Neurocomputing 367 (2019) 346–356

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Age estimation via attribute-region association

Yiliang Chen[a], Shengfeng He[a,*], Zichang Tan[b], Chu Han[c], Guoqiang Han[a], Jing Qin[d]

[a] School of Computer Science and Engineering, South China University of Technology, China
[b] Institute of Automation, Chinese Academy of Sciences, China
[c] Department of Computer Science and Engineering, the Chinese University of Hong Kong, China
[d] Department of Nursing, the Hong Kong Polytechnic University, China

ARTICLE INFO

Article history:
Received 18 December 2018
Revised 28 April 2019
Accepted 14 August 2019
Available online 27 August 2019

Communicated by Prof. Liu Guangcan

Keywords:
Age estimation
Multi-task learning
Attribute-region association

ABSTRACT

Human age has been treated as an important biometric trait in many practical applications. In this paper, we propose an Attribute-Region Association Network (ARAN) to tackle the challenging age estimation problem. Instead of performing prediction from a global perspective, we delve into the relationship between face attributes and regions. First, the proposed network is guided by the auxiliary demographic information, as different demographic information (*e.g.*, gender and ethnicity) intrinsically correlates to human age. Second, different face components are separately handled and then involved in the proposed ensemble network, as these components vary differently along with human age. To explore both global and local information, the proposed network consists of several sub-network, each of them takes the global face and a face sub-region as input. Each sub-network leverages the intrinsic correlation between different face attributes (*i.e.*, age, gender, and ethnicity), and it is trained in a multi-task manner. These attribute-region sub-networks are associated to yield the final predictions. To properly train and coordinate such a complex network, a new hierarchical-scheduling training method is proposed to balance the learning complexity in the multi-task learning. In this way, the performance of the most difficult task (*i.e.*, age estimation) can be significantly improved. Extensive experiments on the MORPH Album II and FG-NET show that the proposed method outperforms the state-of-the-art age estimation methods by a significant margin. In particular, for the challenging age estimation, the Mean Absolute Errors (MAE) are decreased to 2.51 years compared to the state-of-the-arts on the MORPH Album II dataset.

1. Introduction

Human facial attributes are used as the biometric traits in many applications. For example, age, gender, and ethnicity information can be used for precise advertising, human computer interaction (HCI), and security control. However, predicting facial attributes suffers from different challenging factors like wrinkles, lighting or occlusion, which makes it a difficult task even for a human.

Early research [20] on age estimation of face image uses handcrafted facial geometric features and facial wrinkles to classify children, teenage and old people. Due to the publicly available large scale age datasets like FG-NET [21] and MORPH Album II [26], handcrafted features are replaced by learned feature representations.

Recently, regression based methods and ranking based methods become popular and shown to be useful for improving the performance of age estimation. Regression based methods [8,24,27] usually adopt a loss function like L2 loss to penalize the differences between the predicted ages and the ground-truth. This kind of method pays more attention to investigate attribution relationship among face images and age prediction. Differently, ranking based methods [2,3,6] regard the age value as a rank ordered data, and utilize multiple binary classifiers to determine the rank of the age in a face image. These methods focus on the ordinal relationship among face images with age values.

The majority of the age estimation methods [7,30] focus on the age attribute of an input face image. However, other face attributes reveal human age from various perspectives. For instance, gender and ethnicity attributes are closely related to ages, and our human predict ages takes into account these two source of information.

On the other hand, the correlation between the holistic face image and different face components is also useful for face attributes estimation. For example, face aging process can be observed from the shape of face in childhood, while the aging process is more conspicuous in the skin of texture in adulthood. Therefore, face components reveal age information, and they should not be

* Corresponding author.
E-mail address: hesfe@scut.edu.cn (S. He).

https://doi.org/10.1016/j.neucom.2019.08.034
0925-2312/

considered globally under the same criterion. Therefore, attribute-correlation based methods [16,18,31] are widely used to estimate the exact age of the subject in a face image.

In this paper, we aim to integrate different face attributes and components in an end-to-end unified network for efficient and accurate prediction. To this end, we propose an Attribute-Region Association Network (ARAN) to learn the correlation between these two additional information. First, our proposed network is guided by the auxiliary demographic information such as gender and ethnicity. Second, different face important components like eyes, noses and mouths are utilized to explore the correlation between the holistic face region and the local components. To properly train such a complex network, a new hierarchical-scheduling training (HST) method is proposed to balance the learning complexity in the multi-task learning. The contributions of the proposed work are summarized below:

- A novel Attribute-Region Association Network (ARAN) and a hierarchical scheduling-training (HST) methods are proposed for age estimation, which simultaneously utilizes different face attributes and components in unified network.
- Extensive experimental results demonstrate that the proposed framework is significantly better than the state-of-the-art methods for age estimation, gender classification, and ethnicity recognition on the MORPH Album II and FG-NET datasets.

2. Related Work

Age estimation methods. Human face attributes estimation has been explored for over 20 years. In the early research of age prediction, geometry features [20,25], *e.g.*, chin skin wrinkles, nose or eye, are usually used to predict the range of human age (*i.e.*, child, young or senior adult). Then, some methods, such as AGing PattErn Subspace (AGES) [11], Biologically Inspired Features (BIF) [17], BIF+SVM [17] and BIF+CCA [15], are proposed for precise age estimation. These approaches adopt handcrafted features (*i.e.* BIF), and however they are difficult to obtain a rich representation of human face.

To address this problem, several deep learning based methods are proposed for age estimation [23,27,28,34]. The most representative method is Deep Expectation (DEX) [27] model, which is based on the VGG16 structure and adopts an Expected Value (EV) to calculate the final prediction. This method ranks the 1st place at the CheaLeARN LAP challenge 2015. Rothe et al. then improve their simple and elegant DEX method without using facial landmarks and decreased the MAE to 4.785 years on the CACD [4] dataset. The champion [1] of the ChaLearn Lap challenge 2016 further improves DEX method, by adding a separate model for the images of children. Shen et al. [29] propose an end-to-end CNNs method named Deep Regression Forests (DRFs) for age estimation, which learns nonlinear regression between heterogeneous facial feature space and ages.

Ordinal information is used by Niu et al. [23] to reduce the MAE to 3.63 years on the MORPH Album II, by adopting an end-to-end deep model to address the ordinary regression problem. Similarly, Chen et al. [6] build multiple binary CNNs to learn the ordinal information of ages, and the final result is the aggregation of these binary CNN outputs.

Multi-region methods. Gidaris *et al.* [12] propose a multi-region deep convolutional neural network for object detection, and it achieves a very impressive detection result. This similar idea is also adopted for age estimation using pre-partitioned facial regions. Yi et al. [34] train 46 parallel CNNs with different facial sub-regions, and such method successfully decreased the MAE to 3.63 years on the MORPH Album II. After that, Ting et al. [22] try to simplify and improve [34], but the improvement is not significant as it ignores the differences among different sub-regions.

Multi-task learning. Existing approaches [18,19] utilize attribute correlation to enhance the performance of age estimation. Wan et al. [19] propose a divide-and-conquer method, which takes advantage of gender and race attributes to improve the performance of age estimation. Han et al. [18] design a Deep Multi-Task Learning method for both attribute correlation and attribute heterogeneity in a single neural network. Tian et al. [31] use orthogonalizing to calculate the attribute correlation between human gender and age.

