Auxiliary Demographic Information Assisted Age Estimation With Cascaded Structure

Jun Wan, *Member, IEEE*, Zichang Tan, Zhen Lei, *Senior Member, IEEE*,
Guodong Guo, *Senior Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Owing to the variations including both intrinsic and extrinsic factors, age estimation remains a challenging problem. In this paper, five cascaded structure frameworks are proposed for age estimation based on convolutional neural networks. All frameworks are learned and guided by auxiliary demographic information, since other demographic information (i.e., gender and race) is beneficial for age prediction. Each cascaded structure framework is embodied in a parent network and several subnetworks. For example, one of the applied framework is a gender classifier trained by gender information, and then two subnetworks are trained by the male and female samples, respectively. Furthermore, we use the features extracted from the cascaded structure frameworks with Gaussian process regression that can boost the performance further for age estimation. Experimental results on the MORPH II and CACD datasets have gained superior performances compared to the state-of-the-art methods. The mean absolute error is significantly reduced from 3.63 to 2.93 years under the same test protocol on the MORPH II dataset.**

***Index Terms*—Age estimation, convolutional neural networks (CNNs), demographic, Gaussian process regression (GPR).**

I. Introduction

HUMAN face attributes play fundamental roles in real-world applications [1], such as video surveillance, security control and human–computer interaction. As a part of face attributes, facial age estimation has gained a lot of attentions. Age estimation from face images has started since 1994 by the work in [2]. Later, some public age datasets, (FG-NET [3] and MORPH II [4]) were released and some new methods [5], [6] or popular features [7], [8] are proposed, which have improved the performance a lot for age estimation. In recent years, because of the big achievements in the imagenet classification challenge [9], the deep convolutional neural network (CNN) has gained increasing attentions in many areas, such as image classification [9] and segmentation [10], object detection [11], facial point detection [12], and face recognition [13]. Therefore, some works [14]–[16] based on the deep CNN are proposed for age estimation, which have indicated very promising results on MORPH II [4] and ChaLearn apparent age dataset [17].

Fig. 1. (a) and (b) or (c) and (d) Comparison between different genders (male versus female) having the same race. (a) and (c) or (b) and (d) Comparison between different races having the same gender (white versus black). All the faces appeared in this figure are labeled with 19 years old on MORPH II. It shows that the appearances of face images from different genders or races are very varied, even though they are of the same year.

However, owing to the variations including both intrinsic and extrinsic factors (i.e., illumination, expression, occlusion, race, or gender), accurate age estimation from face images is still a challenging problem. As shown in Fig. 1, the appearances vary a lot even though all the faces are labeled with 19 years old. We can see that:

1) when the faces are from the same gender or race, the appearances will be effected by facial expression, illumination or partial occlusions (extrinsic factors);
2) the appearances of face images from different genders or races (intrinsic factors) are very varied, although they are of the same year;
3) the effects of intrinsic and extrinsic factors are both important for age estimation in face images.

For extrinsic factors, there are a lot of works [18]–[21] to deal with them. Gross and Brajovic [18] proposed a image preprocessing algorithm that compensates for illumination variations in images by estimating and compensating the illumination field, while Li *et al.* [19] presented a novel solution

Manuscript received March 23, 2017; revised June 27, 2017; accepted August 16, 2017. Date of publication January 23, 2018; date of current version August 16, 2018. This work was supported in part by the National Key Research and Development Plan under Grant 2016YFC0801002, in part by the Chinese National Natural Science Foundation Projects under Grant 61502491, Grant 61473291, Grant 61572501, and Grant 61572536, in part by the Science and Technology Development Fund of Macau under Grant 112/2014/A3, and in part by the NVIDIA GPU Donation Program and AuthenMetric Research and Development Funds. This paper was recommended by Associate Editor L. Rutkowski. *(Corresponding author: Jun Wan.)*

J. Wan, Z. Tan, Z. Lei, and S. Z. Li are with the Center for Biometrics and Security Research and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@nlpr.ia.ac.cn; zichangtan2016@nlpr.ia.ac.cn; zlei@nlpr.ia.ac.cn; szli@nlpr.ia.ac.cn).

G. Guo is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: guodong.guo@mail.wvu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2741998

2168-2267 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

for illumination invariant using near-infrared images. Then, the sparse coding-based method [21] is proposed, which is robust to partial occlusion, varying facial expression, illumination, and disguise. Later, to tackle difficult lighting conditions under uncontrolled environments, the work [20] combined the strengths of illumination normalization, local texture and kernel-based feature representations.

Nevertheless, there are relatively less research works [14], [22] on tackle the effects of intrinsic factors. Han *et al.* [22] explored automatic demographic estimation face images using demographic information which consists of age, gender, and race as explained in [22]. They proposed a hierarchical age estimator including a classification stage and a following regression stage. In the classification stage, a two-level binary decision tree is build to classify the query face image into one of the four groups (i.e., male-white, male-black, female-white, and female-black). Then a separate support vector machine (SVM) regressor is trained within each group to make an age prediction in the regression stage. Although it has gained low mean absolute error (MAE) (3.8 years) without quality assessment on MORPH II from experimental results in [22], it is still the traditional methods that is feature extraction with classification or regressor training.

Yi *et al.* [14] proposed a multiscale CNN for age estimation. The main metric in [14] is an end-to-end system to estimate age from image pixels directly, instead of hand-crafted feature designing. The multiscale CNN [14] is treated as a multitask learning work, which means it can simultaneously estimate age, gender and race from a face image. It has gained lower MAE (3.6 years) on MORPH II. Although the multiscale CNN used demographic information, it still cannot assess the effect of gender or race for age estimation.

Inspired by Han *et al.* [22] which proposed a hierarchical age estimation [i.e., binary decision tree for classifying nonoverlapping groups (e.g., male versus female and white versus black) and within-group age regressors learned from overlapping age groups] and deep learning technique, we proposed different cascade networks using demographic information. However, the differences between our method and [22] are: 1) we applied the demographic information in CNNs with cascade structures and 2) we used the linear regression or Gaussian process regression (GPR) to estimate age value of each face image.

In this paper, we propose a novel cascaded structure framework for age estimation to reduce the effect of intrinsic factors (i.e., demographic information) on age estimation. As shown in Fig. 1, the appearances are very diverse for different demographics. So we explore the demographic information to guide the learning of the proposed cascade frameworks. Here, we explore five cascaded structure frameworks which are Gender2AgeNet, Race2AgeNet, Age2AgeNet, GenderRace2AgeNet, and RaceGender2AgeNet, as shown in Fig. 2. To evaluate the proposed frameworks, we use two networks: 1) a popular deep network (VGG-16 [23]) and 2) a shallow network (see Fig. 3) proposed in this paper. The main contributions of this paper are summarized below.

1) The proposed frameworks use a divide-and-conquer strategy to greatly improve the accuracy of age estimation.
2) The proposed frameworks are designed for age estimation, but they can also be used for age or gender classification, such as Gender2AgeNet or RaceGender2AgeNet (see Fig. 2).
3) In order to demonstrate the effectiveness of the features extracted from the proposed frameworks, GPR instead of linear regression is used, which can further boost the performance of age estimation.

The rest of this paper is organized as follows. Related works are reviewed in Section II. The proposed algorithm is introduced in Section III. Then, extensive experiments are provided in Section IV to evaluate and compare our method with the state-of-the-art. Section V gives some discussion of the proposed method. Finally, the conclusion is given in Section VI.

