Multimedia Tools and Applications (2022) 81:22695–22713
https://doi.org/10.1007/s11042-021-10630-8

1169: INTERDISCIPLINARY FORENSICS: GOVERNMENT, ACADEMIA AND INDUSTRY INTERACTION

Multimodal soft biometrics: combining ear and face biometrics for age and gender classification

Dogucan Yaman[1] · Fevziye Irem Eyiokur[1] · Hazım Kemal Ekenel[1]

Accepted: 3 February 2021 /
Published online: 15 March 2021

Abstract

In this paper, we present a multimodal, multitask deep convolutional neural network framework for age and gender classification. In the developed framework, we have employed two different biometric modalities: ear and profile face. We have explored three different fusion methods, namely, data, feature, and score fusion, to combine the information extracted from ear and profile face images. In the framework, we have utilized VGG-16 and ResNet-50 models with center loss to obtain more discriminative features. Moreover, we have performed two-stage fine-tuning to increase the representation capacity of the models. To assess the performance of the proposed approach, we have conducted extensive experiments on the FERET, UND-F, and UND-J2 datasets. Experimental results indicate that ear and profile face images contain useful features to extract soft biometric traits. We have shown that when frontal face view of the subject is not available, use of ear and profile face images can be a good alternative for the soft biometric recognition systems. The presented multimodal system achieves very high age and gender classification accuracies, matching the ones obtained by using frontal face images. The multimodal approach has outperformed both the unimodal approaches and the previous state-of-the-art profile face image or ear image-based age and gender classification methods, significantly in both tasks.

Keywords Multimodal learning · Multitask learning · Soft biometrics · Age estimation · Gender classification · Convolutional neural networks

Dogucan Yaman and Fevziye Irem Eyiokur have equally contributed.

Dogucan Yaman and F. Irem Eyiokur did this work as students at Istanbul Technical University. Now, they are working at Karlsruhe Institute of Technology.

✉ Dogucan Yaman
yamand16@itu.edu.tr

Fevziye Irem Eyiokur
eyiokur16@itu.edu.tr

Hazım Kemal Ekenel
ekenel@itu.edu.tr

[1] Istanbul Technical University, Istanbul, Turkey

1 Introduction

Extracting soft biometric traits is an important research topic in biometrics [16, 17, 29, 33]. It has been found that by utilizing soft biometric traits, subjects can be described better and this way the identification performance can be improved [16, 17, 24]. Two of the most widely used soft biometric traits are subjects' age and gender.

Recently, with the convolutional neural networks (CNN) based approaches, superior results have been obtained for person identification, age estimation, and gender classification using ear and profile face images [22, 28, 35, 36]. There have been several works about extracting soft biometric traits, age and gender, using these modalities [2, 12, 14, 18, 21, 23, 25, 27, 30, 35]. In addition to the unimodal approaches, there has been one multimodal work that utilized both ear and profile face images to perform gender classification [39]. Most of these previous works have focused on gender classification. There is only one recent work [35] that has performed age classification from ear images, one that has performed age estimation from profile face images [5], and one multimodal approach [36].

Frontal face images have been used in many different biometric studies [3, 6, 8, 9, 22, 24, 28]. However, depending on the application, it may not be possible to always obtain frontal face images. For example, in a surveillance scenario, it would be very rare to acquire frontal faces of the subjects. In these situations, profile face and ear could be useful as alternative biometric modalities. Besides, ear biometrics has certain advantages, for instance, toleration to lighting changes compared to profile and frontal face images [10].

In this paper, extending our previous work [36], we have presented a comprehensive study on age and gender classification using ear and profile face images as input biometric modalities. We have developed and investigated several multimodal and multitask deep CNN frameworks. We have explored three different fusion methods: data fusion —intensity fusion, spatial fusion, channel fusion—, feature fusion, and score-level fusion. In the study, we have employed two well-known deep CNN models, namely VGG-16 [32] and ResNet-50 [13]. To obtain more discriminative features, besides softmax loss, we have also benefited from center loss. In addition, we have performed domain adaptation via a two-stage fine-tuning approach to increase the representation capability of the models. In this work, in addition to [36], we examined age regression which is one of the most important task in soft-biometric analysis. We also deeply analyzed the performances such as $\pm$ age classification accuracy, effect of context information over performance, effect of accessories, and etc. Besides, we trained a model using frontal face data for age and gender classification on the datasets that are used in this work in order to compare different biometric modalities. Please note that, profile face images can already contain ear region, when they are cropped using a large bounding box. However, when we enlarge the bounding box of the profile face images, then hair and background information are also included. In our experiments we have observed that including these irrelevant information degrades the performance.

We have conducted extensive experiments on the UND-F, UND-J2, and FERET datasets [26, 37]. Experimental results have shown that profile face images contain a valuable source of information for age and gender classification. The proposed multimodal system has achieved very high age and gender classification accuracies. Moreover, we attained superior results compared to the state-of-the-art profile face image or ear image-based age and gender classification methods.

The contributions of the study can be summarized as follows:

– We have presented a multimodal, multitask deep convolutional neural network approach for age and gender classification.

- We have thoroughly explored several ways of benefiting from multimodal input for age and gender classification. We have investigated three different data fusion methods, as well as feature and score level fusion.
- We have adapted the utilized deep CNN models, namely VGG-16 and ResNet-50, to the ear domain by using a two-stage fine-tuning approach. For this purpose, we have generated the extended version of the Multi-PIE ear dataset that was presented in our previous work [35] and named it Multi-PIE extended-ear dataset. Moreover, we have employed center loss in combination with the softmax loss to learn more discriminative features.
- We have provided class activation maps to observe the CNN behaviour under different circumstances.
- We have conducted a comprehensive experimental analysis. We have used the UND-F, UND-J2, and FERET datasets for gender classification, and only the FERET dataset for age classification, since the UND-F and UND-J2 datasets do not contain age labels. We have achieved state-of-the-art age and gender classification results on these datasets.
- We have also performed age and gender classification experiments using the frontal face images of the same subjects in order to determine the usefulness of profile face and ear images as an alternative to the frontal face image. We have found that the accuracies achieved by the proposed multimodal approach are very close to the ones obtained by using the frontal face image as input.

The remainder of the paper is organized as follows: In Section 2, we provide a brief overview of the previous works on the topic. In Section 3, we explain the used CNN architectures, proposed fusion methods, and two-stage fine-tuning strategy. In Section 4, we present the datasets, experimental setups, and experimental results. Finally, Section 5 concludes the paper.