In this paper, we combine the advantages of multi-region and multi-task learning methods as a new framework. However, these two type of methods cannot be heuristically combined due to their complex network structures. Therefore, we propose a specifically designed network and a training scheduling method to overcome this barrier.

3. Attribute-Region association network

In this section, we introduce the proposed attribute-region association framework. Our pipeline is shown in Fig. 1. We describe each stage of our framework in detail below. Besides, we introduce a hierarchical scheduling training technique for multi-task learning.

3.1. Face alignment

Face alignment is of great importance to age estimation, as it can eliminate irrelevant factors from facial images and reduce the ambiguities from processing. Therefore, at the first stage in Fig. 1 we use active shape model [9] to locate the important facial points from the images. Then we crop and rotate the input image according to the center location of two eyes and the middle point of the upper lip. The regions of the eyes, nose and mouth are then cropped and aligned according to the facial key points. All the cropped and aligned images are resized to 224×224. Similar to [23], we use color face image as input. In addition, we follow the setting of [8,27,30,34] that discard the face images which cannot be detected by a face detector. Some examples of the aligned and cropped images are presented in Fig. 2.

3.2. Architecture

To leverage the information from attribute-region association learning, the proposed method consists of three sub-networks (see Fig 1). Each of them corresponds to the region of eye, nose, and mouth. Each sub-network takes a pair of images as input, *i.e.*, a global face image and a sub-region image. Three softmax classifiers are connected to each sub-network for age estimation, gender classification and race recognition respectively. The outputs of each task are then combined in an ensemble layer, and this layer yields a final output for a specific task.

The proposed ARAN sub-networks are constructed based on two classical architecture: AlexNet and VGG-16 networks. The example sub-network structure is shown in Fig. 3. Besides, we also adopt a single AlexNet or VGG-16 baseline model in our experiments for better comparison. The detailed architectures of each sub-network are discussed below.

3.2.1. Baseline network architecture

To deploy our multi-task learning strategy, we slightly modify the original AlexNet structure. Our baseline network includes 5 convolutional layers and 3 max pooling layers, where the first four layers are shared across different tasks. After that, the output from Conv4 layer is fed to three branches for three different tasks. Each branch contains an independent convolutional layer, a max pooling

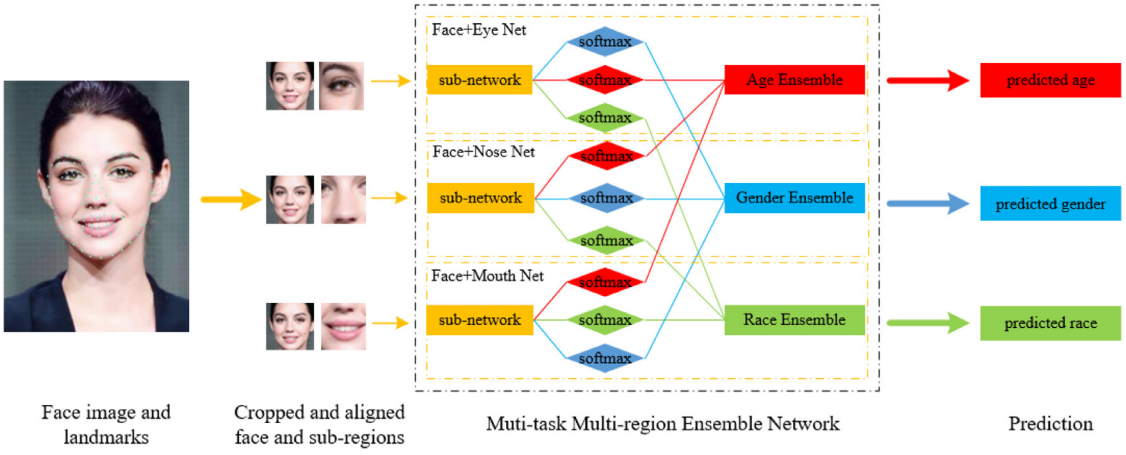


Fig. 1. Our attribute-region association framework. (Red: age estimation, blue: gender classification, green: race recognition.)

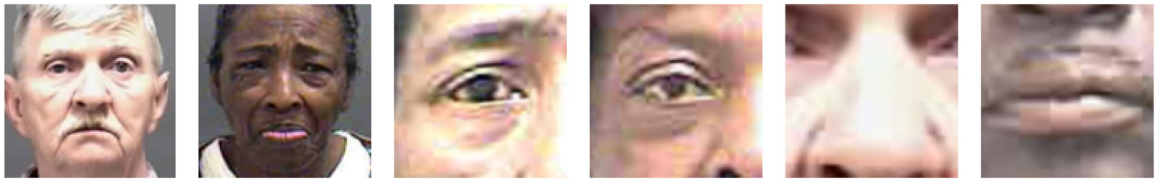


Fig. 2. Results of face alignment.

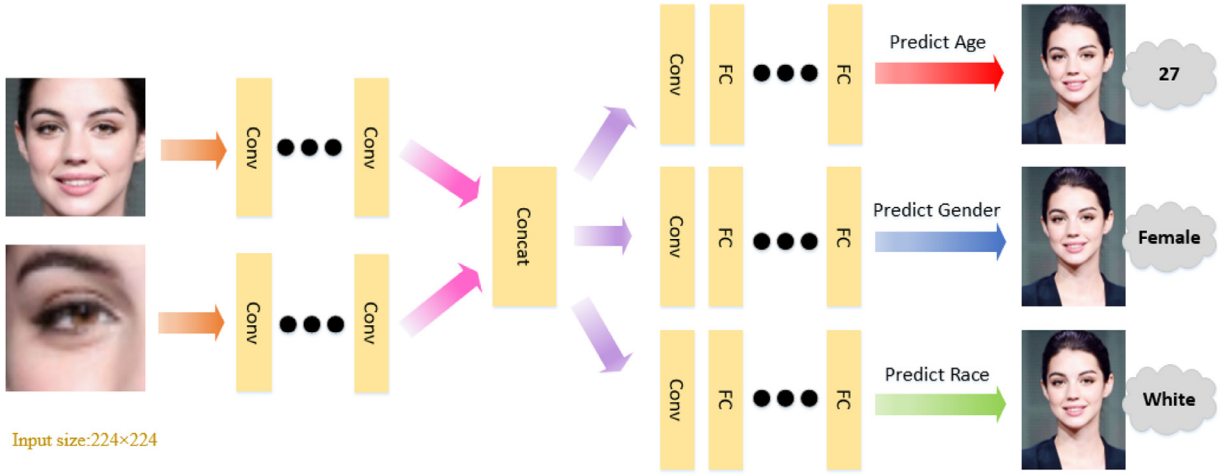


Fig. 3. The detailed architecture of a face+eye sub-network in ARAN.

layer and two fully connected layers. The last fully connected layer of each branch is followed by a loss function for a specific task. The original Alexnet includes three fully connected layers, but we find that if our network contains only two fully connected layers, it can reach a better performance for age estimation. Besides, in the following experiments, our single-task baseline network is identical to our multi-task baseline architecture, but only one branch is activated.