II. Related Work

Human age estimation from face images has been studied for over 20 years. In the general methods, these works [2], [6], [7], [24] includes two stages: 1) feature extraction (feature representation from face images) and 2) regression or classification (age prediction with the extracted features).

For feature extraction, some works used geometric features to classify the age into three groups (i.e., baby, young, or senior adult). The popular geometric features [2], [25] included chin drop, skin wrinkles, nose drop or mustache. Although geometric features can discriminate baby and adult, it cannot deal with the adult and old people. Later, the most representative feature, biologically inspired feature (BIF), proposed by Guo *et al.* [7], is widely used by many works [6], [22] for age estimation. They used Gabor filters with smaller sizes and suggested to determine the number of bands and orientations in a problem-specific manner [7]. However, while the BIF feature is carefully designed in a handcrafted way, we explore the integrations of automatic feature extraction and regression (or classification) based on the proposed cascade structure frameworks.

Then, the next stage is to achieve age estimation. Usually, age estimation can be treated as a classification or regression problem. Kwon and Lobo [2] categorized facial images as age group classification. But there were only 47 images in the experimental dataset, and the correct accuracy for baby group was below 68%. The work in [26] used five classifiers to predict the age group by adopting the majority decision rule. The final accuracy can achieve 74% for three groups: 1) 0–15; 2) 15–30; and 3) above 30. Then, Sai *et al.* [27] applied the extreme learning machine [28]–[30] for age groping using the extracted features, namely local Gabor binary pattern [31], BIF, and Gabor features. The reported accuracy was about 70% from their experiments on MORPH II. From the above discussions, most of the existing age estimation methods use handcrafted features (such as BIF and geometric features) and shallow models (i.e., SVM). However, with the combination

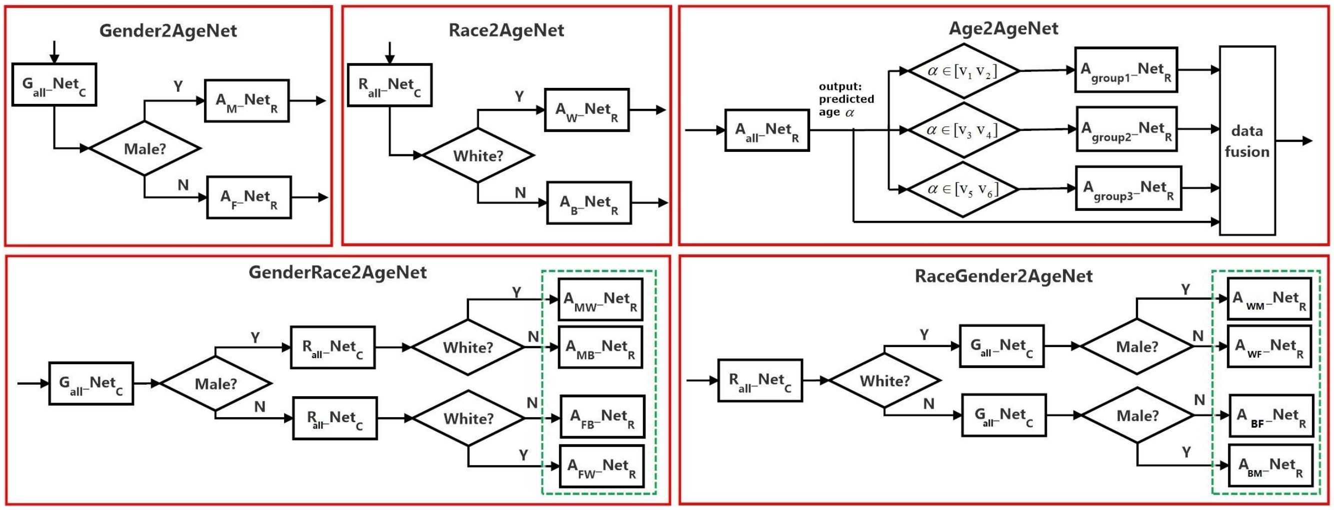


Fig. 2. Schematic of five cascaded structure frameworks (left to right, top to bottom) used the demographic information for age estimation: Gender2AgeNet, Race2AgeNet, Age2AgeNet, GenderRace2AgeNet, and RaceGender2AgeNet.

of these low-level features, the shallow models are very hard to improve the accuracy for age estimation.

Hence, in the past two years, there have been a few deep models for age estimation proposed in [14]–[16], [32], and [33]. The deep models are usually based on CNN and have achieved promising results. Yi *et al.* [14] proposed a multiscale CNN method trained from local aligned face patches to simultaneously achieve age estimation, race and gender classification. Wang *et al.* [32] proposed a CNN framework to extract high level features obtains in different layers of CNN for age estimation. Levi and Hassner [33] used a deep-CNN for age and gender classification and achieved the state-of-the-art result on the recent Adience [34] benchmark. Rothe *et al.* [16] proposed a deep expectation method for apparent age from a single image based on VGG-16 architecture [23] and obtained the first place of the ChaLearn LAP 2015 challenge on apparent age estimation.

III. Proposed Method

A. Cascaded Structure Frameworks

To facilitate the description of different networks, the symbol $\mathrm{DI}_{\mathrm{data}}\_\mathrm{Net}_f$ is defined, where DI is demographic information (i.e., A-age, G-gender, and R-race); net represents CNN chosen from one of our two basic networks (deep or shallow net); $_{\mathrm{data}}$ denotes training samples from different gender or race (i.e., all-all training samples, M-male, F-female, W-white, B-black, WM-white male, BM-black male, BF-black female, and WF-white female); $_f$ is defined as the network function (i.e., $_C$-gender or race classification and $_R$-regression for age estimation). For example, a gender classifier $G_{\mathrm{all}}\_\mathrm{Net}_C$ (see Gender2AgeNet in Fig. 2) is trained using all the training samples from the dataset, while the age regressor $A_{\mathrm{WM}}\_\mathrm{Net}_R$ is trained using training samples from white male persons.

As shown in Fig. 2, we explore five cascaded structure frameworks using demographic information. In the structure frameworks, two basic nets are alternative: 1) a shallowed network shown in Fig. 3 and 2) a popularly deep network, namely VGG-16 [23]. The shallow network is similar to Alexnet [9], but the length of our shallow network is less than that of Alexnet. The proposed frameworks apply a divide-and-conquer strategy to improve the accuracy of age estimation. Before the description of the proposed frameworks in detail, we first describe two basic nets as regressor or classifier used in our algorithm.

1) Basic Nets for Classification and Regression: For age estimation, we treat it as the regression problem using linear regression as the objective function in both shallow and deep nets. That is because the predicted age is a real value. For gender or race recognition, it can be treated as a binary classification problem. Hence, we use softmax as the objective function in basic nets for gender and race recognition.

Specifically, in the training set $\{(x_{(1)}, y_{(1)}), \ldots, (x_{(n)}, y_{(n)})\}$ of n labeled examples, the input features are $x_{(i)} \in R^n$ and labels are $y_{(i)} \in [0, k-1]$. For age estimation, the loss function is defined

$$l = -\frac{1}{2n}\left(\sum_{i=1}^{n}\left\|\hat{y}_i - y_i\right\|_2^2\right) \tag{1}$$

where y_i is the true age and $\hat{y}_i$ is a prediction for the ith training sample.