2 Related work

In this section, we have reviewed the previous studies that employed ear and profile face images for age estimation and gender classification.

In [12], feature extraction is performed on ear images by calculating distances between identified seven points and ear-hole. The internal dataset, which contains 342 ear images, is employed for the experiments and 90.42% gender classification accuracy is achieved with k-nearest neighbor classifier. In [39], the unimodal and multimodal experiments are conducted with ear and profile face modalities using support vector machines (SVM) for gender classification. The experiments are conducted on the UND-F dataset. While the multimodal system achieves 97.65% accuracy, ear-only and profile face-only performances are 91.78% and 95.42%, respectively. According to this work, while profile face is found to be more useful than ear, multimodal system's performance surpasses the accuracies obtained by both of the modalities. In [18], the UND-J2 dataset, which is the extended version of the UND-F dataset, is used to study gender classification task using ear images. 89.49% classification accuracy is obtained with majority voting using Gabor filter as a feature extractor. In [21], the UND-F and UND-J2 datasets are employed for gender classification experiments. In contrast to the previous works, the authors used both 2D and 3D ear images for the experiments with SVM as the classification method and Histogram of Indexed Shapes as a feature extractor. They achieved 92.94% classification accuracy with 3D ear data. Besides, they showed that 3D ear images contain rich information and are more useful than the 2D ear images for gender classification. In [5], profile face images are used as an alternative modality to the frontal face. ResNet-50, ResNet-101, and ResNet-152 are utilized as feature extractors and sparse partial least-squares regression is employed. The best age estimation

result, 5.50 mean absolute error (MAE), on FERET dataset is achieved with ResNet-152 features. In [35], age and gender classification experiments are presented using ear images. Distance measurements between ear landmarks are used for the geometric-based representation. Extracted geometric-based representations are then utilized as inputs to several classification algorithms. Appearance-based representation is learned with CNNs.The highest accuracies are achieved using appearance-based representation, and they are 52% for age classification and 94% for gender classification. In [23], the authors addressed gender classification task with both deep learning-based and model-based methods. The evaluations are conducted on the USTB dataset [15] and their own dataset. The obtained gender classification performance is 82.9% with the model-based approach and 93.8% with the deep learning-based transfer learning approach. Besides, the authors investigated the effect of different parts of the ear on the gender classification performance. They found that the upper helix is significantly more important than the middle and lower helix for gender classification. In [36], ear and profile face images are employed to perform age and gender classification. In this work, different fusion methods are explored using VGG-16 and ResNet-50 deep CNN architectures. The proposed methods are tested on benchmark datasets. While age classification accuracy on FERET dataset is 67.59% with VGG-16 model using spatial fusion, the gender classification is 99.11% on FERET, 100% on UND-F, and 99.79% on UND-J2 datasets.

3 Methodology

In this work, we have developed a multimodal, multitask learning framework to combine ear and face biometrics and to benefit from the relationship between age and gender information. We have investigated both unimodal approaches, ear-only and profile face only, and proposed a multimodal approach. One might think that profile face images already contain ear appearance, therefore, instead of having a multimodal approach, it would be sufficient to have a larger crop of a profile face including the ear region and process this input image in a unimodal way. However, there are two drawbacks of such an approach. The first one is due to the noise, for example, hair and background regions are included when having a larger crop, which degrades the performance. Moreover, since ear is a small area on the profile face, the deep CNN model may extract less amount of features from ear region. By taking these factors into account, we preferred to have a tight crop of a profile face, which does not contain any part of the ear, and ear region separately.

3.1 Learning approaches

In order to extract age and gender information from ear and profile face images, we focus on benefiting from ear and profile face both individually and jointly. While the CNN model is fed with the ear or profile face image in the unimodal approach, different combinations of them are provided to the deep CNN models in the multimodal approach. With this work, the superiority of the multimodal approach over unimodal ones is demonstrated.

Additionally, all the deep networks are trained in a multitask learning manner, which is useful to exploit the relation between age and gender information. In the end of the CNN model, the age and gender classifications are performed using the same features, and different loss functions are calculated for each task. Later, the combination of individual loss values are computed as a final total loss. This combined loss is used in backpropagation step to update all model weights. Although the combination of age and gender cost functions

makes the final cost function complex to optimize, the age and gender estimation results in the multitask approach are similar with using both tasks individually according to our experiments. This inference indicates that there is a joint feature set that can be utilized to extract age and gender information effectively. This multitask approach can also help to gain from processing time and reduce storage & computational costs.