3.2.2. ARAN Network architecture

Our ARAN network is made up of three sub-networks. The input of each sub-network is a pair of images with a holistic face and a face component. These two images are fed to a sub-network. Each sub-network is a variant of either Alexnet or VGG16. The difference between the multi-task baseline architecture and our ARAN sub-network Fig. 3 is that our sub-network is a two-stream network, and each input image is processed separately with two different set of network parameters. The outputs from two streams are concatenated, and then they fed to three branches. Our ARAN framework with AlexNet architecture is trained from scratch, and no pre-trained model is used in our experiments. Similarly, for our ARAN (VGG16) architecture which contains a very deep structure with smaller kernel size (3x3), it is pre-trained on the ImageNet [10] dataset, and fine-tuned on the MORPH Album II dataset in our following experiments.

3.3. Ensemble inference

In Fig. 1, the red block of softmax function is for age estimation and its cross entropy loss function is defined as:

$$l_{\text{Age}}(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{u} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{u\_age} e^{\theta_l^T x^{(i)}}}\right], \tag{1}$$

where θ is the parameter matrix of the softmax function; $1\{\bullet\}$ is indicator function, which means $1\{\text{a true statement}\} = 1$; $\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{u} e^{\theta_l^T x^{(i)}}}$ is the predicted probability distribution. Besides, every input sample is represented as $(x^{(i)}, (y^{(i)}, g^{(i)}, r^{(i)}))$ and $x^{(i)}$ indicates the features of the i-th sample, and $x^{(i)} \in R^n$. $y^{(i)}$ is the age label of the ith sample, $g^{(i)}$ is the gender label of the ith sample and $r^{(i)}$ is the race label of the i-th sample. *u_age* is the class numbers for our age prediction. Similarly, the blue and green blocks of the softmax functions are used for gender classification and race classification respectively, and their corresponding loss functions are Eqs. (2) and (3).

$$l_{\text{Gender}}(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{u} 1\left\{g^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{u\_gender} e^{\theta_l^T x^{(i)}}}\right]. \tag{2}$$

$$l_{\text{Race}}(\theta) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{u} 1\left\{r^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{u\_race} e^{\theta_l^T x^{(i)}}}\right]. \tag{3}$$

Therefore, our objective function of each sub-network is defined as:

$$l_{all}(\theta) = \alpha l_{Age}(\theta) + \beta l_{Gender}(\theta) + \gamma l_{Race}(\theta), \tag{4}$$

where α, β and γ are the loss weights of their own loss functions (age estimation, gender classification, and ethnicity recognition respectively). In our experiments, α, β and γ are set to 1, 0.1 and 0.1, respectively. We also have a brief discussion about different combinations of the loss weights in our experiment section.

For the age prediction sub-network, the metric Expected Value [27] is widely adopted to calculate the predicted age. The predicted age of P_k in Eq. (5) from each softmax function of the sub-network is $\sum_{i=0}^{100} p_i i$, where p_i is the predicting probability of the corresponding age i. Its subindex i ranges from 0 to 100, as our softmax function is a hundred-and-one-dimensional vector [22,28,30,34]. Finally, three predicted ages are combined in the age ensemble function:

$$P_{\text{age}} = \sum_{k=1}^{m} W_k P_{k,age}, \tag{5}$$

where W_k is the weight for each sub-network k and m is defined as the number of the sub-regions. Afterwards we combine the predictions of three sub-networks and get a final prediction P_{age}.

For gender and ethnicity predictions, both of them are binary classifications in our experiments. Unlike age predictions, an integer value is required for evaluating these two tasks, therefore in here we use a rounding approach. Their own softmax outputs are two numbers which can indicate the accuracies of our prediction. Therefore, the class of higher accuracy is our predicted gender or race. Subsequently, both tasks go through their individual ensemble layers, and their predictions are combined respectively according to the following function:

$$P_{gender,race} = round\left(\frac{1}{m}\sum_{k=1}^{m} P_{k,gender,race}\right). \tag{6}$$

3.4. Hierarchical scheduling

Apparently, age estimation task is far more difficult than gender classification and ethnicity recognition. As shown in Fig. 4, the MAE of age estimation (red curve) requires a large number of iterations to converge, while the gender classification and the ethnicity recognition (purple and green curves) take a few iterations to get the highest accuracies. If the other two tasks have converged and the age estimation task still requires further training, the overfitted gender and ethnicity tasks may prevent age estimation from training towards optimum performance. Therefore, we introduce a hierarchical scheduling training method to ease the burden of unbalanced multi-task learning.

Our method is divided into two stages. Firstly, we pre-trained our network for the age estimation task independently. In other words, the branches of the gender classification and ethnicity recognition are not updated during the pre-training period. Particularly, the shared layers across tasks do not update in pre-training. During the second stage, we only keep the parameters from the shared layers of the pre-trained network, and the rest of layers are initialized by random gaussian distribution. As can be seen in Fig. 3, we keep the pre-trained parameters of shared convolutional layers for initialization in the second stage. After that, we finetune our pre-trained network for all the three tasks.

In this way, the difficult age estimation task is individually pre-trained first, which mitigates the learning ambiguities of learning three tasks simultaneously. Furthermore, these three tasks share similar mid-level and high-level knowledge of human face, and the pre-trained shared layers learn rich representations of different face attributes, and thus boost the performance of all the three tasks.

Table 1
The number of images in the three splits of the MORPH Album II dataset.

	Male			Female		
Black	S1:4012	S2:4012	S3:28835	S1:1305	S2:1305	S3:3166
White	S1:4012	S2:4012	S3:0	S1:1305	S2:1305	S3:0
Other	S1:0	S2:0	S3:1845	S1:0	S2:0	S3:130

4. Experiments

4.1. Experimental setup

In our experiments, the proposed framework is evaluated mainly on the MORPH Album II [26], FG-NET [21] and CACD [4] datasets. Both datasets are public popular datasets for human face age estimation. We use a learning rate of 0.0001, a weight decay of 0.0005 and a momentum of 0.9.

MORPH Album II contains approximately 55,000 face images and their ages range from 16 to 77 years. CACD is the biggest public cross-age dataset and it is collected from the Internet Movie DataBase (IMDB). Besides, FG-NET contains 1002 facial images of 82 individuals and CACD includes more than 160 thousands images of 2000 celebrities.

MORPH Album II is the only dataset that contains age, gender and race attributes. According to the age, gender and race distributions of MORPH Album II in [32], the dataset contains approximately 77% black and 19% white, while the corresponding percentage for the gender is 15% female and 85% male. Since MORPH Album II is highly unbalanced in terms of race and gender distribution, we follow [13,29,34] to use the same test protocols[1] provided by Yi et al. [34]. The dataset is randomly partitioned into three non-overlapped subsets S1, S2 and S3. Therefore, there are two different combinations of training set and testing set: 1) Training set is S1, and testing sets are S2+S3; 2) Training set is S2, and testing sets are S1+S3. For multi-task learning, we discard the images with non-black and non-white races ("other" in Table 1), because our ethnicity recognition is only for white and black faces.

