An Ensemble CNN2ELM for Age Estimation

Mingxing Duan, Kenli Li, *Senior Member, IEEE*, and Keqin Li, *Fellow, IEEE*

***Abstract*— Age estimation is a challenging task, because it can be easily affected by gender, race, and other intrinsic and extrinsic attributes. At the same time, performing age estimation for a narrow age range may lead to better results. In this paper, to achieve robust age estimation, an ensemble structure referred to as CNN2ELM, which includes convolutional neural network (CNN) and extreme learning machine (ELM), is proposed for age estimation. The three-level system includes feature extraction and fusion, age grouping via an ELM classifier, and age estimation via an ELM regressor. Age-Net, Gender-Net, and Race-Net are trained using different targets, such as age class, gender class, and race class, respectively, and the three networks are used to extract features corresponding to age, gender, and race from the same image of a person during validation and test stages. Features related to the age property are enhanced by fusing these of race and gender properties. Then, to achieve a narrow age range, the ELM classifies the fusion results into one of the age groups. Afterward, an age decision is made using an ELM regressor. Our network is pretrained on an ImageNet database and then fine-tuned on the IMDB–WIKI database. The recently released Adience benchmark, ChaLearn Looking at People 2016 (LAP-2016), and MORPH-II are used to verify the performance of "Race-Net + Age-Net + Gender-Net + ELM classifier + ELM regressor (RAGN)." RAGN outperforms the existing state-of-the-art age estimation methods. The mean absolute error of the age estimation of RAGN for MORPH-II is determined to be 2.61 years; the accuracy of the age estimation for the Adience benchmark is 0.6649; and the normal score (ϵ) for the sequestered test set of the LAP-2016 data set is 0.3679.**

***Index Terms*— Age estimation, convolutional neural network, ensemble, extreme learning machine.**

I. Introduction

A. Motivation

IN RECENT years, facial age estimation has drawn a large amount of attention in computer vision due to its crucial applications in video surveillance, internet access control, security, and demography [1]. For example, the Age Specific Human Computer Interaction (ASHCI) system has helped to control underage drinking and smoking and prevented children from surfing harmful web pages.

However, human age estimation is a complicated process that is easily affected by many factors, such as identity, gender, race, and extrinsic factors, including lifestyle, environment, body pose, and facial expression. Human facial image datasets have different statistical characteristics [2], *e.g.*, regarding two people of the same age, one may be a female with a face that appears young, while the other may be a male with a face that appears old. Gender and race greatly influence age estimation. For example, [3] performed a systematic and quantitative study on the performance of age estimation across races and/or genders. The mean absolute error (MAE) of age estimation without crossing is 4.96 years, while that for crossing gender, crossing race, and crossing gender and race combined is 7.41, 8.38, and 9.77 years, respectively. Therefore, age estimation is closely related to gender and race.

All information, including race, gender, age, and other intrinsic and extrinsic attributes, exists in the form of pixels of input images, and methods used to extract wanted features are based on the corresponding targets, such as race labels, gender labels, and age labels. For example, during the process of age estimation, when the target is age class, the extracted features are closely related to the age attribute from the pixels of the input image, while most gender, race, and other traits that are not associated with the target are not extracted, even though age estimation is easily affected by these attributes. More importantly, before age estimation, a higher accuracy result will be achieved when the input features have more discriminative information. Few studies have fullly considered race and gender traits for age estimation. Therefore, for an image of human face, when we extract different kinds of features related to race, gender, and age, and then fuse them, a more robust feature set will be obtained. Of course, the origin features are enhanced via the approach, and the fused features are beneficial to age estimation. We believe that this research can offer a general guide for age estimation using large databases.

At the same time, a good age estimation method should predict age from a face image precisely, and age estimation methods have drawn the attention of researchers. Currently, many estimation systems have been developed and can be divided into two classes: age feature extraction and age estimation. Obtaining a low MAE for age estimation has heavily relied on the quality of the extracted features [4]. It not only demands that the features have the most differentiable characteristics among different classes but also that they retain unaltered characteristics within the same class. A large amount of

Manuscript received April 6, 2017; revised August 14, 2017 and October 11, 2017; accepted October 20, 2017. Date of publication October 25, 2017; date of current version December 19, 2017. This work was supported in part by the National Outstanding Youth Science Program of National Natural Science Foundation of China under Grant 61625202, in part by the International (Regional) Cooperation and Exchange Program of National Natural Science Foundation of China under Grant 61661146006, and in part by the National Key R&D Program of China under Grant 2016YT80201900. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anderson Rocha. *(Corresponding author: Kenli Li.)*

M. Duan is with the College of Information and Engineering, Hunan University, Changsha 410082, China, also with the School of Computer Science, National University of Defense Technology, Changsha 410073, China, and also with the Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China (e-mail: duanmingxing16@nudt.edu.cn).