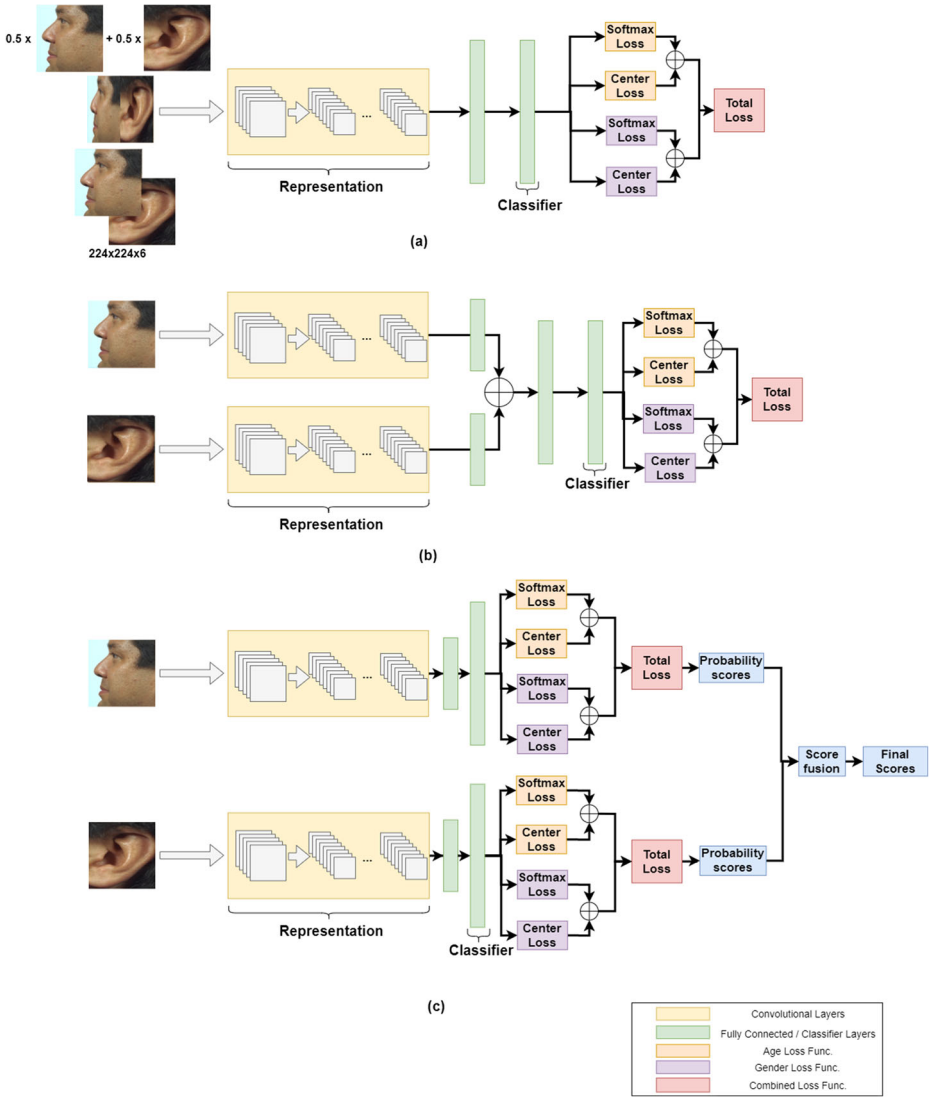

Fig. 1 Proposed multimodal, multitask age and gender prediction system. **a** shows data fusion method. Three different data fusion methods which are intensity, spatial, and depth, are performed separately. **b** shows feature fusion method. Two different networks take ear and profile face images separately. Then, the representation parts, which are convolutional parts of the networks, process images and extract features. These extracted features from different modalities are concatenated in the end of the convolutional part. **c** shows the score fusion method. Two different CNN models, that take different modalities as input, are trained as an end-to-end system. Then, in the test phase, final prediction for each sample is calculated by combining the scores obtained from both networks

The proposed multimodal, multitask approach is illustrated in Fig. 1. In Fig. 1a, three different data fusion methods are shown. A CNN model takes combined data to extract features and learn relationship between these features and age & gender in a multitask manner. The visualized three data fusion methods are performed individually. In Fig. 1b, two separate convolutional parts of the employed CNN models are utilized to extract features from profile face and ear images with the intent of performing feature fusion in the next step. Afterwards, the extracted feature maps are flattened and concatenated, and then transferred to the learning part. In Fig. 1c, two individual CNN models are trained with profile face and ear images in an end-to-end manner. Fusion is performed by using the probability scores of these individual models.

3.2 Fusion types

In order to investigate the multimodal approach, three different fusion methods are performed to utilize two different modalities, ear and profile face. The effect of these fusion methods on the performance is examined in detail. In the following sections, the explanation of these fusion types are presented.

3.2.1 Data fusion

To combine two modalities at the data level, three different data fusion methods are proposed. These fusion methods: intensity, spatial, and depth fusion are illustrated in Fig. 2.

Intensity fusion is a fusion method that is used to combine ear and profile face images. In this method, ear and profile face images' intensity values are multiplied by 0.5 and summed up to obtain the final image. In fact, this is a pixel-wise average version of both modalities.

Spatial fusion is another version of the employed data fusion. In this fusion method, the ear and profile face images are concatenated side-by-side. The generated image contains both ear and profile face images. That is, left half part of the concatenated image is a profile face image, while the right half part of it is an ear image. Although the aspect ratios of the ear and profile face images are changed for concatenation, it was the best fusion method in terms of age and gender classification performance.

Depth fusion is the last method used for data fusion. In this method, the ear and profile face images are concatenated in depth and the resulting image has six channels. While the first three channels contain profile face image, other three channels include ear image. In order to use this concatenation type in the model, the layer parameters are updated to take six channels input and the weights in the initialization step are duplicated.

After performing these data fusion methods, the obtained input images are then provided to the deep CNN models to learn correlation between extracted features for age and gender classification.

3.2.2 Feature fusion

In the feature fusion, ear and profile face images are passed through two different CNN models. While the first model takes profile face image as the input, the second one takes the ear image. At the end of the representation part (convolutional part) of the network, features

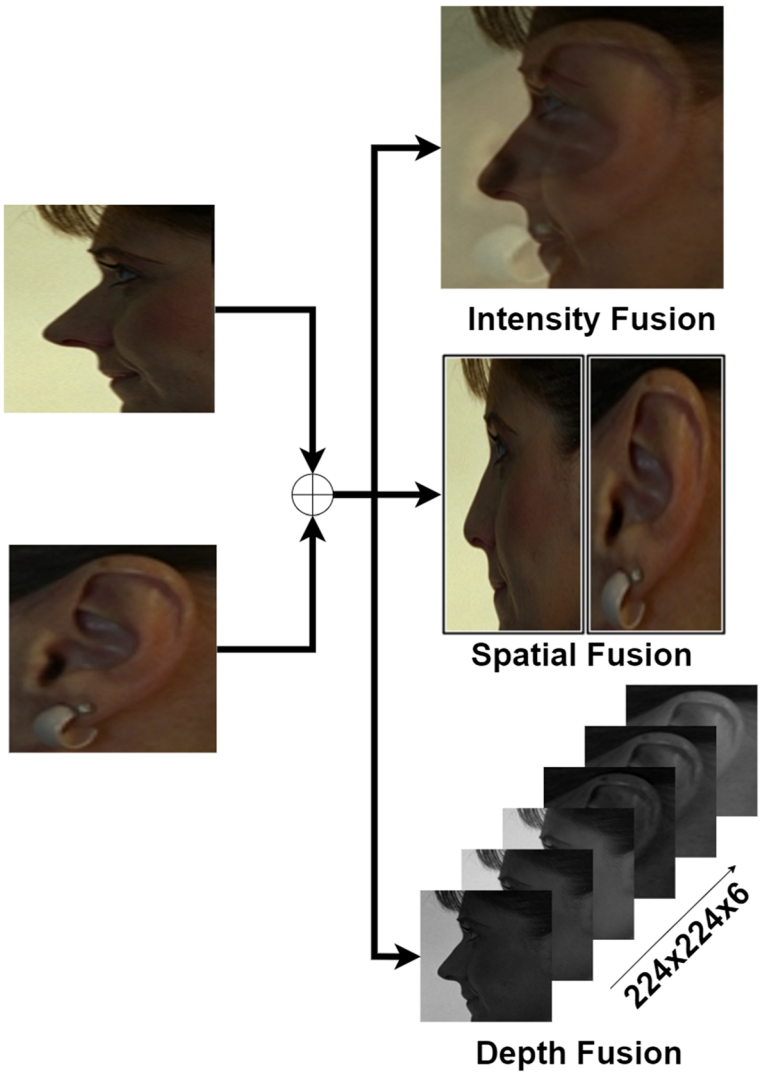

Fig. 2 Data fusion methods. Images are from FERET dataset [26]. In the intensity fusion, the average of profile face and ear images are computed as pixel-wise. While the profile face and ear are concatenated side-by-side in the spatial fusion, both modalities are combined along channels in the depth fusion. The images in the depth fusion visualization represent R, G, and B channels of the profile face image and ear image, respectively. In the end, the final image has six channels