[1] http://www.cbsr.ia.ac.cn/users/dyi/agr.html.

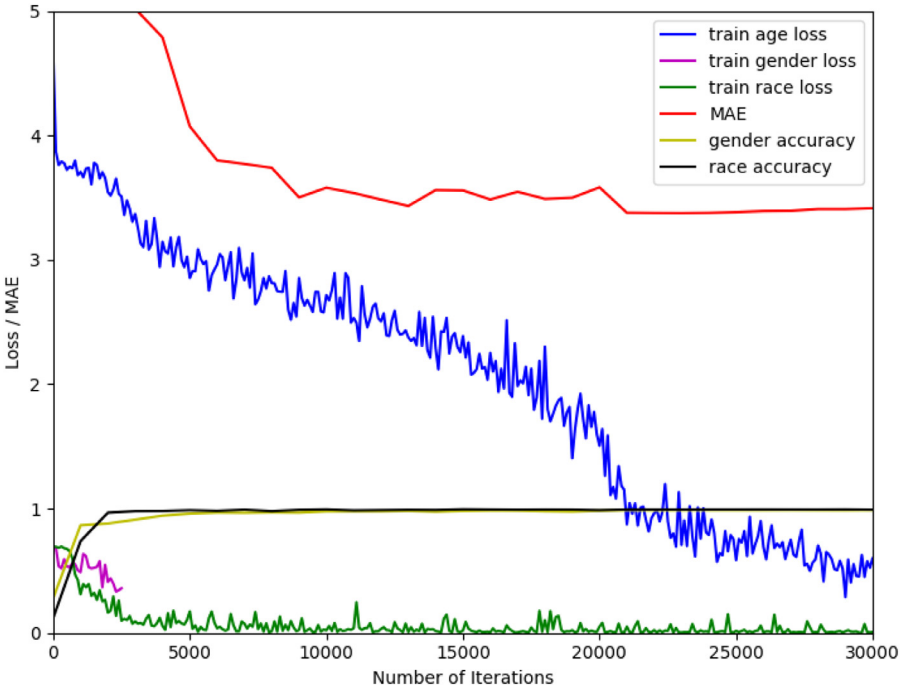


Fig. 4. The training processing without HST technique on the Morph Album II dataset (Training Set: S1, Testing Set: S2+S3).

As a result, the filtered test set contains 42,635 samples, and the number of samples in the train set remains the same.

For FG-NET datast, we manually annotate gender and ethnic labels on this dataset. In our experiments, we follow the testing setting used in [3,5,33], and we perform "leave one out" cross validation on FG-NET dataset. In other word, we leave images of one person for testing and take the remaining images for training.

As the CACD dataset images are collected from IMDB, which introduces noise labels to the dataset. We follow the setting of [30] to use 200 celebrities that with less noise for testing, and the other for training in our experiments. Due to the lack of gender and race attributes, CACD dataset is only used for our single-task evaluations.

As discussed in Section 3.1, we crop the regions from the human face and all the non-face images are removed from the dataset. After this operation, MORPH Album II includes 55,244 images, and CACD contains 162,941 images [30].

To avoid overfitting on the MORPH Album II dataset, we augment the training images by flipping, rotating with $\pm$ 8º and $\pm$ 5º, and adding Gaussian white noises with variance of 0.001, 0.005, 0.01, 0.015 and 0.02.

4.2. Evaluation metrics

In our experiments, we adopt Mean Absolute Error (MAE) and cumulative score (CS) as the evaluation criteria of age estimation [17]. For gender classification and ethnicity recognition, we only compute their accuracies. MAE is defined as:

$$MAE=\frac{1}{N}\sum_{t=1}^{N}|\widehat{y}_i - y_i|, \tag{7}$$

where N denotes the number of testing samples. $\widehat{y}_i$ and y_i are the ground truth age and predicted age of the k-th image respectively. The lower the MAE, the better the estimation performance. CS is defined as:

$$CS(j) = \frac{N_{e\leq j}}{N} \times 100\%, \tag{8}$$

where $N_{e\leq j}$ denotes the number of testing images whose absolute error between the ground truth age and the predicted age is more than j years. On the contrary to MAE, a higher CS value indicates the better result.

4.3. Multi-region evaluation

We take age estimation task as an example, by setting the learning rate of the other two branches to zero, to show the performance of our ARAN on single task learning.

4.3.1. Relation between regions and age groups

We investigate the performance of age estimation among different single-region networks, which are based on our single-task baseline architecture on the Morph Album II dataset (training set: S1, testing set: S2). The results are shown in Fig. 5a. We can see that the global face outperforms the other regions significantly, and other regions get bad performances on their own. Besides, we also show the age estimation performances of different multi-region networks on the Morph Album II (training set: S1, testing set: S2) and CACD datasets. We can see in Fig. 5b and c that the performance of a single-face net is close to a Face+LeftEye net, a Face+Nose net or a Face+Mouth net on both datasets. Surprisingly, for these four networks, each of them can outperform others at some ages, and these ages are different between two datasets. It is difficult to say which region is benefit to some ages all the time for different datasets. Hence, we adopt a simple method to ensemble the networks by calculating their average value. We can see that for the Morph Album II dataset, there are only about 100 samples existing at the range from 55 to 60, and therefore we can consider single-task ARAN (AlexNet) almost outperforms other networks at all ages on both datasets.

4.3.2. Evaluation on the importance of regions

Here we show the importance of different regions. We can see from Table. 2, it shows that our results are relatively stable w.r.t different weights in a reasonable range (around 1/3), which indicates that their combination strategies are not sensitive to the

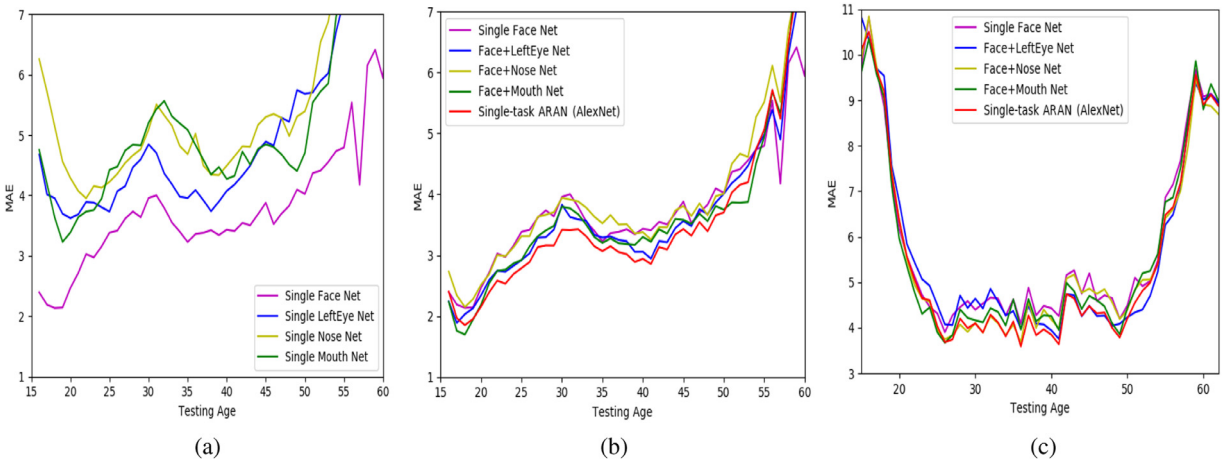


Fig. 5. Age estimation evaluations with different settings: (a) Single region nets at different testing ages on the Morph Album II dataset. (b) Multi-region nets at different testing ages on the Morph Album II dataset. (c) Multi-region nets at different testing ages on the CACD dataset.