K. Li is with the College of Information Science and Engineering, Hunan University, Changsha 410082, China (e-mail: lkl@hnu.edu.cn).

K. Li is with the College of Information Science and Engineering, Hunan University, Changsha 410082, China, and also with the Department of Computer Science, State University of New York, New Paltz, NY 12561, USA (e-mail: lik@newpaltz.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/TIFS.2017.2766583

research on age estimation has achieved great success through age feature extraction. For example, researchers have successfully used an anthropometric model [5], active appearance model (AAM) [6], AGing pattern Subspace (AGES) [7], age manifold [8], patch-based appearance model [9], local binary patterns (LBP) [10], biologically inspired features (BIF) [11], and deep neural networks [12]. In recent years, due to its strong ability to learn robust and discriminative features, CNN has been highlighted in machine learning and pattern recognition fields. It has achieved a state-of-the-art performance in image recognition and can automatically extract facial features.

Additionally, the process of making an age decision makes full use of features extracted from a facial dataset, so that a high discrimination age prediction system can be built (*i.e.*., age classifier or age regressor). By utilizing machine learning approaches to train a model for extracting features, an age estimation system can make an age prediction of queried faces. In general, age estimation can be considered to be a classification problem [13], [14], regression problem [8], [9], or combination of both [15]. We can model age estimation as a multi-classification and SVM is one of the most popular approaches for this type of classification. Recently, a large amount of researchers have studied age regression models, such as support vector analysis (CCA) [16], partial least squares (PLS) [17], support vector regression (SVR) [11], among others. ELM has been proven to be an efficient and fast classification algorithm because of its good generalization performance, fast training speed, and need for little human intervention [18]. Additionally, ELM and improved ELM, including mixing with other methods, have been widely used to process pattern recognition tasks to achieve a good performance [19].

Decision fusing and age grouping are two successful age estimation methods. Decision fusing is widely used to fuse multiple decisions to achieve a more robust decision. With more discriminative features and more powerful age estimation models, higher recognition rate will be obtained. As we know, a set of age estimation models with similar training facial datatsets will have different generalization performances. In such cases, combining the outputs of different models may lower the risk of an unsuitable selection of a poorly performing estimation model [20]; for example, asking different doctors' suggestions before performing a major operation or considering users' evaluations before purchasing a product as we usually make decisions by combining different opinions in our everyday lives. Our goal is to improve the probability of making the right decision by weighing different decisions and combining them to reach a final goal.

For age grouping, a large number of methods have been proposed to conduct age estimation, and many of them make age estimations over a very wide age range. In general, making an age estimation from a narrow age range may lead to better results. For example, we can estimate an age in the range of 30 to 35 more easily than in the range of 0 to 70. The goal of age grouping is to classify a facial image into an age group, and a final age decision is made based on the age groupings.

B. Our Contributions

Therefore, age estimation should not only consider race and gender as factors but also yield a novel structure in which differentiable features can be extracted. Motivated by the analysis above, we propose a novel age estimation architecture called CNN2ELM, which is used to estimate age based on face images. CNN2ELM includes three convolutional neural network (CNN) structures and two extrame learning machine (ELM) structures. Our proposed networks are pretrained on the ImageNet dataset [21] and then fine-tuned on the IMDB-WIKI dataset [22] before being fine-tuned on MORPH-II, Adience benchmark, and ChaLearn Looking at People 2016 (LAP-2016) datasets. During the validation and test stages, the three networks are used to extract features from the same image, and a discriminative and robust feature set is achieved by fusing these features. Based on the fusion features, ELM classifies these features into one of the age groups. Then, our ELM regressor makes a final decision. It not only sufficiently exploits the CNN but also utilizes the outstanding classification and regression properties of the ELM. The major contributions of this paper are as follows:

- We propose an ensemble CNN2ELM approach for performing age estimation. The structure combines the CNN with the ELM to perform age estimation in a hierarchical fashion, in which the CNN is used to extract features, while the ELM predicts age based on the image of a person age. The system full utilizes the advantages of the CNN and ELM.
- The highlights of our CNN2ELM structure are feature enhancement and age grouping. Our system comprises three CNNs and two ELMs (*i.e.*, Race-Net + Age-Net + Gender-Net + ELM classifier + ELM regressor (RAGN)"). To enhance the features extracted using Age-Net, RAGN fuses the features corresponding to the race and gender characteristics for the same face images into discriminative and robust features. For the age prediction system, the ELM classifier classifies the face image into one of the age groups to carry out an age estimation for a narrow age range, while the ELM regressor performs the final age estimation.
- Finally, we present the process of integrating the synergy of the hybrid structure in detail, including the design of the layers in the CNN, feature extraction and fusion, and age grouping. More importantly, extensive experiments are conducted using the MORPH-II, Adience benchmark, and LAP-2016.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III provides preliminary information. Section IV discusses the architecture of the CNN2ELM model. The experiments and results are illustrated in Section V. Finally, we draw conclusions in Section VI.