For gender or race recognition, the loss function is defined as

$$l(\theta) = -\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=0}^{k-1} p_j \log\left(\hat{p}_j\right)\right) \tag{2}$$

where θ denotes the softmax layer parameters; p_j is the target probability distribution, where if $y_{(i)} = j$, then $p_j = 1$, otherwise $p_j = 0$; $\hat{p}_j = (e^{\theta_j^T x_i})/(\sum_{l=0}^{k-1} e^{\theta_l^T} x_i)$ is the predicted probability distribution.

Therefore, we can train three basic models via (1) and (2).

1) $G_{all}\_Net_C$*:* In Fig. 3, the shallow net has one full connection layer with size 64. We added a new full connection layer with size 2 in the shallow net, because it is used to train a binary classifier for gender recognition. Here, we

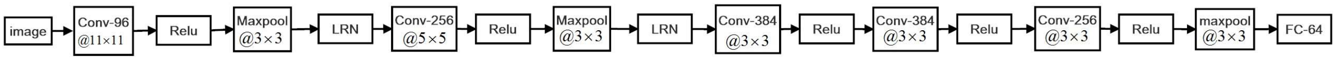


Fig. 3. It shows a shallow network used in this paper. This network includes five convolutional layers, three max pooling layers, and one fully connected layer. The kernel size (number values after the symbol @ in this figure) and the number of feature maps of each convolutional layer (value after the symbol *conv* in this figure) are given for each convolutional and max pooling layers. LRN denotes local response normalization.

utilized the gender information of all training samples and chose Softmax as the objective function.

2) R_{all}_*Net*$_C$*:* This net is the same as G_{all}_Net$_C$. The only difference is the race information used to train R_{all}_Net$_C$, which is a classifier for race recognition.
3) A_{all}_*Net*$_R$*:* Considering age estimation as a regression problem, we chose the Euclidean loss [see (1)] as the objective function in the deep and shallow nets. When the deep VGG-16 net is used, the size of the last full connection is 1 for age regression (instead of 1000 in original VGG-16 net). If the shallow net in Fig. 3 is used, one new full connection with size 1 is added in the last layer of the shallow net. A_{all}_Net$_R$ is used to predict a real value for a face image to do age estimation.

As shown in Fig. 2, three basic models as parent networks are utilized in the proposed frameworks.

a) Gender2AgeNet: In the training stage, besides G_{all}_Net$_C$ for gender recognition, we trained two age regressors: 1) A_M_Net$_R$ and 2) A_F_Net$_R$ for age estimation. The training set S is divided into two sets: 1) training images from male S_M and 2) training images from female S_F, where $S = \{S_M, S_F\}$. Then A_M_Net$_R$ and A_F_Net$_R$ are trained by S_M and S_F, respectively. We note that A_M_Net$_R$ and A_F_Net$_R$ are finetuned by A_{all}_Net$_R$. In the testing stage, the query face image is first recognized as male or female by G_{all}_Net$_C$. If the recognition result is male, then A_M_Net$_R$ is used to predict a real value for this image. Otherwise, this image is fed into A_F_Net$_R$.

b) Race2AgeNet: The second framework is similar to Gender2AgeNet. In the training stage, besides R_{all}_Net$_C$ for race classification (white or black), two age regressors are trained: 1) A_W_Net$_R$ and 2) A_B_Net$_R$ for age estimation. The training set S is divided into two sets: 1) training images from white persons S_W and 2) training images from black persons S_B, where $S = \{S_W, S_B\}$. Then A_W_Net$_R$ and A_B_Net$_R$ are trained by S_W and S_B, respectively. Besides, A_W_Net$_R$ and A_B_Net$_R$ are finetuned by A_{all}_Net$_R$.

c) Age2AgeNet: The third framework is different from the above two frameworks which include a classifier to recognize gender or race and two subnetworks for age estimation. The data is divided into n overlapped groups, each of which is used for training an age regressor. Each group has training images if the truth label satisfies the following conditions:

$$\begin{aligned} S_{\text{group1}} &= \{S \in [v_{11}\ v_{12}]\}, \quad \text{s.t. } v_{11} < v_{12} \\ S_{\text{group2}} &= \{S \in [v_{21}\ v_{22}]\}, \quad \text{s.t. } v_{11} < v_{21} < v_{12} < v_{22} \\ &\cdots \\ S_{\text{group}n} &= \{S \in [v_{n1}\ v_{n2}]\}, \ \text{s.t. } v_{(n-1)1} < v_{n1} < v_{(n-1)2} < v_{n2} \end{aligned} \tag{3}$$

where S denotes training set, $S_{\text{group1}}, S_{\text{group2}}, \ldots, S_{\text{group}n}$ are the overlapped training sets, and v_{ij}, $i = 1, 2, \ldots, n, j = 1, 2$ is the predefined threshold. After getting n overlapped groups, we train n subnetworks from their corresponding groups, which are A_{group1}_Net$_R$, A_{group2}_Net$_R, \ldots, A_{\text{group}n}$_Net$_R$. All three subnetworks are finetuned by A_{all}_Net$_R$.

In the testing stage, we first predict a real value α for a query face image using the parent network A_{all}_Net$_R$. Then, according to (3), we can find α belonging to which age range. Then we fuse the predicted results via (4) to obtain the predicted age $\overline{\alpha}$ for the query image

$$\overline{\alpha} = \frac{\alpha + \sum_{i=1}^{n} \alpha_i * \delta(\alpha_i)}{1 + \sum_{i=1}^{n} \delta(\alpha_i)} \tag{4}$$

where α_i is calculated from $A_{\text{group}i}$_Net$_R$ $(i = 1, 2, \ldots, n)$; $\delta(\alpha_i) = 1$, if $\alpha \in [v_{i1}\ v_{i2}]$, otherwise $\delta(\alpha_i) = 0$.

d) GenderRace2AgeNet: There are three classifiers and four regressors. As the first classifier, G_{all}_Net$_C$ is one of the basic models for gender recognition. Then, R_{all}_Net$_C$ is trained with all samples for race classification. Later, the training set S is divided into four sets: 1) training images from white male S_{WM}; 2) training images from black male S_{BM}; 3) training samples from white female S_{WF}; and 4) training samples from black female S_{BF}, where $S = \{S_{\text{WM}}, S_{\text{BM}}, S_{\text{WF}}, S_{\text{BF}}\}$. And we trained four regressors, A_{MW}_Net$_R$, A_{MB}_Net$_R$, A_{FW}_Net$_R$ and A_{FB}_Net$_R$ using the corresponding training set, respectively. In the training stage, A_{MW}_Net$_R$ and A_{MB}_Net$_R$ are finetuned by A_M_Net$_R$, while A_{FW}_Net$_R$ and A_{FB}_Net$_R$ are finetuned by A_F_Net$_R$.

e) RaceGender2AgeNet: In Fig. 2, RaceGender2AgeNet also has four regressors which are the same regressors as GenderRace2AgeNet. In the testing stage, the query face image is first recognized as white or black via R_{all}_Net$_C$. Then, if the recognition result is white (or black), then the image is continuously recognized as male or female via G_{all}_Net$_C$. Finally, a real value as the predicted age is obtained via one of four regressors. Similar to GenderRace2AgeNet, the four regressors are finetuned by A_W_Net$_R$ and A_B_Net$_R$ according to the race information.

Note that finetuning is an important strategy in our cascaded frameworks. With the cascaded networks deepening, the age regressors are more special and designed for certain group of people. However, as shown in Fig. 2, after the database is divided into several parts by race or gender information, the images of each part decreases dramatically. Due to this, we use finetuning strategy to cope with the scarcity of face images for each age regressor and the details have been mentioned above.