Table 2
Different ARAN (AlexNet) ensemble methods based on the CACD and Morph Album II (Training set: S1) datasets.

Face+LeftEye	Face+ Nose	Face+ Mouth	CACD MAE ↓	MOR. Alb. II MAE↓
1	0	0	5.12	3.25
0	1	0	5.11	3.25
0	0	1	5.08	3.48
1/3	1/3	1/3	4.85	3.04
1/2	1/3	1/6	4.87	3.04
1/2	1/6	1/3	4.87	3.05
1/3	1/2	1/6	4.86	3.05
1/3	1/6	1/2	4.86	3.08
1/6	1/2	1/3	4.86	3.07
1/6	1/3	1/2	4.86	3.10

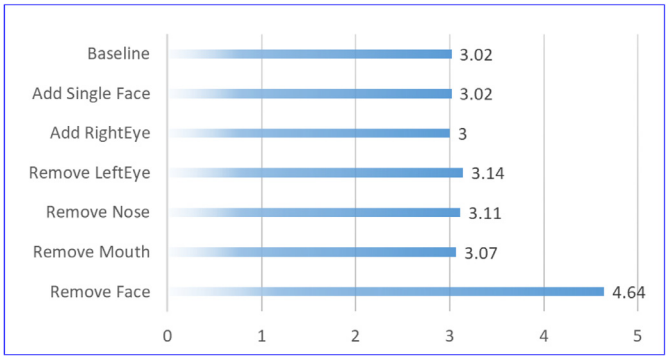


Fig. 6. Evaluation on the number of regions.

combination parameters. The best parameters set is 1/3, 1/3 and 1/3 (averaging). Especially on CACD, the differences between different combinations are no more than 0.02. Though the results of the Morph Album II dataset is not as stable as CACD, but the difference is very small and the best performance is still around 1/3. Therefore, in our experiment, the averaging ensemble is very suitable for the ensemble weighting coefficients W_k of our ARAN framework.

4.3.3. Evaluation on the number of regions

We also investigate the influence of increasing or decreasing sub-regions with our single-task baseline architecture. Our baseline structure only adopt left eye, nose and mouth as sub-regions in our experiments. Fig. 6 illustrates that our framework does not get a significant improvement if we add the sub-regions like the global face or the right eye. The reason may be these regions are very similar to our adopted regions. However, if we decrease one of the regions we adopted, the performance drops dramatically to no less than 3.07 years. Therefore, the eyes, the noses and the mouths are very important for our framework. A better performance may be achieved if we can get more unique face information, like ears and hairs (they cannot be obtained using face landmark). Besides, we can see that the global face is the most important feature, and it cannot be removed from the pipeline.

4.4. Evaluation on the attribute-Region association network

We evaluate the proposed attribute-region association framework only on the Morph Album II dataset, as the other datasets do not contain age, gender, race labels at the same time. Our attribute-region association sub-network is built on Fig. 3 for AlexNet version and VGG16 version respectively. ARAN (AlexNet) is directly trained form the training set of the Morph Album II dataset, while the VGG16 version is pre-trained with ImageNet [10]. In Table 3, we compare our ARAN frameworks and some baseline architectures. Comparing with the results between ARAN (AlexNet) and single-task methods (based on modified AlexNet), we can find that ARAN (AlexNet) method can hugely improve the performance of age estimation and slightly increase the accuracy of ethnicity recognition. The MAE value is improved from 3.20 to 2.96. Moreover, it is noticeable that our multi-task baseline network has a slight decrease (0.02) compared with the single-task baseline network, and therefore our ARAN (AlexNet) provides more help for the multi-task learning networks. Besides, the performance of gender classification and ethnicity recognition also have a slight increase from our multi-task baseline from ARAN (AlexNet). Though there is still a slight difference on gender classification between our ARAN (AlexNet) and the single task, our ARAN (AlexNet) framework can significantly enhance the performance of the most difficult task (age estimation). For our VGG-16 version, it outperforms all the methods on three tasks, and further improves the performance of age estimation to 2.63.

4.5. Comparison with the state-of-the-arts

In order to show the effectiveness of our method, we compare our methods with other state-of-the-art algorithms on the Morph Album II and FG-NET dataset.

For Morph Album II, the results are summarized in Table 4 and Fig. 7. Similar to previous works [8,19], we adopt the mirroring

Table 3
Comparisons with the baseline methods on the Morph Album II dataset. Single task methods are based on our sinlge-task baseline architecture.

Architecture	Train Set	Test Set	Gender Acc.	Race Acc.	Age MAE ↓
Age (Single task)	S1	S1+S3	–	–	3.34
	S2	S1+S3	–	–	3.05
	Average		–	–	3.20
Gender (Single Task)	S1	S2+S3	98.94%	–	–
	S2	S1+S3	98.99%	–	–
	Average		*98.97%*	–	–
Race (Single Task)	S1	S2+S3	–	99.02%	–
	S2	S1+S3	–	99.13%	–
	Average		–	99.08%	–
Baseline (Multi-task)	S1	S2+S3	98.58%	99.33%	3.40
	S2	S1+S3	98.83%	99.26%	3.03
	Average		98.71%	99.30%	3.22
Face+LeftEye (AlexNet)	S1	S2+S3	98.59%	99.17%	3.39
	S2	S1+S3	98.72%	99.30%	3.04
	Average		98.66%	99.24%	3.22
Face+Nose (AlexNet)	S1	S2+S3	98.59%	99.17%	3.46
	S2	S1+S3	98.72%	99.30%	3.03
	Average		98.66%	99.24%	3.25
Face+Mouth (AlexNet)	S1	S2+S3	98.58%	99.12%	3.72
	S2	S1+S3	98.75%	99.26%	3.29
	Average		98.67%	99.19%	3.51
Face+LeftEye+Nose+Mouth	S1	S2+S3	–	–	3.43
	S2	S1+S3	–	–	3.13
	Average		–	–	3.28
ARAN (AlexNet)	S1	S2+S3	98.81%	99.27%	3.15
	S2	S1+S3	98.92%	99.37%	2.76
	Average		98.87%	**99.32%**	*2.96*
Face+LeftEye (VGG16)	S1	S2+S3	98.20%	99.02%	2.96
	S2	S1+S3	98.62%	98.97%	2.70
	Average		98.41%	99.00%	2.83
Face+Nose (VGG16)	S1	S2+S3	98.63%	98.60%	2.98
	S2	S1+S3	98.64%	99.22%	2.67
	Average		98.64%	98.91%	2.83
Face+Mouth (VGG16)	S1	S2+S3	98.15%	99.05%	2.96
	S2	S1+S3	98.67%	98.95%	2.75
	Average		98.41%	99.00%	2.86
ARAN (VGG16)	S1	S2+S3	99.03%	99.14%	2.77
	S2	S1+S3	98.94%	99.22%	2.48
	Average		**98.99%**	*99.18%*	**2.63**

* Bold and italic fonts indicate the #1 and #2 performances.