II. Related Work

A. Hybrid Neural Network System

CNN has been successfully applied to various fields, and image recognition, in particular, is a hot area of research.

However, few researchers have studied hybrid neural networks. Lawrence *et al.* [23] presented a hybrid neural-network solution for face recognition that made full use of the advantages of self-organizing map (SOM) neural networks and CNN. That approach showed a higher accuracy compared with other methods of face recognition. In 2012, Niu and Suen [24] introduced a hybrid classification system for objection recognition by integrating the synergy of CNN and SVM, and their experimental results showed that the method improved the classification accuracy. Liu *et al.* [25] used CNN to extract features and used a Conditional Random Field (CRF) to classify the deep features. Through extensive experiments on different datasets, such as the Weizmann horse, Graz-02, MSRC-21, Stanford Background, and PASCAL VOC 2011, the hybrid structure obtained a better segmentation performance compared with other methods on the same datasets. Xie *et al.* [26] used a hybrid representation method to process scene recognition and domain adaption. Convolutional parts were used to extract features, and the mid-level local representation (MLR) and convolutional Fisher vector (CFV) representation made the most of local discriminative information in the input images. Next, a SVM classifier was used to classify the hybrid representation and achieved better accuracy. Recently, Tang *et al.* [27] proposed a hybrid structure of a Deep Neural Network (DNN) and ELM to detect ships on space-borne images. DNN was used to process high-level feature representation and classification, while ELM was used for effective feature pooling and decision making. Furthermore, extensive experiments were presented to demonstrate that the hybrid structure required the least detection time and achieved a higher detection accuracy compared with existing relevant methods. Liu *et al.* [1] fused regression models and classification models with a large-scale deep convolutional neural network to perform an apparent age estimation and the experimental results showed that this approach obtained a state-of-the-art performance. Gürpinar *et al.* [28] combined kernel ELM with CNN to estimate the age of a person based on their face and achieved good results. Abdulnabi *et al.* [29] used a multitask CNN model to extract features corresponding to attributes in images and SVM models to perform attribute prediction. The method was shown to be effective for two popular attribute datasets. Zhang *et al.* [30] proposed a novel task-constrained deep model for face alignment, in which a deep model is used to extract a high-level representation and both landmark detection and attribution are learned using generalized linear models. A good performance was achieved compared with existing face alignment methods. Han *et al.* [31] presented a deep multitask learning (DMTL) method that can be used to jointly estimate multiple heterogeneous attributes from a single face image. First a face image is projected onto a high-level representation via a deep network and then refined by using shallow subnetworks for individual attribute estimation tasks. Extensive experimental results conveyed the superior performance of the DMTL method compared with the state of the art. Hand and Chellappa [32] proposed a multitask deep convolutional neural network (MCNN) with an auxiliary network (AUX) for facial attribute classification in which, AUX utilizes all attribute scores from a trained MCNN to capture attribute relationships at the score level. The approach reduces the number of parameters in the network and reduces the training time. Based on the above results, we decided to integrate CNN with other classifiers to improve the classification accuracy. In Sections IV, we will present the ensemble CNN2ELM in detail and show that it has better performance compared with other methods for processing the same tasks.

B. Fusion Methods

Fusion is a popular technology in biometrics, and it is most commonly used to fuse decisions or features in a hierarchical learning system. One of the most successful examples is the ensemble system [20]. This system proves that several classifiers with similar training characteristics have different generalization performances, and fusing these classifiers may or may not lead to a better performance than that of the best classifier in this ensemble system while reducing the global risk of making a poor decision. Recently, in the field of computer vision, many researchers have begun to study ensemble systems as these hybrid structures perform well. Malli *et al.* [33] proposed an ensemble CNN structure, that combined the outputs of learning models and used facial features to perform an apparent age estimation, they obtained a 0.3668 error in the final ChaLearn LAP 2016 dataset [34]. Rothe *et al.* [35] designed Deep EXpectation (DEX) of apparent age, which detected the facial images first, and then extracted CNN predictions from an ensemble network. The DEX won 1*st* place in the ChaLearn LAP 2015 challenge for apparent age estimation. Liu *et al.* [1], 2*nd* place, also proposed using ensemble CNNs that were based on the GoogleNet architecture [36]. Liu *et al.* [4] used a grouping estimation fusion (GEF) system to perform human age estimation. By fusing diverse decisions, GEF obtained better results, which reduced the overall risk of making a poor decision.