are extracted and concatenated to obtain a single feature vector. Then, this combined feature vector is provided to the joint classifier part of the model. This lets us to train these two separate convolutional parts of the networks simultaneously.

Table 1 These formulas are used for calculating confidence score of the model. Then, score fusion is performed using these confidence scores. In the each formula, c represents confidence scores, while s represents the prediction scores of the model and it is sorted from descending order. In the last three formulas, M is the number of class

Method	Formula
Basic	$c = s[0]$
d2s	$c = s[0] - s[1]$
d2sr	$c = 1 - (s[1]/s[0])$
avg-diff	$c = \frac{1}{M-1} \sum_{i=1}^{M} (s[0] - s[i])$
diff1	$c = \sum_{i=1}^{M-1} (\frac{s[i-1]-s[i]}{i})$

3.2.3 Score fusion

In the score fusion method, basically two parallel CNN models are trained end-to-end separately. While one network is trained with ear data, the other one takes profile face image as input. In the test phase, we followed different approaches for age classification and regression. In the age classification experiment, the probabilities of both models are extracted to calculate confidence score of the models using score fusion formulas that are presented in Table 1. In the formulas, array s contains prediction scores of the classification model. This array is sorted in descending order. Later, confidence score c is calculated from the array s by applying the listed formulas. In the *basic* method, the confidence score is calculating based on the prediction that has highest score. In the *d2s*, the normalized difference between two best candidates is considered to calculate confidence score. In the *d2sr*, the ratio between best two candidates is taken into account. In the *avg-diff*, the mean of the difference between highest score and all scores except itself is measured to obtain confidence score. Finally, in the *diff1*, the separation between consecutive scores is considered. The CNN model that has the highest confidence score for an image is assumed as the most reliable model for the related image and its prediction is selected as the final prediction in the test phase. On the other hand, for the age regression, since the models are trained for regression purposes, instead of classification, the outputs of them are estimated ages. Therefore, we could not calculate confidence scores from them. Instead of this, we just applied basic score fusion method in a different manner, in which we calculated average of predictions of both models as a final result.

3.3 Transfer learning and training procedure

In this work, VGG-16[32] and ResNet-50 [13], which are well-known deep CNN architectures are utilized to perform age and gender classification from ear and profile face modalities. These two models are selected based on our experiments on the ear domain [11]. Due to the significant efficiency and performance improvement provided by transfer learning, as shown in many previous works [11, 24, 31, 38], it is also benefited to improve the performance of the proposed system. In order to perform transfer learning, at first, the network is initialized with the weights of the pretrained model, which was trained on the ImageNet dataset [7]. Afterwards, this initialized network is fine-tuned on target datasets to acquire age and gender predictions. Using weights of the pretrained network in the initialization step instead of random initialization leads to a more efficient training procedure

and a more powerful model, especially, when small datasets are used for training. Since our target datasets are not large enough, we obtained better performance using transfer learning instead of training from scratch. Moreover, in order to adapt pretrained CNN model to the ear domain, two-stage fine-tuning process is applied using Multi-PIE Extended ear dataset [36]. Multi-PIE Extended ear dataset [36] [1] is a large-scale ear dataset, which is the extended version of the Multi-PIE Ear dataset [11], and it contains around 40000 ear images of more than 200 subjects. For this adaptation, first, pretrained model on the ImageNet dataset [7] is fine-tuned on the Multi-PIE Extended ear dataset [36] in order to learn characteristics of the ear domain. Then, this model is employed for model initialization in the last fine-tuning step on the target datasets. According to the experimental results, this intermediate fine-tuning step helps deep CNN models to adapt to the ear domain and increases their performance.

In the training, the initial learning rate is selected as 0.0001. The model is trained for 100 epochs and the learning rate is decreased by 10 times in the epoch 50. For the training, Adam optimizer [20] is employed. B1 and B2 parameters are assigned as 0.9 and 0.999, respectively. For the experiments, we used TensorFlow framework [1] and NVIDIA GTX 1080 Ti GPU. While batch size is chosen as 32 for unimodal representation, it is selected as 16 for multimodal representation, since two networks are employed in the multimodal approach, instead of a single network. In the VGG-16 architecture, the network has 13 convolutional layers and 3 fully-connected (FC) layers after the convolutional part. The max pooling and ReLU activation function are utilized inside the network. Moreover, the dropout with 0.5 drop rate is employed after each fully-connected layer except the output layer. For age and gender classification tasks, the output layer has a softmax activation function followed by the combination of cross-entropy loss and center loss. However, for age regression, the mean absolute error (L1 loss) is used as the loss function for age regression, which measures the difference between exact age and predicted age by model. The combination of cross-entropy loss and center loss stay same for gender classification. In contrast to the VGG-16, ResNet architecture has no fully-connected layers except output layer. In ResNet-50, there are 50 layers, 49 of them are convolutional layers and one of them is a global average pooling layer. At the end of the network, there is also a FC layer as the output layer for producing the output of the network. The outputs and loss functions are built as in VGG-16 network. The input images are resized since both networks take input in 256×256 pixel resolution. Besides, apart from depth fusion, for all other fusion methods, the network takes three channels images as input. On the other hand, for depth fusion, since the input data has six channels, the filter parameters of the first convolutional layer of the both networks are arranged to take and process six channels images instead of a three channels one. In addition, as we discussed before, the networks are initialized with pretrained models that are trained on Multi-PIE Extended ear dataset [36]. The model weights can be available.[2]