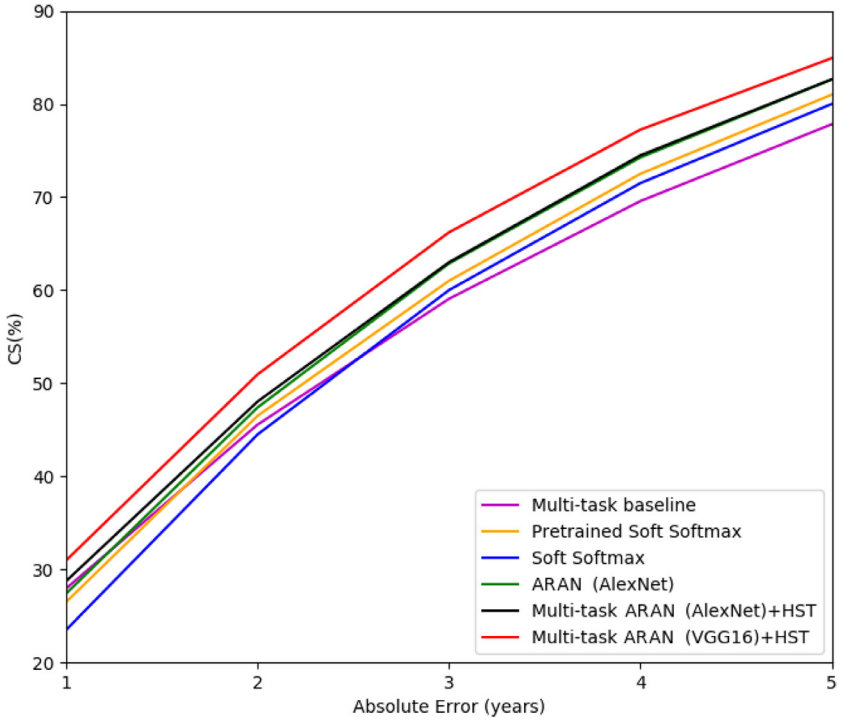


Fig. 7. CS curves based on the Morph II dataset.

Table 4
Comparisons with the state-of-the-art methods on the Morph Album II dataset.

Architecture	Train Set	Test Set	Gender Acc.	Race Acc.	Age MAE ↓
BIF+KCCA [15]	S1	S1+S3	98.5%	98.9%	4.00
	S2	S1+S3	98.4%	99.0%	3.95
	Average		98.5%	99.0%	3.98
BIF+KPLS [14]	S1	S2+S3	98.4%	99.0%	4.07
	S2	S1+S3	98.3%	99.0%	4.01
	Average		98.4%	99.0%	4.04
Soft Softmax[a] [30]	S1	S2+S3	–	–	3.14
	S2	S1+S3	–	–	2.92
	Average		–	–	3.03
Multi-scale CNN[b] [34]	S1	S2+S3	98.0%	99.1%	3.63
	S2	S1+S3	97.8%	98.1%	3.63
	Average		97.9%	98.6%	3.63
Net^{VGG}_{Hybrid}[a] [32]	S1	S2+S3	–	–	2.96
	S2	S1+S3	–	–	2.95
	Average		98.7%	99.2%	2.96
RaceGender2Age[a] [19] (VGG16)	S1	S2+S3	98.23%	97.78%	3.15
	S2	S1+S3	98.70%	97.99%	2.84
	Average		98.47%	97.89%	2.99
Fused Method[a] [19] (AlexNet)	S1	S2+S3	–	–	3.25
	S2	S1+S3	–	–	3.05
	Average		–	–	3.15
Fused Method[a] [19] (VGG16)	S1	S2+S3	–	–	3.09
	S2	S1+S3	–	–	2.81
	Average		–	–	2.95
DRFs [29]	S1	S2+S3	–	–	–
	S2	S1+S3	–	–	–
	Average		–	–	2.98
ARAN (AlexNet)	S1	S2+S3	98.81%	99.27%	3.15
	S2	S1+S3	98.92%	99.37%	2.76
	Average		*98.87%*	**99.32%**	2.96
ARAN+MP[b] (AlexNet)	S1	S2+S3	98.81%	99.27%	3.11
	S2	S1+S3	98.92%	99.37%	2.72
	Average		*98.87%*	**99.32%**	2.92
ARAN (VGG16)	S1	S2+S3	99.03%	99.14%	2.77
	S2	S1+S3	98.94%	99.22%	2.48
	Average		**98.99%**	*99.18%*	*2.63*
ARAN+MP[b] (VGG16)	S1	S2+S3	99.03%	99.14%	2.75
	S2	S1+S3	98.94%	99.22%	2.45
	Average		**98.99%**	*99.18%*	**2.60**

[a] The IMDB-WIKI database [27] was used for network pre-training.
[b] Adopting mirroring prediction techniques [8] in testing.
* Bold and italic fonts indicate the #1 and #2 performances.

prediction (MP) technique in our experiments, and it is only used for age estimation. Besides, most approaches use the Morph Album II dataset tend to pretrain their networks using the IMDB-WIKI dataset [27], which is a much larger dataset than the Morph Album II dataset. We train on a smaller dataset, but the proposed method show superior performance than the state-of-the-art approaches with a considerable margin. In Table 4, there is no doubt that our methods outperform the state-of-the-art traditional methods [14,15]. In terms of deep model methods, both of our network structures still outperform these methods. For instance, the fused method [19] (VGG16), which is fused with five cascaded structure frameworks, reduces the MAE of Morph Album II to 2.95, while our VGG16 framework boost the record to 2.63 under the same testing protocol without using IMDB-WIKI for pre-training. The mirroring prediction brings not significant improvement of approximately 0.03 or 0.04 for age estimation. The cumulative score (CS) curves in Fig. 7 show similar results. Our methods also outperforms all the state-of-the-art methods with a significant margin from CS1 to CS20.

For FG-NET, The quantitative comparisons are demonstrated in Table 5. We do not show our performances of gender and race, because the results are very close to 100%. As can be seen, ARAN achieves the state-of-the-art result with 3.79 MAE and 81.0% CS. It also illustrates that our ARAN can learn attribute-region association from a small dataset.

Table 5
Comparisons with the state-of-the-art methods on the FG-NET dataset.

