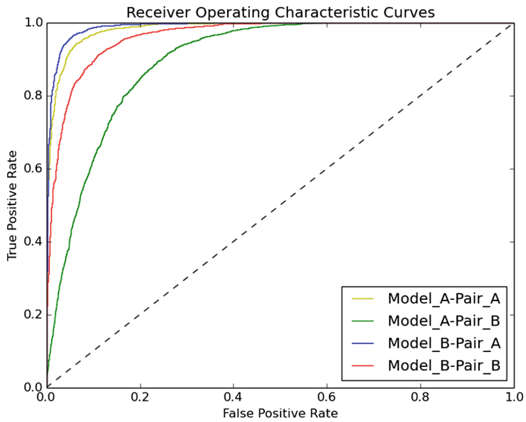

Fig. 7. The ROC curves of the present models on two validation set.

6 Conclusions

In this paper, we present a new large ear database which can be utilized to train and test ear recognition and verification systems. The images in this database were taken under uncontrolled conditions with illumination and pose variation. All of individuals were required to not particularly care about ear occlusions. Therefore, researchers can utilize the images to train a deep learning model to represent ear feature. The experiments demonstrate that, the capacity of ear feature representation can be obtained via feeding the CNN with images in this database. This database will be public and freely available from our web site: http://www1.ustb.edu.cn/resb/en/index.htm.

Acknowledgement. This article is supported by the National Natural Science Foundation of China (Grant No. 61472031). The authors would like to thank the Xi'an Mu sheng Electronic Technology Co., LTD. for their cooperation.

References

1. Jain, A., Flynn, P., Ross, A.A.: Handbook of Biometrics, pp. 131–150. Springer Science & Business Media, Berlin (2007). https://doi.org/10.1007/978-0-387-71041-9
2. Yuan, L., Mu, Z., Xu, Z.: Using ear biometrics for personal recognition. In: Li, S.Z., Sun, Z., Tan, T., Pankanti, S., Chollet, G., Zhang, D. (eds.) IWBRS 2005. LNCS, vol. 3781, pp. 221–228. Springer, Heidelberg (2005). https://doi.org/10.1007/11569947_28
3. Yuan, L., Mu, Z.C.: Ear recognition based on local information fusion. Pattern Recogn. Lett. **33**(2), 182–190 (2012)
4. Zhang, B., Mu, Z., Li, C., et al.: Robust classification for occluded ear via Gabor scale feature-based non-negative sparse representation. Opt. Eng. **53**(6), 061702 (2014)
5. Taigman, Y., Yang, M., Ranzato, M.A., et al.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)

6. Zhou, E., Cao, Z., Yin, Q.: Naive-deep face recognition: touching the limit of LFW benchmark or not? arXiv preprint arXiv:1501.04690 (2015)
7. Yi, D., Lei, Z., Liao, S., et al.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
8. Sun, Y., Chen, Y., Wang, X., et al.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996 (2014)
9. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
10. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2892–2900 (2015)
11. Sun, Y., Liang, D., Wang, X., et al.: Deepid3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
12. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
14. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
15. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
19. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 (2016)
20. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
21. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
22. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Huang, G.B., Ramesh, M., Berg, T., et al.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07-49, University of Massachusetts, Amherst (2007)
24. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. BMVC **1**(3), 6 (2015)
25. Wu, X., He, R., Sun, Z.: A lightened cnn for deep face representation. In: 2015 IEEE Conference on IEEE Computer Vision and Pattern Recognition (CVPR) (2015)
26. Yosinski, J., Clune, J., Bengio, Y., et al.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)

27. Hurley, D.J., Nixon, M.S., Carter, J.N.: Force field energy functionals for image feature extraction. Image Vis. Comput. **20**(5), 311–317 (2002)
28. Arbab-Zavar, B., Nixon, M.S.: Robust log-Gabor filter for ear biometrics. In: International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
29. Abate, A.F., Nappi, M., Riccio, D., et al.: Ear recognition by means of a rotation invariant descriptor. In: International Conference on Pattern Recognition, pp. 437–440. IEEE (2006)
30. Kisku, D.R., Mehrotra, H., Gupta, P., et al.: SIFT-based ear recognition by fusion of detected keypoints from color similarity slice regions. In: 2009 International Conference on Advances in Computational Tools for Engineering Applications, ACTEA 2009, pp. 380–385. IEEE (2009)
31. Prakash, S., Gupta, P.: An efficient ear recognition technique invariant to illumination and pose. Telecommun. Syst. **52**, 1435–1448 (2013)
32. Yuan, L., Mu, Z.C.: Ear recognition based on local information fusion. Pattern Recogn. Lett. **33**(2), 182–190 (2012)
33. Chen, L., Mu, Z.: Partial data ear recognition from one sample per person. IEEE Trans. Hum.-Mach. Syst. **46**, 799–809 (2016)
34. Frejlichowski, D., Tyszkiewicz, N.: The West Pomeranian university of technology ear database – a tool for testing biometric algorithms. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, pp. 227–234. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13775-4_23
35. Zhang, Y., Mu, Z.: Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. Symmetry **9**(4), 53 (2017)
36. http://caffe.berkeleyvision.org/