We performed age classification as five class classification task as in the previous works [35, 36]. For age regression task, we used CNN models to predict exact age of the subjects, instead of classifying the age groups. We used mean absolute error (MAE) as the evaluation metric. In all the classification experiments, we utilized center loss [34] in addition to softmax loss. The main motivation behind center loss is to obtain more discriminative features. The center loss identifies the class center for each class and calculates the L2 distance between features and their corresponding class center. Then, the center loss is calculated for

[1] https://github.com/iremeyiokur/multipie_extended_ear_dataset

[2] https://github.com/yamand16/age_and_gender_classification

each feature and final loss is summed up with softmax loss to get the final loss. Besides, there is a λ coefficient in order to adjust the effect of the center loss over total loss. The overall loss formula is given in (1).

$$L = -\sum_{i=1}^{m} log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} ||x_i - c_{y_i}||_2^2 \tag{1}$$

According to the experimental results obtained on the validation set, the best performance is achieved with 0.1 λ coefficient. Since the proposed framework is trained in a multitask learning manner, there is a specific loss for each task, age and gender classification. The used loss function for both tasks are the same as mentioned before. However, in order to get the final loss, the age and gender losses are added and their contribution to the overall loss is weighted by the β coefficient. Since the gender classification task is easier than age classification, at the beginning of the training, the gender loss reduces very fast. Decreasing age loss takes more iterations than gender loss. Therefore, when we took average of both losses, it caused poor convergence. Because the gender loss, which has a low value, causes the total loss value to be low, which means that the total loss is more influenced by gender loss. In order to improve the convergence of the system, we determined the β coefficient to decrease the effect of gender loss and increase the effect of age loss over total loss. Experimental results obtained on the validation set indicate that using β coefficient to limit the effect of gender loss causes performance improvement of the system, especially for the age classification. The best performance is achieved with 0.75 β coefficient. The proposed final loss function for age and gender classification is shown in (2).

$$L_{total} = \beta * (L_{total_{age}}) + (1 - \beta) * (L_{total_{gender}}) \tag{2}$$

In addition to the age classification experiments, we configured our loss function based on mean absolute error (MAE) for age regression task. Besides, since the value ranges of MAE loss and cross-entropy loss are different, we limited MAE loss to be similar with cross-entropy loss in order to prevent performance degradation in gender classification during the training. For this, we empirically chose to shrink the mean absolute error multiplying by 0.1.

4 Experimental results

In this section, we first present used datasets and the experimental setups, then we provide the experimental results and discuss them.

4.1 Datasets

In this work, FERET [26], UND-F, and UND-J2 [37] datasets are utilized for the experiments. While FERET, UND-F, and UND-J2 datasets are employed for gender classification experiments, for age classification, solely FERET dataset is used, since UND-F and UND-J2 datasets do not have age labels. Sample images from the datasets are presented in Fig. 3.

FERET [26] contains 14,126 face images that belong to 1199 subjects. Both frontal and profile images of the subjects are available in the dataset. We obtained ear and face regions by running dlib face detector [19] and OpenCV [4] ear detector. This way, we acquired 1397 ear and face images of 596 subjects.

Fig. 3 Sample images from the employed datasets: **a** FERET, **b** UND-F, **c** UND-J2. While FERET dataset contains both profile face and frontal face, UND-F and UND-J2 datasets include only profile face images

ND-Collection-F (UND-F) & ND-Collection-J2 (UND-J2) [37] contain 942 and 2430 profile face images, respectively. In ear biometrics literature, UND-F and its extended version UND-J2 are benchmark datasets for gender classification. As in the FERET dataset, we applied dlib face detector [19] and OpenCV [4] ear detector to 2D images in order to get profile face and ear images.

In the experiments, we splitted the datasets into train, validation, and test sets using the 80%, 10%, and 10% of the samples in the datasets, respectively. We selected images randomly with respect to the data distribution per class. For age classification experiment, we have five age classes defined as follows: 18-28, 29-38, 39-48, 49-58, 59-68+. We determined these boundary values according to the previous age classification works [35, 36]. Number of samples per age class in FERET dataset is imbalanced. These age classes con-

Table 2 Age classification and regression results on FERET dataset

		Age Classification		Age Estimation	
		VGG-16	ResNet-50	VGG-16	ResNet-50
Fusion					
Data	Intensity	62.05%	58.53%	8.15	8.75
	Spatial	67.59%	62.71%	5.89	7.48
	Channel	61.83%	57.49%	8.22	9.35
Feature		67.28%	66.44%	5.33	5.91
Score	Basic	63.76%	63.06%	7.81	9.65
	d2s	63.06%	62.02%	–	–
	d2sr	63.06%	62.02%	–	–
	avg-diff	63.76%	63.06%	–	
	diff1	63.76%	63.06%	–	–

These results are obtained with two-stage fine-tuning strategy. For both tasks, data fusion, feature fusion, and *basic* version of score fusion are performed. Besides, additional score-based fusion methods are only used for age classification task. Since the classification scores were not available in the age regression task, the score-based fusion methods could not be performed except *basic* one. In the *basicmethod*, the average of the estimated exact ages of both models is calculated as a final prediction

tain 419, 435, 316, 169, and 58 profile face and ear images, respectively. As can be noticed, most of the subjects belong to the first two age groups. This imbalanced distribution affects the performance of age classification and age regression systems. While the classification accuracies are low in the last classes, it is more accurate in young ages. Similarly, the mean absolute error between ground truth and prediction gets higher for the age estimation task as the age increases.

4.2 Age estimation results

In Table 2, we presented the experimental results of different fusion types for age classification and regression tasks on FERET dataset [26]. In the table, the first column lists the used fusion types, and the other columns lists corresponding age classification and age regression results for both VGG-16 [32] and ResNet-50 [13] CNN models. According to the experimental results, the best age classification performance, when VGG-16 model is employed, is obtained with spatial fusion. Feature fusion also lead to a similar performance.