Methods	Age MAE ↓	CS ↑
CA-SVR	4.67	74.5%
OHRank	4.48	74.4%
CAM	4.12	73.5%
DEX	4.63	N/A
DRFs	3.85	80.6%
ARAN (AlexNet)	3.79	81.0%

4.6. Evaluation on the hierarchical scheduling

We investigate the improvement of hierarchical scheduling training (HST) method and perform the experiments based on our multi-task baseline architecture on Morph Album II dataset. The results of the experiments are shown in Fig. 8 and Fig. 9. Firstly, we can see from Fig. 9 that compared with Fig. 4, HST can prevent early convergence of the ethnic and the gender tasks, and therefore prevent the age task from stacking in a local minimum. The high-

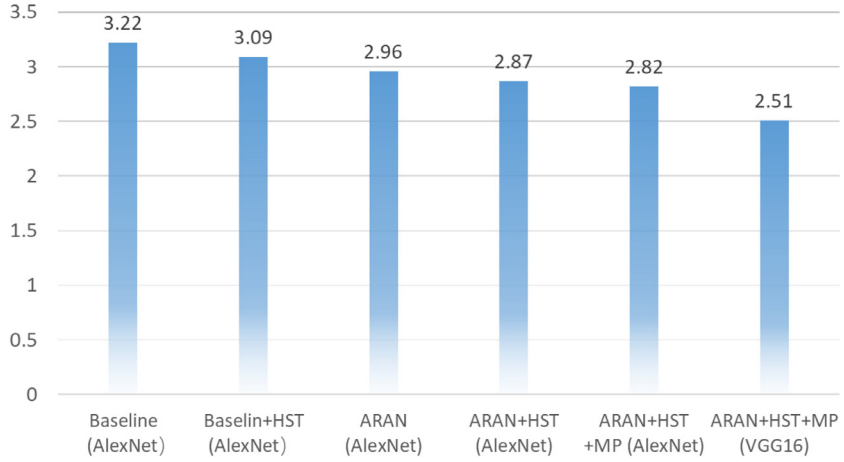


Fig. 8. Results of the improvement of HST on the Morph Album II dataset.

Fig. 9. The training processing with HST technique on the Morph Album II dataset (Training Set: S1, Testing Set: S2+S3).

est MAE point of HST method is around 6000th iteration, while the corresponding iteration in regular method is around 26,000th iteration. And the tasks of gender classification and ethnicity recognition remain the same trend in Figs. 4 and 9, which remain stable at around 2000th iteration. The Fig. 8 shows that our HST method greatly increases the performance of our multi-task learning. For instance, compared with the result in Fig. 4, the MAE of our multi-task baseline is improved by 0.13. Even for the single task of age estimation, our HST method can enhance the result from 3.20 to 3.09. Noticeably, owing to the drop of the testing samples, our result experience a small setback from 3.16 to 3.20 compared with the result of age estimation in [8], and however our method still can greatly improve the performance. For ARAN (AlexNet), our HST method successfully boosts the result to 2.87. With the helping of mirroring technique, our ARAN (VGG-16) achieves its lowest figure 2.51 MAE in Fig. 8. Though our HST method slightly reduces the performance (about 0.2%) of gender classification and ethnicity recognition, our method can greatly enhance performance the most difficult task (age estimation) in our method. Therefore, it can be argued that our hierarchical scheduling training is very helpful for age estimation in multi-task learning.

4.7. Evaluation on the loss functions in multi-task learning

In this section, we briefly discuss the influence of different weight combinations of loss function in multi-task learning. We conduct this experiment based on our multi-task baseline architecture on Morph Album II dataset. According to Eq. (4), we can set different weight for different loss function in multi-task learning, but it is very hard to find out a best combination for each loss function. In our experiment, we adopt a similar idea with our hierarchical scheduling training. We try to make our network pay more attention to age estimation which is the most difficult task, and slow down their speed of the convergence. In Fig. 4, we can see that the gender classification is hard to converge than the ethnicity, so we also try to assign a higher weight to gender classification, and a lower weight to ethnicity recognition. The results are shown in Fig. 10. We can see that if β and γ contain a same loss

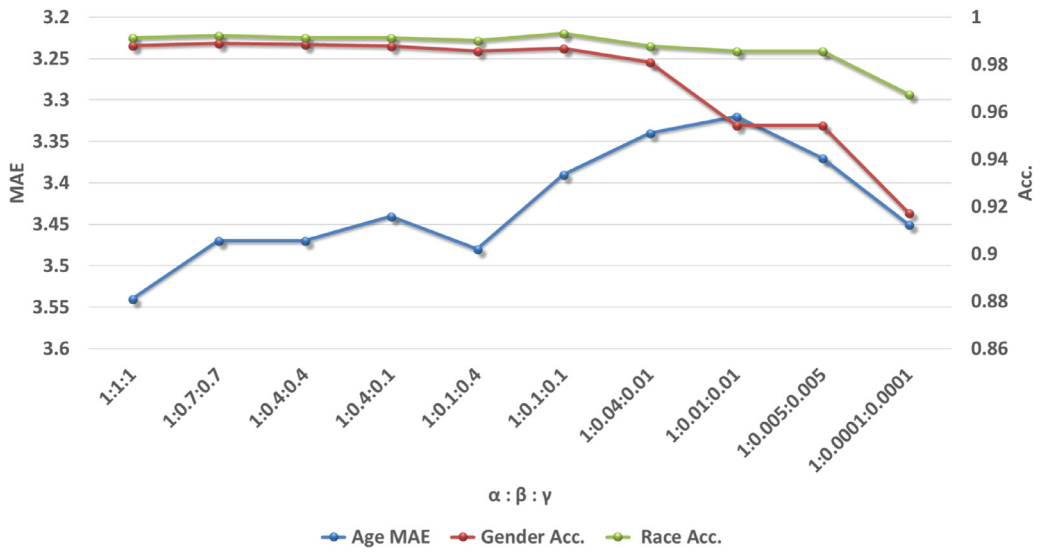


Fig. 10. Comparisons with different weight combination of loss function in attribute-Region association learning on the Morph Album II dataset (Training Set: S1, Testing Set: S2+S3).

weight with α, our performance of age estimation is dropped considerably. By contrast, if β and γ contain a very small loss weight, our performance of age estimation is increased significantly, and the MAE is even higher than the MAE of single-task baseline architecture. However, if loss weight of β and γ are too low, the performance of gender classification is also decreased dramatically. Therefore, in our experiments, we assign 1, 0.1 and 0.1 to α, β and γ so as to achieve a good compromise.

5. Conclusion

In this paper, we present an Attribute-Region Association Network (ARAN) for age estimation by learning the association of the face attributes and components. To this end, we design a novel framework which adopts different critical regions from the human face and combines different demographic information together in an unified network. Our framework balances and takes full advantage of these regions and demographic information for accurate prediciton. Extensive experiments show the proposed method outperforms existing state-of-the-art methods by a significant margin on the Morph Album II dataset. Also, we propose a hierarchical scheduling training method to address the complexity unbalancing problem in multi-task learning. Together with the proposed scheduling method, we achieve the highest performance of age estimation on the Morph Album II dataset. Our further research may focus on transferring the age knowledge to other face attributes recognition with one-shot or few-shot learning.

Declaration of competing interest

None.

Acknowledgments

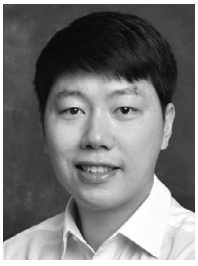
This project is supported by the National Natural Science Foundation of China (No. 61472145, No. 61972162, and No. 61702194), the Innovation and Technology Fund of Hong Kong (Project No. ITS/319/17), the Special Fund of Science and Technology Research and Development on Application From Guangdong Province (SF-STRDA-GD) (No. 2016B010127003), the Guangzhou Key Industrial Technology Research fund (No. 201802010036), the Guangdong Natural Science Foundation (No. 2017A030312008), and the CCF-Tencent Openfund.