B. Gaussian Process Regression

In Section III-A, the regressors of the proposed frameworks used linear regression [see (1)]. Unlike combining linear regression function with the shallow or deep (VGG-16) net to train a CNN model, we directly use features extracted from the trained regressor models in Section III-A. For example, refer to Gender2AgeNet of Fig. 2, the feature x of a face image is extracted from the trained model A_M_Net$_R$ or A_F_Net$_R$. For the shallow nets, the size of feature x is 64 (the length of the last full connection in Fig. 3). For the deep net (VGG-16), the feature x is of size 4096.

Gaussian process is convenient for flexible nonlinear regression problems. In this paper, age estimation is considered as a regression problem with Gaussian processes after feature extraction.

Each observation y (i.e., ground truth age label) can be equal to an function $f(x)$ through a Gaussian noise model at input features x

$$y = f(x) + N\left(0, \sigma_n^2\right) \tag{5}$$

where we assume that the observation noise has a Gaussian distribution $N(0, \sigma_n^2)$ with its mean value 0 and standard deviation σ. The object of inference is the latent function f, which is given a Gaussian process prior. This implies that any finite subset of latent variables, $F = \{f(x_i)\}_{i=1}^n$ have a multivariate Gaussian distribution. In particular, for the given feature set $X = \{x_i\}_{i=1}^n$, the latent variables have a distribution

$$p(F|X) = N\left(F|\mu, K_{f,f}\right) \tag{6}$$

where $K_{f,f}$ is the covariance matrix and μ is the mean function. Each element in the covariance matrix is a realization of covariance function $[K_{f,f}]_{i,j} = k(x_i, x_j)$, which represents the prior assumptions of the smoothness of the laten function [35]. In our approach, the covariance function is the stationary squared exponential

$$k\left(x_i, x_j\right) = \sigma_f{}^2 \exp\left[\frac{-(x_i - x_j)^2}{2l_d^2}\right] \tag{7}$$

where σ_f is the scaling parameter and l_d is the length scale. So we can know that all the hyperparameters of GPR can be represented as $\theta = \{l_d, \sigma_f, \sigma_n\}$.

According to the Bayes' theorem, assuming we have little prior knowledge about θ, we maximize marginal likelihood $\log p(y|x, \theta)$ by

$$\log p(y|x, \theta) = -0.5 y^T K^{-1} y - 0.5 \log |K| - 0.5 n \log 2\pi. \tag{8}$$

To optimize (8), the conjugate gradient is applied to seek hyperparameters θ.

IV. Experiments

A. Datasets

We evaluate the performance of our method on three age datasets, MORPH II database for controlled environment and CACD database for uncontrolled environment. By the way, ChaLearn apparent age competition also offer an age database in the uncontrolled environment, and it is more close to the real life comparing with the CACD database. Because it does not provide age labels of test set, we only provides the performances on validation set of the ChaLearn database.

1) MORPH II: As far as we known, MORPH II is the only large age dataset with accurate age labeling. This dataset includes about 55 000 face images and age ranges from 16 to 77 years. Although it is a good and large database, the distributions of gender and race are uneven. The male–female ratio is about 5.5:1 and the white–black ratio is about 4:1. Except for white and black, the proportion of other race is very low. Therefore, in our experiments, we employ two typical protocols for evaluation on MORPH II dataset.

1) *S1-S2-S3 Protocol:* We follow the work [14], [36] to split MORPH II into three nonoverlapped subsets S_1, S_2, S_3 randomly. These three subsets are constructed by two rules: a) male–female ratio is equal to three and b) white–black ratio is equal to one. In our experiments, we totally use the same test protocols[1] provided by Yi *et al.* [14]. That is all experiments are repeated two times: a) training set: S_1 and testing sets: $S_2 + S_3$ and b) training set: S_2 and testing sets: $S_1 + S_3$.
2) *80-20 Protocol:* Following the experimental setting in [37]–[39], a subset of 5493 images was used, where the images are selected from Caucasian descent to reduce the cross-race influence. All face images in this protocol are white, and no other race images, which means we cannot use race information in our cascade frameworks. We also randomly split the whole dataset into two nonoverlapped parts: a) 80% images for training and b) 20% images for testing. In this evaluation way, the number of testing images is a quarter of training images.

2) CACD: CACD database is collected from the Internet movie database (IMDB), and it is the largest public cross-age database. This database includes more than 160 thousands images of 2000 celebrities taken from 2004 to 2013 (ten years in total). The age ranges from 16 to 62. However, unlike MORPH II database with precise age labels, the CACD database contains much noise and only 200 celebrities were checked and their noisy images were removed. So far only very few people conducted evaluation on this database because of the noise. We evaluate this database through this way: we randomly split those clean images from 200 celebrities, of which 150 celebrities are used for training and left 50 celebrities for testing. And those noisy images of other 1800 celebrities are used for pretrain.

3) ChaLearn LAP: The ChaLearn LAP dataset [17] contains 4691 images in total, which is released by ChaLearn LAP competition 2015. This is the first age dataset for apparent age estimation, where each image was labeled by at least ten users with two Web-based applications and then the averaged age is used as the final annotation. This dataset offers the standard deviation for each age label. It is split into three subsets, where training set has 2476 images, validation set includes 1136 images and the left 1079 images are used for testing.

B. Evaluation Metrics

In this paper, we use the MAE and cumulative score (CS) as the evaluation criterion. For the MAE calculation, it computes

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

Fig. 4. Results of face alignment.

the MAE between the true age and the predicted age in the testing set. Formally, MAE is calculated as

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N}\left|l_k - \hat{l}_k\right| \tag{9}$$

where l_k and $\hat{l}_k$ denote the ground truth age and predicted age of the kth image, respectively, and N is the number of testing images. If the value of MAE is lower, the performance is more better.

The CS is calculated as follows:

$$\mathrm{CS}(j) = \frac{N_{e\leq j}}{N} \times 100\% \tag{10}$$

where $N_{e\leq j}$ is the number of the testing facial images whose absolute error between the estimated age and the ground truth age is not greater than j years. The value of CS is higher, the performance is better.

For apparent age estimation, the ϵ-error is used as a quantitative measure, which is proposed by the ChaLearn LAP competition. The ϵ-error is computed as

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{11}$$

It not only measures the error between the predicted value x and the averaging labeled age μ, but also takes into consideration the standard deviation σ. The final ϵ-error is the average over all predictions.

C. Preprocessing

The preprocessing of face images includes face alignment and data augmentation steps.

1) Face Alignment: In order to align the faces, we first detect the facial landmarks by active shape models [40]. Then, we crop and rotate face image according to the center positions of two eyes (see Fig. 4).

2) Data Augmentation: Because of the difficulty in collecting face images with accurate ages, the popular age datasets are limited. For example, the MORPH II has about 55 000 face images while training set only includes about 10 000 face images. Therefore, we attempted to augment data from the training images. For every training image, we flip it left to right and rotate it by $\pm 5°$, $\pm 10°$. And we add gaussian white noises of mean M and variance V, where $M = 0$ and $V = \{0.001, 0.005, 0.1, 0.015, 0.02\}$ for the original, flipped, and rotated images. Therefore, we can obtain 36 images from a training sample.