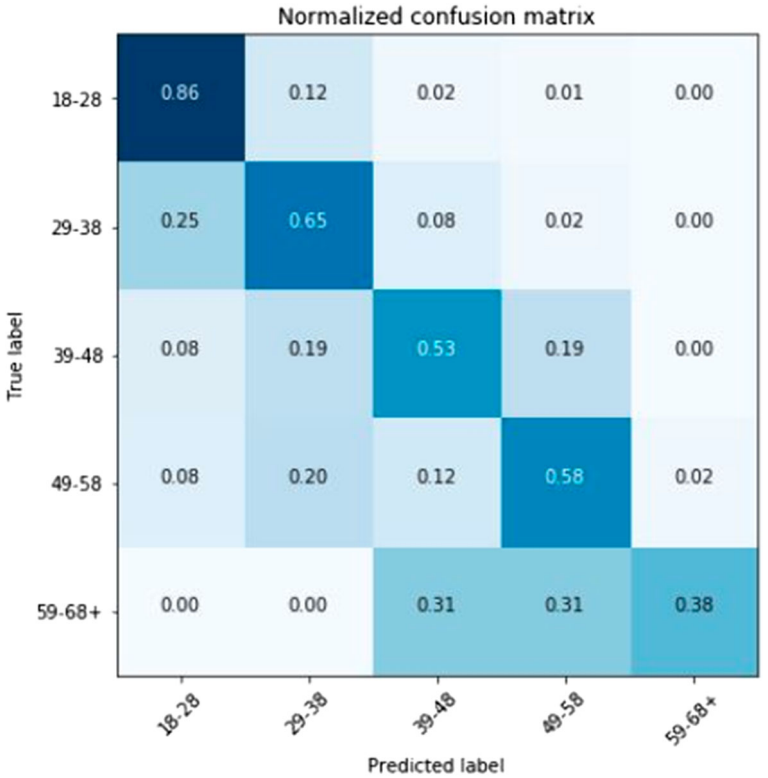

Fig. 4 Confusion matrix of age classification experiment on FERET dataset. As age goes up, the classification performance reduces. And the main reason for this outcome is found as unbalanced data distribution. Besides, as can be seen from the matrix, the misclassified samples are predicted as one of the neighbouring class

On the other hand, when ResNet-50 model is employed, feature fusion performs better than the other fusion methods and the obtained results using both models are close to each other. When the obtained results with score fusion are examined, they are found to be inferior to the ones achieved using feature fusion.

To analyze age classification results further, according to the outcomes of the best performing model, we generated the confusion matrix, as shown in Fig. 4. As can be seen from the confusion matrix, classification accuracy for the subjects, whose ages are between 18-28 is 86%, whereas, it is only 38% for the age group of 59-68+. One of the main reasons of this result is dataset imbalance, that is, the number of images per class is not the same and for the older age groups, there are less number of images. Besides, there are other factors related with images in the dataset such as lightning conditions, accessories, and etc. From the confusion matrix, it can also be observed that most of the errors occur at the class borders.

Therefore, we also calculated $\pm$ 1 class accuracy, which means if misclassified images are predicted as one of the neighbouring class, we accept it as correctly classified. The obtained $\pm$ 1 age classification accuracy is 89.33%, which validates that most of the misclassified images are assigned to the neighboring classes.

In addition to the age classification, we have also performed age regression to predict exact age of the subjects from their face and ear images The results are given in the third column of Table 2. Achieved lowest MAE is 5.33, which is obtained using VGG-16 model and feature fusion. This result indicates that profile face image and ear image contain sufficient information to perform age estimation.

In order to compare the performance of the multimodal approach with the ones obtained by the unimodal approaches, we have conducted age classification and age estimation experiments using ear images only and profile face images only. The results of these experiments are presented in Table 3. As one can see from the table, multimodal approach has outperformed both of the unimodal approaches. According to the results, profile face image is found to contain more useful information for age classification. Ear-only approach attains a lower performance compared to the one achieved by the profile face-only approach. However, the experimental results indicate that these two modalities complement each other, thus, when they are combined, performance has been improved.

Table 3 Age classification and regression results using different modalities on FERET dataset

Model	Modality	Accuracy	MAE
VGG-16	Ear	60.97%	9.91
VGG-16	Profile	65.73%	5.46
VGG-16	Ear + Profile	67.59%	5.33
ResNet-50	Ear	60.97%	10.02
ResNet-50	Profile	62.37%	8.99
ResNet-50	Ear + Profile	66.44%	5.44

These results are obtained with two-stage fine-tuning approach. Multimodal approach is outperformed unimodal methods both age classification and regression tasks

Table 4 Gender classification results on FERET dataset. These results are obtained with two-stage fine-tuning approach as in age tasks

		VGG-16	ResNet-50
Fusion			
Data	Intensity	93.03%	92.33%
	Spatial	99.11%	99.11%
	Channel	92.33%	91.63%
Feature		98.16%	97.56%
Score	Basic	97.90%	98.00%
	d2s	97.90%	98.00%
	d2sr	97.90%	98.00%
	avg-diff	97.90%	98.00%
	diff1	97.90%	98.00%

4.3 Gender classification results

The multimodal gender classification results are presented in Table 4. All these results are obtained on FERET dataset. The best gender classification result, which is 99.11%, is achieved with both VGG-16 and ResNet-50 models using spatial data fusion method. Compared to the age classification results, gender classification performance is found to be very high. This outcome could be due to several factors. First of all gender classification is just a two-class classification problem. Moreover, it does not suffer from the data imbalance problem, since the amount of images for males and females are close to each other. Lastly, also for humans, it is easier to estimate the gender from the profile views, that is, gender information is more evident in facial appearance, this way providing more useful features for gender classification.

Similar to age classification experiments, we also run unimodal systems for gender classification and presented the results in Table 5. Again multimodal approach has been found to be superior to both of the unimodal approaches. However, this time, ear image is found to contain more useful information for gender classification. Profile face-only approach obtains a lower performance compared to the one achieved by the ear-only approach.One of the main reason for this outcome could be earrings in the female subjects. In these datasets, generally, while male subjects do not have any earrings or similar accessories in the ear, female subjects have. Therefore, although the accessories are distracting factors for our models, it caused to build a relationship between accessories and gender because of the data distribution. Nevertheless, these experimental results also indicate that the two modalities complement each other.