References

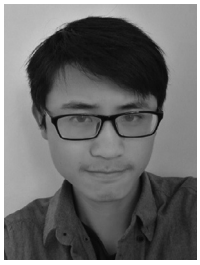
[1] G. Antipov, M. Baccouche, S.A. Berrani, J.L. Dugelay, Apparent age estimation from face images combining general and children-specialized deep learning models, in: Proceedings of the CVPR Workshops, 2016, pp. 801–809.
[2] K.-Y. Chang, C.-S. Chen, A learning framework for age rank estimation based on face images with scattering transform, IEEE TIP 24 (3) (2015) 785–798.
[3] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of the CVPR, IEEE, 2011, pp. 585–592.
[4] B.-C. Chen, C.-S. Chen, W.H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: Proceedings of the ECCV, Springer, 2014, pp. 768–783.
[5] K. Chen, S. Gong, T. Xiang, C. Change Loy, Cumulative attribute space for age and crowd density estimation, in: Proceedings of the CVPR, 2013, pp. 2467–2474.
[6] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-CNN for age estimation, in: Proceedings of the CVPR, 2017, pp. 5183–5192.
[7] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-CNN for age estimation, in: Proceedings of the CVPR, 2017.
[8] Y. Chen, Z. Tan, A.P. Leung, J. Wan, J. Zhang, Multi-region ensemble convolutional neural networks for high-accuracy age estimation, in: Proceedings of the BMVC, 2017.
[9] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Underst. 61 (1) (1995) 38–59.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the CVPR, IEEE, 2009, pp. 248–255.
[11] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE TPAMI 29 (12) (2007) 2234–2240.
[12] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware CNN model, in: Proceedings of the ICCV, 2015, pp. 1134–1142.
[13] G. Guo, G. Mu, Human age estimation: what is the influence across race and gender? in: Proceedings of the CVPR Workshops, IEEE, 2010, pp. 71–78.
[14] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: Proceedings of the CVPR, IEEE, 2011, pp. 657–664.
[15] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: CCA vs. PLS, in: Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.
[16] G. Guo, G. Mu, A framework for joint estimation of age, gender and ethnicity on a large database, Image Vis. Comput. 32 (10) (2014) 761–770.
[17] G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features, in: Proceedings of the CVPR, IEEE, 2009, pp. 112–119.
[18] H. Han, K.J. A, S. Shan, X. Chen, Heterogeneous face attribute estimation: a deep multi-task learning approach., IEEE TPAMI PP (99) (2017) 1–.
[19] W. Jun, T. Zichang, L. Zhen, G. Guodong, S.Z. Li, Auxiliary demographic information assisted age estimation with cascaded structure, IEEE Trans. Cybern. 48 (9) (2018) 2531–2541.
[20] Y.H. Kwon, et al., Age classification from facial images, in: Proceedings of the CVPR, IEEE, 1994, pp. 762–767.
[21] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, IEEE Trans. Syst. Man, Cybern., Part B (Cybern.) 34 (1) (2004) 621–628.
[22] T. Liu, J. Wan, T. Yu, Z. Lei, S.Z. Li, Age estimation based on multi-region con-

volutional neural network, in: Proceedings of the Chinese Conference on Biometric Recognition, 2016, pp. 186–194.
[23] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proceedings of the CVPR, 2016, pp. 4920–4928.
[24] H. Pan, H. Han, S. Shan, X. Chen, Mean-variance loss for deep age estimation from a face, in: Proceedings of the CVPR, 2018, pp. 5285–5294.
[25] N. Ramanathan, R. Chellappa, Modeling age progression in young faces, in: Proceedings of the CVPR, 1, IEEE, 2006, pp. 387–394.
[26] A.W. Rawls, K. Ricanek, Morph: Development and optimization of a longitudinal age progression database, in: Proceedings of the European Workshop on Biometrics and Identity Management, Springer, 2009, pp. 17–24.
[27] R. Rothe, R. Timofte, L. Van Gool, Dex: deep expectation of apparent age from a single image, in: Proceedings of the ICCV Workshops, 2015, pp. 10–15.
[28] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, IJCV (2016) 1–14.
[29] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, A.L. Yuille, Deep regression forests for age estimation, in: Proceedings of the CVPR, 2018, pp. 2304–2313.
[30] Z. Tan, S. Zhou, J. Wan, Z. Lei, S.Z. Li, Age estimation based on a single network with soft softmax of aging modeling, in: Proceedings of the ACCV, Springer, 2016, pp. 203–216.
[31] Q. Tian, S. Chen, Joint gender classification and age estimation by nearly orthogonalizing their semantic spaces, Image Vis. Comput. 69 (2018) 9–21.
[32] J. Xing, K. Li, W. Hu, C. Yuan, H. Ling, Diagnosing deep learning models for high accuracy age estimation from a single image, Pattern Recognit. 66 (2017) 106–116.
[33] S. Yan, H. Wang, X. Tang, T.S. Huang, Learning auto-structured regressor from uncertain nonnegative labels, in: Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[34] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Proceedings of the ACCV, Springer, 2014, pp. 144–158.

Yiliang Chen is a research assistant in the School of Computer Science and Engineering, South China University of Technology. He obtained his B.Sc. degree from Macau University of Science and Technology. His research interests include computer vision and image processing.



Shengfeng He is an Associate Professor in the School of Computer Science and Engineering, South China University of Technology. He was a Research Fellow at City University of Hong Kong. He obtained his B.Sc. degree and M.Sc. degree from Macau University of Science and Technology, and the Ph.D. degree from City University of Hong Kong. His research interests include computer vision, image processing, computer graphics, and deep learning.



Zichang Tan received the B.E. degree from the department of Automation, Huazhong University of Science and Technology (HUST), Wuhan, China in 2016. He was named as outstanding graduates of the college when he graduated. He is currently studying for a Ph.D. degree at the Institute of Automation, Chinese Academy of Science (CASIA). His main research interests include deep learning, face attribute analysis and face recognition.

Chu Han graduated from South China Agricultural University in 2011 with a B.Sc. degree in computer science. He received his M.Phil. degree in computer science from South China University of Technology in 2014, under the supervision of Prof. Xuemiao Xu. He is now pursuing his Ph.D. degree in the Department of Computer Science and Engineering of the Chinese University of Hong Kong, under the supervision of Prof. Tien-Tsin Wong. His current research interests include computer graphics, image processing, pattern recognition, and computer vision.

Guoqiang Han received the B.Sc. degree from the Zhejiang University, Hangzhou, China, in 1982, and the master's and Ph.D. degrees from the Sun Yat-sen University, Guangzhou, China, in 1985 and 1988, respectively. He is a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou. He was the dean of the School of Computer Science and Engineering. He has published over 100 research papers. His current research interests include multimedia, computational intelligence, machine learning, and computer graphics.

Jing Qin received his Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2009. He has been an assistant professor at The Hong Kong Polytechnic University from 2016. His research interests include visualization, human-computer interaction and deep learning.