D. Results on MORPH II

When conducting experiments on MORPH II, we divide images into three groups and set $v_{11} = 16, v_{12} = 40$, $v_{21} = 30$, $v_{22} = 60$, $v_{31} = 50$, and $v_{32} = 77$ for Age2AgeNet. We use SGD and mini-batch size of 64. The shallow net shown in Fig. 3 is directly trained from the training samples of MORPH II while the VGG-16 net is first pretrained using IMDB-WIKE dataset[2] [16]. That is because it is difficult to train a good model only used MORPH II for the deep net. For instance, the average MAE is 7.96 when the deep net A_{all}_Net$_R$ is trained on MORPH II. It means that the deep net A_{all}_Net$_R$ is underfitting. When we first trained the deep net using IMDB-WIKE dataset and then finetuned A_{all}_Net$_R$ on MORPH II, the average MAE is drastically decreased to 3.13.

1) Experiments Under S1-S2-S3 Protocol: Table I shows the results of our proposed cascaded frameworks for the shallow and deep net. We can see the following.

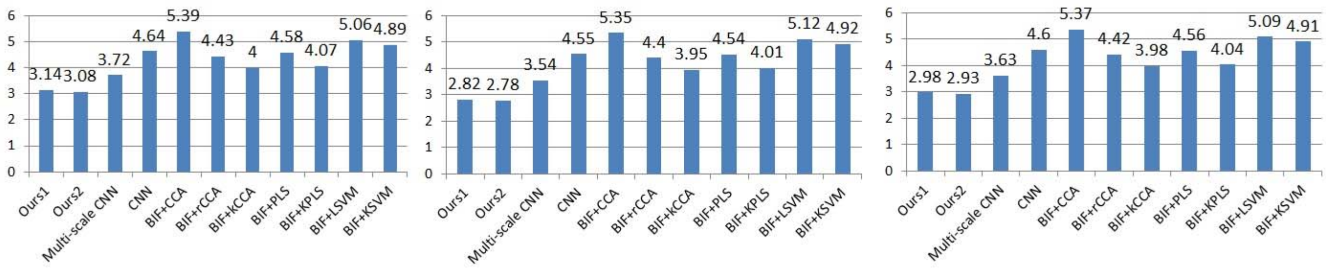
1) The performances of the deep net are better than the shallow net for the same cascade framework.
2) Any of our cascaded frameworks are better than the single model whether the shallow or deep net is used. For example, under the framework of shallow net, compared with A_{all}_Net$_R$, the average MAE without GPR of RaceGender2AgeNet is below 0.22.
3) The cascaded framework is deeper, the result is better. For instance, the MAE of GenderRace2AgeNet (or RaceGender2AgeNet) is better than Gender2AgeNet, Race2AgeNet, and Age2AgeNet.
4) The performances of GenderRace2AgeNet are comparative to RaceGender2AgeNet, and Gender2AgeNet, Race2AgeNet, and Age2AgeNet also have similar results. Moreover, it demonstrates that deep cascade networks (i.e., GenderRace2AgeNet and RaceGender2AgeNet) can get better results than shallow cascade networks (i.e., Race2AgeNet and Gender2AgeNet). That is because age estimation can be beneficial from other gender and race information. If both gender and race information is used, the better performance will be occurred (i.e., GenderRace2AgeNet and RaceGender2AgeNet).
5) The cascaded networks combined with GPR can get better performances. And the average MAE is reduced by 0.02 from Table I.
6) The best result of the proposed five cascade frameworks is 2.98 which is from GenderRace2AgeNet and RaceGender2AgeNet with GPR by the deep net.
7) It also shows that the gender and race accuracies are more than 98% and 97%, respectively. Owing to their high recognition accuracies, the gender and race have a little effect on age estimation. Therefore, our proposed cascade framework is reasonable.

Furthermore, we can obtain five predicted values from the proposed five cascaded frameworks for any query face image, and then simply calculate the average value as the final predicted age. The performances of the fused method are also provided in Table I. We can see that the fused method can still boost the performances and the best average MAE is 2.93, which is better than the cascaded frameworks at least 0.05 in MAE.

[2]https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

TABLE I
COMPARISONS WITH DIFFERENT NETWORKS BY THE SHALLOW AND DEEP NETWORK ON THE MORPH II WITH $S1$-$S2$-$S3$ PROTOCOL

Nets	Train Set	Test Set	Shallow net				Deep net				Gender accuracy	Race accuracy
			MAE	Average MAE	MAE (GPR)	Average MAE (GPR)	MAE	Average MAE	MAE (GPR)	Average MAE (GPR)		
$A_{all}\_Net_R$	$S1$	$S2+S3$	3.447	3.40	3.423	3.37	3.212	3.13	3.203	3.13	–	–
	$S2$	$S1+S3$	3.355		3.326		3.054		3.052		–	–
Gender2AgeNet	$S1$	$S2+S3$	3.413	3.29	3.377	3.25	3.166	3.03	3.017	2.95	98.23%	–
	$S2$	$S1+S3$	3.167		3.128		2.901		2.891		98.70%	–
Race2AgeNet	$S1$	$S2+S3$	3.397	3.27	3.363	3.23	3.165	3.03	3.160	3.02	–	97.78%
	$S2$	$S1+S3$	3.140		3.105		2.902		2.888		–	97.99%
Age2AgeNet	$S1$	$S2+S3$	3.409	3.28	3.374	3.25	3.212	3.08	3.201	3.07	–	–
	$S2$	$S1+S3$	3.151		3.121		2.943		2.935		–	–
GenderRace2Age	$S1$	$S2+S3$	3.289	3.18	3.285	3.15	3.143	2.99	3.140	2.98	98.23%	97.78%
	$S2$	$S1+S3$	3.071		3.021		2.839		2.825		98.70%	97.99%
RaceGender2Age	$S1$	$S2+S3$	3.287	3.18	3.279	3.14	3.145	2.99	3.141	2.98	98.23%	97.78%
	$S2$	$S1+S3$	3.070		3.003		2.838		2.824		98.70%	97.99%
Fused method	$S1$	$S2+S3$	3.248	3.15	3.214	3.11	3.089	2.95	3.078	2.93	–	–
	$S2$	$S1+S3$	3.051		2.998		2.811		2.782		–	–



(a) (b) (c)

Fig. 5. Comparisons with other state-of-the-art methods for age estimation on MORPH II with $S1$-$S2$-$S3$ protocol. All the methods are compared under the same test protocol. It shows that the average MAE is greatly reduced from 3.63 to 2.93. The methods are: Ours1-RaceGender2AgeNet by the deep net and Ours2-Fused method by the deep net; multiscale CNN [14]; CNN [41]; BIF+CCA [6]; BIF+rCCA [6]; BIF+KCCA [6]; BIF+PLS [42]; BIF+KPLS [42]; BIF+LSVM [6]; and BIF+KSVM [6]. (a) MAE (train set: $S1$; test set: $S2+S3$). (b) MAE (train set: $S2$; test set: $S1+S3$). (c) Average MAE of (a) and (b).