Table 5 Gender classification results using different modalities on FERET dataset. These results are obtained with two-stage fine-tuning approach. As in age tasks, multimodal approach performs better than unimodal method

Model	Modality	Accuracy
VGG-16	Ear	97.56%
VGG-16	Profile	95.81%
VGG-16	Ear + Profile	99.11%
ResNet-50	Ear	98.00%
ResNet-50	Profile	94.05%
ResNet-50	Ear + Profile	99.11%

Table 6 Comparison of age classification and regression results with previous works

Method/Model	Dataset	Modality	Acc.	MAE
GoogLeNet [35]	Internal	Ear	52.00%	–
GoogLeNet [35] *	FERET	Ear	58.53%	–
ResNet-152 [5]	FERET	Profile	–	5.50
Ours	FERET	Multi	67.59%	5.33
Ours	FERET	Frontal	71.00%	4.89

Since the employed dataset in [35] is not FERET, we implemented it and run it on FERET dataset. This result is presented with * symbol

4.4 Comparison with state-of-the-art

We also compared the proposed approach with the previous works. We achieved state-of-the-art results for all tasks. For age classification experiment, the employed dataset in this work is not the same as the one used in the previous work [35]. For this reason, to have a fair comparison, we run the method presented in [35] on FERET. We showed our implementation of [35] with * in Table 6. According to presented results in Table 6, we achieved state-of-the-art results on age classification and age regression tasks. We outperformed the previous ear only age classification and profile face only age regression results with the presented multimodal approach. Moreover, we performed age classification and regression experiments using frontal face images in order to compare the usability of profile face and ear images as an alternative to the frontal face image.We performed frontal face experi-

Table 7 Comparison of gender classification results with previous works and frontal face

Method/Model	Dataset	Modality	Acc.
Distance + KNN [12]	Internal	Ear	90.42%
GoogLeNet [35]	Internal	Ear	94.00%
Gabor + Voting [18]	UND-J2	Ear	89.49%
HIS + SVM [21]	UND-J2	Ear	91.92%
BoF + SVM [39]	UND-F	Ear	91.78%
HIS + SVM [21]	UND-F	Ear	92.94%
BoF + SVM [39]	UND-F	Profile	95.43%
BoF + SVM [39]	UND-F	Multi	97.65%
Ours	FERET	Ear	98.00%
Ours	FERET	Profile	95.81%
Ours	FERET	Multi	99.11%
Ours	UND-F	Ear	98.33%
Ours	UND-F	Profile	97.50%
Ours	UND-F	Multi	100%
Ours	UND-J2	Ear	99.16%
Ours	UND-J2	Profile	98.33%
Ours	UND-J2	Multi	99.79%
Ours	FERET	Frontal	100%

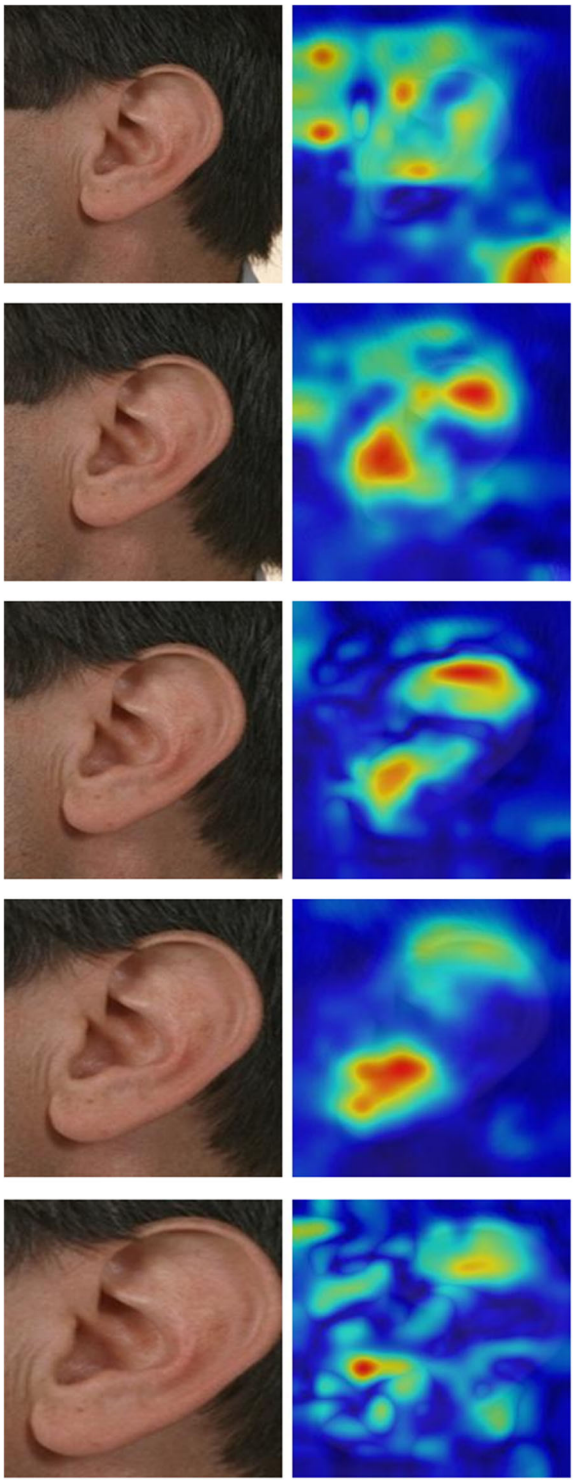

Fig. 5 Illustration of class activation map (CAM) which belongs to ear images with different crop margins. From the bottom up, the background part increased and the ear part decreased. In order to obtain images with different amount of context information, we added different amount of margin to the bounding box of ear detection outcome

ments with both VGG-16 and ResNet-50 models and since the best result is obtained with ResNet-50, we mentioned only that result on the table. According to the results, the age classification using frontal face images is 71% and our best result, 67.59%, shows that it is comparable with frontal face. This slight difference between proposed multimodal and frontal results indicates that the profile face and ear images can be employed as alternative modalities to the frontal face.

The comparison of gender classification results with previous works are presented in Table 7. We outperformed the previous ear only and profile face only approaches with our multimodal system and we achieved state-of-the-art results on FERET, UND-F & J2 datasets. Besides, we also outperformed them even with our unimodal approaches. We performed frontal face only experiments as well. We found that our multimodal approach performed almost the same with frontal face only approach and this outcome states that the combination of profile face and ear images can be utilized effectively to predict gender of the subject as an alternative to frontal face image when it is not available.

4.5 Class activation maps

We benefited from class activation maps to investigate the behaviour of the model. We generated class activation maps for the ear using similar strategy with [40] to see which part of ear is crucial for the CNN model in terms of feature extraction. Moreover, we cropped ear images with different margin to investigate the effect of context information over performance. We performed five different cropping strategy and presented their class activation maps and classification performances as well. For each different crop, we executed unimodal age classification experiment. The sample cropped images and corresponding class activation maps for age classification task are shown in Fig. 5. According to these results, it can be easily seen that the irrelevant parts such as background and hair, are distracting factors for the model. As context information gets bigger, the CNN model extracts features from outside of the ear which causes poor results.

The Table 8 lists age classification results from ear using images that contain different amount of context information. These results are obtained with VGG-16 model. The values in the first column are the number of pixels that are added to the four side of the bounding box coordinates to extend it while cropping the ear region. The achieved classification accuracies are in compliance with the class activation maps and as expected the best classification performance is obtained using a tighter crop (zero margin in Table 8).