TABLE II
CSS (%) OF THE CASCADED FRAMEWORKS

Nets	2	4	6	8	10
$A_{all}\_Net_R$	41.09	69.76	85.92	93.83	97.53
Gender2AgeNet	41.97	70.16	86.19	94.12	97.60
Race2AgeNet	41.63	70.13	86.30	94.07	97.56
Age2AgeNet	41.36	69.45	85.69	93.83	97.41
GenderRace2Age	42.06	70.39	86.40	94.30	97.62
RaceGender2Age	41.83	70.49	86.49	94.21	97.61
Fused method	**42.54**	**71.12**	**87.07**	**94.65**	**97.82**

Then, we compare our method with other state-of-the-art methods. The comparisons are shown in Fig. 5 under the same test protocol. Except our method, multiscale CNN [14] has got the best MAE of 3.63 from Fig. 5. When the training set is $S1$ and the testing set is $S2$ and $S3$, the performances of RaceGender2AgeNet and fused method by the deep net are improved by 0.58 and 0.64 [see Fig. 5(a)], respectively. When the training set is $S1$ and the testing set is $S1$ and $S3$, the performances of RaceGender2AgeNet and fused method by the deep net are also improved by 0.72 and 0.76. And the final MAE of our method has achieved the best performance.

Finally, the CSs of the cascaded frameworks with the deep net are shown in Table II where it used the train set $S1$, test set $s2+s3$. It shows that the fusion method can get the best performances. Because the similar conclusion is obtained for the cascaded frameworks with the shallow net and GPR as Table II, the CSs of GPR (or the shallow net) are not given.

2) Experiments Under 80-20 Protocol: Another popular 80-20 protocol is also widely used in previous works. Therefore, some more experiments are shown in Table III. In this table, it shows that the deep learning-based methods (i.e., our methods, DEX [43], and VGG+SVR [44]) are better than traditional methods (AGES [24], CA-SVR [38], OHRank [37], and DLA [32]). Because the race of all faces is white in the 80-20 protocol, there are only results of Race2AgeNet, GenderRace2AgeNet, and RaceGender2AgeNet in Table III. We have reimplemented the DEX method with same processing steps in our method and the result is shown in the table without initialized from other dataset (i.e., IMDB-WIKI dataset), while the MAE result is slightly from the result reported in [43] which may due to some differences in detail (such as face detection and alignment). As shown in this table, the performance of Gender2AgeNet is comparative to DEX [43] and our Age2AgeNet and fused methods are better than other state-of-the-art methods listed.

E. Results on CACD

Unlike MORPH II database with a relatively even distribution among white and black people, almost all the face images of CACD database are white people. Because of this,

TABLE III
EXPERIMENTAL RESULTS ON MORPH II DATABASE WITH 80-20 PROTOCOL. NOTE THAT REIMPLEMENTED DEX AND OUR METHOD ARE EXPERIMENTED WITH THE SAME PREPROCESSING AND EXPERIMENTAL SETTINGS

Method	MAE
AGES [24]	8.83
CA-SVR [38]	5.88
OHRank [37]	5.69
DLA [32]	4.77
VGG+SVR [44]	3.45
DEX [43]	3.33 (3.25 in [43])
Ours(Gender2AgeNet)	3.33
Ours(Age2AgeNet)	**3.32**
Ours(Fused method)	**3.30**

TABLE IV
COMPARISONS WITH DIFFERENT NETWORKS BY THE DEEP NET FOR CACD

Nets	Train Set	Test Set	MAE	MAE (GPR)
BIF+LR [45]	150 celebrities	50 celebrities	7.79	7.75
BIF+SVR [6]	150 celebrities	50 celebrities	7.67	7.65
BIF+SVM [6]	150 celebrities	50 celebrities	8.19	8.16
DFD+LR [45]	150 celebrities	50 celebrities	8.16	8.13
$A_{all}\_Net_R$	150 celebrities	50 celebrities	5.34	5.30
Gender2AgeNet	150 celebrities	50 celebrities	5.27	5.25
Age2AgeNet	150 celebrities	50 celebrities	5.28	5.27
Fused method	150 celebrities	50 celebrities	5.24	5.22

we only perform the Gender2AgeNet and Age2AgeNet on this database, without the cascaded structure related to ethnicity. For Age2AgeNet, we divide faces into four parts and we set $v_{11} = 14, v_{12} = 30, v_{21} = 20, v_{22} = 40, v_{31} = 30, v_{32} = 50, v_{41} = 40$, and $v_{42} = 62$. From the results on MORPH II database, we can see that our cascaded structures are suitable for both shallow and deep nets, and the performance of deep net is better than shallow. So we only conduct the experiments for deep (VGG-16) net on CACD database. Note that we do not use IMDB-WIKI database to pretrain the deep net because of the duplicated images of two databases.

As far as we know, only Liu *et al.* [45] conducted the experiments on CACD database for age estimation. According to it, we also evaluate other age estimators adopting the same training and testing sets as the proposed method. The comparisons are shown in the Table IV.

From Table IV, we can find that our cascaded framework can also be useful for age estimation in the uncontrolled environment. Comparing with $A_{\text{all}}\_\text{Net}_R$, Gender2AgeNet and Age2AgeNet get better performances whether with or without GPR. Furthermore, through the fused method, we can further boost the performance and reduce the MAE to 5.23 on CACD database, which outperforms other methods by a big margin. Moreover, Fig. 6 shows the CSs of different methods, which also demonstrate that our method has significantly improved age estimation accuracy compared to other methods.

F. Results on ChaLearn LAP

Following works [16], [43], [46], we used the pretrained model which is trained on IMDB-WIKI dataset. Then, the pretrained model initialized our VGG-16 network for our

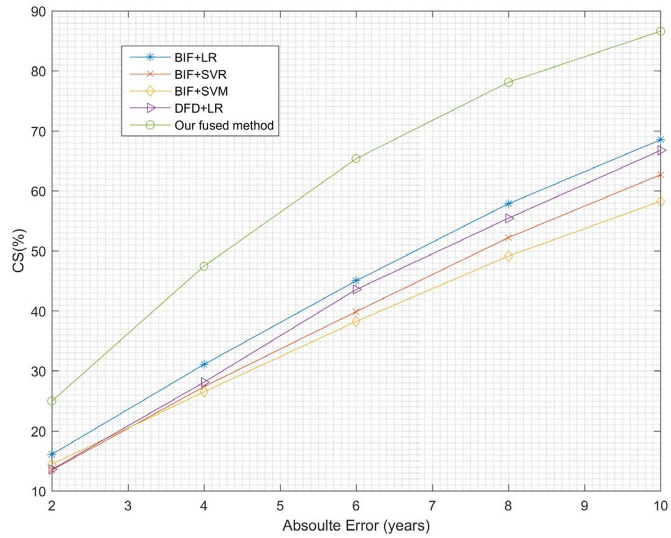


Fig. 6. CSs of different methods, such as BIF+LR [45], BIF+SVR [6], BIF+SVM [6], and DFD+LR [45].

TABLE V
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE CHALEARN APPARENT DATASET

Rank	Team	Validation Set MAE	Validation Set ϵ-error	Pretrain Set	Network
–	$A_{all}\_Net_R$ (Ours)	3.30	0.29	IMDB-WIKI	VGG-16
1	CVL_ETHZ [16], [43]	3.25	0.28	IMDB-WIKI	VGG-16
2	ICT-VIPL [46]	3.33	0.29	FG-NET,Morph, CACD, et al.	GoogleNet
3	WVU_CVL [47]	–	0.31	FG-NET, Morph, CACD, et al.	GoogleNet
4	SEU_NJU [15]	–	0.34	FG-NET, Morph, Adience[33], et al.	GoogleNet

appearance age estimation task. Because this dataset does not include other demographic information (i.e., gender and race), we cannot use the cascaded structure frameworks, such as Gender2AgeNet, Race2AgeNet, GenderRace2AgeNet, and RaceGender2AgeNet. Moreover, there is only about 2400 images in the training set. If we use Age2AgeNet, we need group the training set into at least two subsets. That means each subset will have fewer images that can lead to overfitting easily and cause difficultly in training the CNN. According to the above reasons, we can only provide the performance of $A_{\text{all}}\_\text{Net}_R$ on the ChaLearn LAP dataset.

Experimental results are shown in Table V. It shows that our $A_{\text{all}}\_\text{Net}_R$ has got ϵ-error $= 0.29$, which is comparative to the performances of CVL_ETHZ (ϵ-error $= 0.28$) and ICT-VIPL (ϵ-error $= 0.29$). Based on the evaluation metric of MAE, although our $A_{\text{all}}\_\text{Net}_R$ is slight worst than CVL_ETHZ, $A_{\text{all}}\_\text{Net}_R$ is better than ICT-VIPL. Therefore, our method can achieve high performance in the ChaLearn LAP appearance age estimation.

V. DISCUSSION

In this section, we analyze the effectiveness of augmentation operation, the convergence of the proposed method, and the

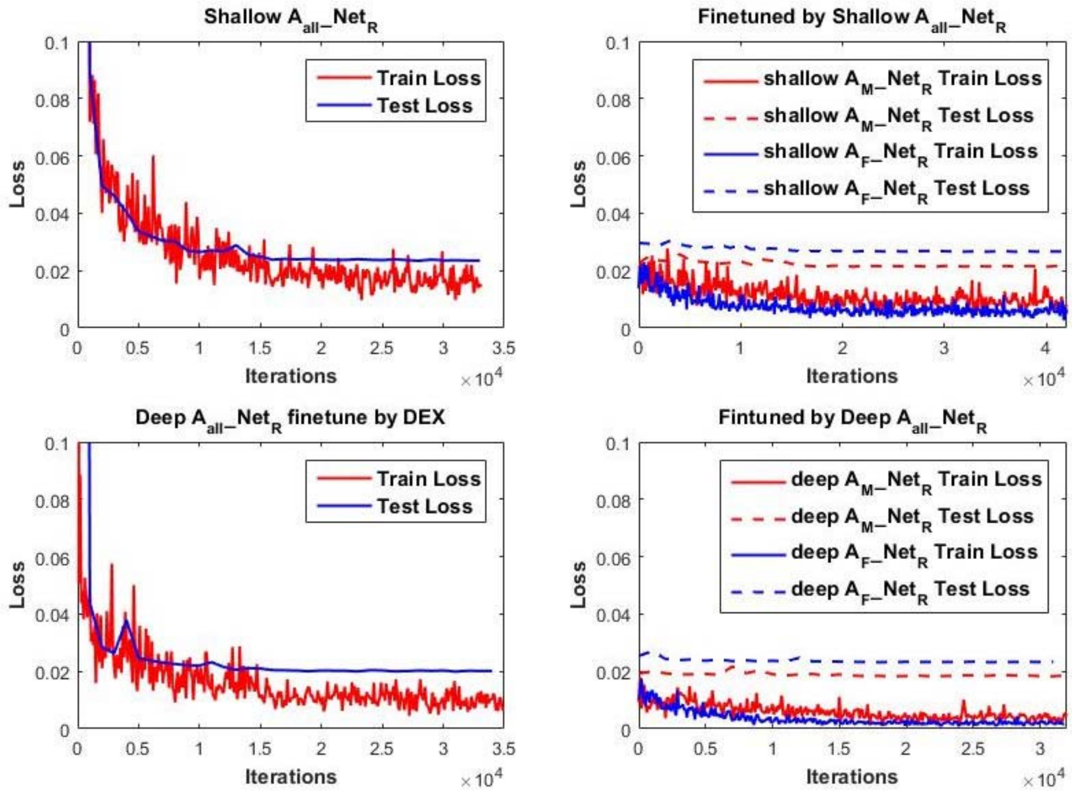


Fig. 7. It shows training loss and testing loss on $S2$ set of MORPH II.

running time of the cascaded frameworks. Besides, a prototype application is introduced using our proposed framework.

A. Data Augmentation

For the testing protocol of MORPH II database we used, the images of training set is only about a quarter of the testing set. We synthesize virtual image samples to make full use of the single training image, and it also could prevent the over-fitting of training model to a certain extent because of the diversity increase in training set. In our experiments, the MAE for training with and without augmentation is 3.40 and 3.49, with $A_{\text{all}}\_\text{Net}_R$ of shallow net for testing, which shows our augmentation method is useful.

B. Convergence Analysis

Fig. 7 shows the training and testing loss of MORPH II when training with $S1 + S3$ and testing with $S2$. Whether the shallow or deep $A_{\text{all}}\_\text{Net}_R$ is used, the network converges to 0.02 for testing loss and 0.01 for the training loss. For the $A_M\_\text{Net}_R$ and $A_F\_\text{Net}_R$, because they are finetuned by $A_{\text{all}}\_\text{Net}_R$, the training loss and the testing loss are very little and the networks also converge.

C. Running Time Analysis

We have tested 2000 images from MORPH II and calculated the average running time (including the time of image loading from the hard disk) of the proposed frameworks under a NVIDIAN TITAN X GPU. The results are shown in Table VI

TABLE VI
RUNNING TIME COMPARISON AMONG DIFFERENT CASCADED STRUCTURE FRAMEWORKS (IN MILLISECOND)

Nets	Shallow net (Alexnet)		Deep net (VGG-16)	
	Time (ms)	MAE	Time (ms)	MAE
$A_{all}\_Net_R$	8.1	3.40	22.6	3.13
Gender2AgeNet	15.3	3.29	31.3	3.03
Race2AgeNet	15.1	3.27	30.2	3.03
Age2AgeNet	16.8	3.28	52.8	3.08
GenderRace2Age	21.8	3.18	35.1	2.99
RaceGender2Age	22.0	3.18	35.7	2.95

where it also provides the average *MAEs* of their corresponding networks. We can see that the cascaded frameworks cost less 22 ms/image in shallow net which means they are suitable for the real-time applications. For the deep net, the cascaded frameworks cost less 36 ms/image except Age2AgeNet. That is because Age2AgeNet includes two VGG-16 nets to predict age value, while other frameworks include only one VGG-16 used for predicting age value with one or two Alexnet nets for gender or race classification. Moreover, compared with the deep net, the cascaded frameworks with shallow net are below about 0.16 in MAE while faster about 18.1 ms/image.

D. Prototype Application Introduction

In order to demonstrate the effectiveness of the proposed method, we also developed a prototype application.[3] In this

[3]http://www.cbsr.ia.ac.cn/users/jwan/face/age.html

Web link, one can upload a face image and test our age estimation system. There are two points one should note. First, the training dataset used in our application includes more that 400 000 images and the dataset is private. Second, owing to the high time complexity of VGG architectures, we selected the Alexnet network as a tradeoff between the accuracy of age estimation and the running time (see Table VI).

VI. Conclusion

In this paper, five structure frameworks based on CNN for age estimation have been proposed, which are learned and guided by demographic information. Also, we have used GPR instead of linear regression to predict age after feature extraction from the CNNs. From our experimental results, our method has greatly improve the accuracy of age estimation under the same testing protocol. Besides the cascaded structure frameworks, a joint framework which does a multitask of multiple face attributes is a good alternative. Our further work will focus on designing a reasonable multitask architecture for age, gender and race estimation jointly.

References

[1] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. London, U.K.: Springer, 2011.