Table 8 Age classification results on FERET dataset with different crop margin

Margin	Age Acc.
35	55.26%
25	57.32%
15	57.09%
5	56.76%
0	60.97%

The first column, margin, indicates the margin pixels that are added to extend the bounding box for ear detection generated by ear detector. The second column represents age classification results. This experiment is conducted using VGG-16 CNN model

5 Conclusion

In this paper, we presented various end-to-end multimodal and multitask deep learning frameworks for age and gender classification using profile face and ear images as input data. We performed domain adaptation using VGG-16 and ResNet-50 deep CNN architectures to adapt them to the ear domain. In order to make features more discriminative in the feature space, we associated center loss with softmax loss. We presented experimental results on FERET, UND-F, and UND-J2 datasets. We achieved state-of-the-art results on these datasets for all tasks with our proposed multimodal approach. We demonstrated that using profile face and ear images together, we can obtain high accuracies that are comparable with the ones obtained by using frontal face and outperforming the state-of-the-art approaches. We also analyzed class activation maps for ear experiments in order to explore the important part of the ear in terms of extracted features. Finally, we investigated the context information for ear and we found that less context data causes better performance. In summary, we have shown that profile face and ear images contain valuable information for age and gender classification, and they can be utilized effectively in a multimodal, multitask deep learning framework.

Acknowledgments This study is supported by the Istanbul Technical University Research Fund, ITU BAP, project no.42547 and Cost Action CA16101 - MULTI-modal Imaging of FOREnsic SciEnce Evidence - tools for Forensic Science (MULTI-FORESEE).

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. (2016) Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp 265–283
2. Abaza A, Ross A, Hebert C, Harrison MAF, Nixon MS (2013) A survey on ear biometrics. ACM Comput Surv 45(2):22
3. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2017) Effective training of convolutional neural networks for face-based gender and age prediction. Pattern Recogn 72:15–26
4. Bradski G, Kaehler A (2000) OpenCV. Dr. Dobb's journal of software tools, 3
5. Bukar AM, Ugail H (2017) Automatic age estimation from facial profile view. IET Comput Vis 11(8):650–655
6. Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp 77–91
7. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Computer vision and pattern recognition. IEEE, pp 248–255
8. Duan M, Li K, Yang C, Li K (2018) A hybrid deep learning cnn–elm for age and gender classification. Neurocomputing 275:448–461
9. Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Trans Inf Forensics Secur 9(12):2170–2179
10. Emeršič Ž, Štruc V, Peer P (2017) Ear recognition: More than a survey. Neurocomputing 255:26–39
11. Eyiokur FI, Yaman D, Ekenel HK (2017) Domain adaptation for ear recognition using deep convolutional neural networks. IET Biometrics 7(3):199–206
12. Gnanasivam P, Muttan S (2013) Gender classification using ear biometrics. In: International conference on signal and image processing. Springer, pp 137–148
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Computer vision and pattern recognition. IEEE, pp 770–778
14. Iannarelli A (1989) Ear identification, forensic identification series. Paramont Publ Company

15. Introduction to USTB ear image databases: http://www1.ustb.edu.cn/resb/en/index.htm. Access date 30 Oct 2018
16. Jain AK, Dass SC, Nandakumar K (2004) Soft biometric traits for personal recognition systems. In: Biometric authentication. Springer, pp 731–738
17. Jain AK, Park U (2009) Facial marks: Soft biometric for face recognition. In: International conference on image processing. IEEE, pp 37–40
18. Khorsandi R, Abdel-Mottaleb M (2013) Gender classification using 2-D ear images and sparse representation. In: Workshop on applications of computer vision. IEEE, pp 461–466
19. King DE (2009) Dlib-ml: A machine learning toolkit. J Mach Learn Res 10(Jul):1755–1758
20. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
21. Lei J, Zhou J, Abdel-Mottaleb M (2013) Gender classification using automatically detected and aligned 3D ear range data. In: International conference on biometrics. IEEE, pp 1–7
22. Levi G, Hassner T (2015) Age and gender classification using convolutional neural networks. In: Computer vision and pattern recognition workshops, pp 34–42
23. Meng D, Mahmoodi S, Nixon MS (2020) Which ear regions contribute to identification and to gender classification? In: 2020 8th international workshop on biometrics and forensics (IWBF). IEEE, pp 1–6
24. Ozbulak G, Aytar Y, Ekenel HK (2016) How transferable are CNN-based features for age and gender classification? In: International conference of the biometrics special interest group. IEEE, pp 1–6
25. Pflug A, Busch C (2012) Ear biometrics: A survey of detection, feature extraction and recognition methods. IET Biometrics 1(2):114–129
26. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104
27. Purkait R, Singh P (2007) Anthropometry of the normal human auricle: A study of adult Indian men. Aesthetic Plast Surg 31(4):372–379
28. Rothe R, Timofte R, Van Gool L (2018) Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis 126(2-4):144–157
29. Saeed U, Khan MM (2018) Combining ear-based traditional and soft biometrics for unconstrained ear recognition. J Electronic Imag 27(5):051220
30. Sforza C, Grandi G, Binelli M, Tommasi DG, Rosati R, Ferrario VF (2009) Age-and sex-related changes in the normal human ear. Forensic Sci Int 187(1-3):110–e1
31. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: An astounding baseline for recognition. In: Computer vision and pattern recognition workshops, pp 806–813
32. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
33. Vaquero DA, Feris RS, Tran D, Brown L, Hampapur A, Turk M (2009) Attribute-based people search in surveillance environments. In: Workshop on applications of computer vision. IEEE, pp 1–8
34. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, pp 499–515
35. Yaman D, Eyiokur FI, Sezgin N, Ekenel HK (2018) Age and gender classification from ear images. In: International workshop on biometrics and forensics. IEEE
36. Yaman D, Irem Eyiokur F, Kemal Ekenel H (2019) Multimodal age and gender classification using ear and profile face images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 0–0
37. Yan P, Bowyer KW (2005) Empirical evaluation of advanced ear biometrics. In: Computer vision and pattern recognition workshops. IEEE, p 41
38. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Advances in neural information processing systems, pp 3320–3328
39. Zhang G, Wang Y (2011) Hierarchical and discriminative bag of features for face profile and ear based gender classification. In: International joint conference on biometrics. IEEE, pp 1–8
40. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929