[2] Y. H. Kwon and N. D. V. Lobo, "Age classification from facial images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 1994, pp. 762–767.

[3] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

[4] A. W. Rawls and K. Ricanek, Jr., "MORPH: Development and optimization of a longitudinal age progression database," in *Biometric ID Management and Multimodal Communication*. Heidelberg, Germany: Springer, 2009, pp. 17–24.

[5] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," *Pattern Recognit.*, vol. 46, no. 3, pp. 628–641, 2013.

[6] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, 2013, pp. 1–6.

[7] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 112–119.

[8] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim, "Age estimation using a hierarchical classifier based on global and local facial features," *Pattern Recognit.*, vol. 44, no. 6, pp. 1262–1281, 2011.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[12] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, 2013, pp. 3476–3483.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 1701–1708.

[14] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis.*, Singapore, 2014, pp. 144–158.

[15] X. Yang *et al.*, "Deep label distribution learning for apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Santiago, Chile, 2015, pp. 344–350.

[16] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Santiago, Chile, 2015, pp. 252–257.

[17] S. Escalera *et al.*, "ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. ChaLearn Looking People Workshop ICCV*, vol. 1. Santiago, Chile, 2015, pp. 243–251.

[18] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Audio- and Video-Based Biometric Person Authentication*. Heidelberg, Germany: Springer, 2003, pp. 10–18.

[19] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[20] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[21] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[22] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

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

[24] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[25] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. New York, NY, USA, 2006, pp. 387–394.

[26] P. Thukral, K. Mitra, and R. Chellappa, "A hierarchical approach for human age estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 1529–1532.

[27] P.-K. Sai, J.-G. Wang, and E.-K. Teoh, "Facial age range estimation with extreme learning machines," *Neurocomputing*, vol. 149, pp. 364–372, Feb. 2015.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[29] N. Wang, M. J. Er, and M. Han, "Generalized single-hidden layer feedforward networks for regression problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1161–1176, Jun. 2015.

[30] N. Wang, M. J. Er, and M. Han, "Parsimonious extreme learning machine using recursive orthogonal least squares," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1828–1841, Oct. 2014.

[31] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1. Beijing, China, 2005, pp. 786–791.

[32] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Win. Conf. Appl. Comput. Vis.*, 2015, pp. 534–541.

[33] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, 2015, pp. 34–42.

[34] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.

[35] J. Vanhatalo, P. Jylänki, and A. Vehtari, "Gaussian process regression with student-t likelihood," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 1910–1918.

[36] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, San Francisco, CA, USA, 2010, pp. 71–78.

[37] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 585–592.

[38] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2467–2474.

[39] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4920–4928.

[40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[41] M. Yang, S. Zhu, F. Lv, and K. Yu, “Correspondence driven adaptation for human profile recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 505–512.

[42] G. Guo and G. Mu, “Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 657–664.

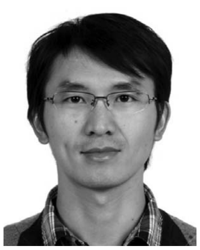
[43] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *Int. J. Comput. Vis.*, pp. 1–14, Aug. 2016.

[44] R. Rothe, R. Timofte, and L. Van Gool, “Some like it hot—Visual guidance for preference prediction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 5553–5561.

[45] T. Liu, Z. Lei, J. Wan, and S. Z. Li, “DFDnet: Discriminant face descriptor network for facial age estimation,” in *Proc. 10th Chin. Conf. Biometrics*, Tianjin, China, 2015, pp. 649–658.

[46] X. Liu *et al.*, “AgeNet: Deeply learned regressor and classifier for robust apparent age estimation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Santiago, Chile, Dec. 2015, pp. 258–266.

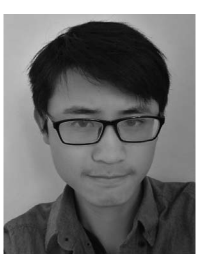
[47] Y. Zhu, Y. Li, G. Mu, and G. Guo, “A study on apparent age estimation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Santiago, Chile, Dec. 2015, pp. 267–273.



Jun Wan (M’16) received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, in 2015.

Since 2015, he has been an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing. He has published papers in top journals, such as *Journal of Machine Learning Research*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON CYBERNETICS. His current research interests include computer vision, machine learning, especially for gesture and action recognition, and facial attribution analysis, such as age estimation, facial expression, gender, and race classification.

Mr. Wan was a recipient of the 2012 ChaLearn One-Shot-Learning Gesture Challenge Award, sponsored by Microsoft and ICPR 2012, and the 2013 and 2014 Best Paper Award from the Institute of Information Science, Beijing Jiaotong University. He has served as a Reviewer for several top journals and conferences, such as *Journal of Machine Learning Research*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, Pattern Recognition, International Conference on Pattern Recognition 2016, International Conference on Computer Vision and Pattern Recognition 2017, International Conference on Computer Vision 2017, and International Conference on Automatic Face and Gesture Recognition 2017.



Zichang Tan received the B.E. degree from the Department of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Science, Beijing, China.

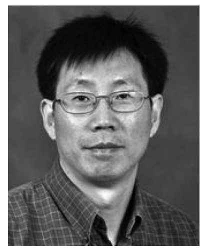
His current research interests include deep learning, face attribute analysis, and recognition.

Mr. Tan was a recipient of the outstanding graduates of the Huazhong University of Science and Technology.

Zhen Lei (SM’16) received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2010.

He is currently an Associate Professor with CASIA. He has published over 90 papers in international journals and conferences. His current research interests include computer vision, pattern recognition, image processing, and face recognition.

Dr. Lei served as the Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.



Guodong Guo (M’07–SM’07) received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, Madison, WI, USA.

He is an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, WV, USA. In the past, he visited and researched in several places, including INRIA, Sophia Antipolis, France, Ritsumeikan University, Kyoto, Japan, Microsoft Research, Beijing, and North Carolina Central University, Durham, NC, USA. He authored a book entitled *Face, Expression, and Iris Recognition Using Learning-Based Approaches* (2008), co-edited a book entitled *Support Vector Machines Applications* (2014), and published about 100 technical papers. His current research interests include computer vision, machine learning, and multimedia.

Dr. Guo was a recipient of the North Carolina State Award for Excellence in Innovation in 2008, the Outstanding Researcher at College of Engineering and Mineral Resources (CEMR) and WVU in 2013 and 2014, and the New Researcher of the Year at CEMR and WVU in 2010 and 2011. He was selected as the “People’s Hero of the Week” by Broadband and Social Justice Blog under the Minority Media and Telecommunications Council in 2013. Two of his papers were selected as “The Best of FG’13” and “The Best of FG’15,” respectively.

Stan Z. Li (F’09) received the B.Eng. degree from Hunan University, Changsha, China, the M.Eng. degree from the National University of Defense Technology, Changsha, and the Ph.D. degree from Surrey University, Guildford, U.K.

He is currently a Professor and the Director of Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was a Researcher with Microsoft Research Asia, Beijing, from 2000 to 2004. Prior to that, he was an Associate Professor with Nanyang Technological University, Singapore. He has published over 200 papers in international journals and conferences, and authored and edited eight books. His current research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance.

Dr. Li was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the Editor-in-Chief of Encyclopedia of Biometrics. He served as the Program Co-Chair for the International Conference on Biometrics 2007 and 2009, and has been involved in organizing other international conferences and workshops in the research fields. He is a member of the IEEE Computer Society.