C. Age Grouping

In recent years, facial age grouping has been widely used for age estimation and many methods have been proposed. Kwon and da Vitoria Lobo [5] first introduced the age grouping method for age estimation, which starts by categorizing images into three age groups: infants, youth, and seniors. During the experiments, 47 images were used to verify the performance of the proposed method, and the classification accuracy for the infant group was below 68%. Horng *et al.* [37] presented a multistage learning system (*i.e.*, primary components detection, feature extraction, and age classification) for age estimation, and facial images were classified into four age groups: infants, youth, middle-aged, and seniors. Two-hundred-thirty facial images were used to test its performance, and the accuracy of the classification rate was 81.58%. By extracting geometric features from facial images and fusing the results from five classifiers, Thukral *et al.* [38] obtained an accuracy of 70.04% for their age groups (*i.e.*., 0-15, 15-30, and 30+). Gunay and Nabiyev [39] presented an automatic age estimation system based on local binary patterns (LBP) [10].

Facial images were divided into small regions, and spatial LBP histograms classified the regions into six age groups: 10 ± 5, 20 ± 5, 30 ± 5, 40 ± 5, 50 ± 5, 60 ± 5. At the same time, nearest neighbor, minimum distance, and k-nearest neighbor were used as the final classifiers and led to a classification accuracy of 80%. Hajizadeh and Ebrahimnezhad [40] used a probabilistic neural network (PNN) to classify images into one of four age groups, and the classification rate was 87.25%. Gürpinar *et al.* [28] proposed a kernel ELM with a CNN structure for age estimation, and with the ELM classifier, facial images were classified into 8 age groups. The age decision was performed by an ELM regressor, and a 0.3740 normal score was obtained for ChaLearn Looking at People 2016 - Apparent Age Estimation challenge dataset.

D. Age Estimation

Recently, age and gender classification has received a large amount of attention, because it provides a direct and quick method for obtaining implicit and critical social information [41]. Fu *et al.* [42] performed a detailed investigation of age classification, and we can learn more about this subject in [43]. Classifying age from human facial images was first introduced by Kwon and da Vitoria Lobo [44], who presented that calculating ratios and detecting the appearance of wrinkles could classify facial features into different age categories. After that study, the same method was used to model craniofacial growth using both psychophysical evidences and anthropometric evidence [45]. This approach demanded accurate localization of facial features.

Geng *et al.* [46] proposed a subspace method called AGing pattErn Subspace, which was used to estimate age automatically, while an age manifold learning scheme was presented in [47] to extract face aging features and a locally adjusted robust regressor was designed to predict the age of humans. Although these methods have many advantages, the requirement that input images are near-frontal and well-aligned is their weakness. It is not difficult to see that the datasets in the above experiments are constrained, so these approaches are not suitable for many practical applications, including unconstrained image tasks.

Last year, many methods were proposed to classify age and gender. Chang and Chen [48] introduced a cost-sensitive ordinal hyperplanes ranking method to estimate human age from facial images, while a novel multistage learning system called the "grouping estimation fusion" (DEF) was proposed to classify human age. Li *et al.* [49] estimated age using a novel feature selection method and showed the advantage of the proposed algorithm through experiments. Although these methods mentioned above have shown numerous advantages, they still rely on constrained images datasets, such as FG-NET [50], MORPH [51], and FACES [52].

All of these methods mentioned above have been verified effectively for age classification while they do not take a full consideration of gender and race factors for age prediction and do not make full use of decision fusion and age grouping methods. Our proposed method not only considers race and gender as factors but also designs a hierarchical structure to estimate age.

III. Preliminary Information

A. Deep Convolutional Neural Networks

Convolutional Neural Networks [53], which usually include an input layer, multi-hidden layers, and an output layer, are deep, supervised learning architectures and are often composed of two parts: an automatic feature extractor and a trainable classifier. CNNs have shown remarkable performance on visual recognition [54]. When we use CNNs to process visual tasks, they first extract local features from the input images. To obtain higher order features, subsequent layers of CNNs will combine these features. Next, these feature maps are finally encoded into 1-D vectors, and a trainable classifier valuates the vectors. Because of the need to consider the size, slant, and position variations of images, feature extraction is a key step during classification. Therefore, with the purpose of ensuring some degree of shift, scale, and distortion invariance, CNNs offer local receptive fields, shared weights, and downsampling. The purpose of training CNNs is to adjust all of the parameters of the system, *i.e.*, the weights and biases of the convolution kernel after which we will use the the well-tuned CNNs to predict the classes, such as label, age, and so on, from unknown input image datasets.

B. Short Review of the Extreme Learning Machine

Huang *et al.* [19] first proposed ELM for single hidden layer feedforward neural networks (SLFNs). Their approach was extended to the *generalized* SLFNS, and its hidden layer is not requires to be neuron-alike [55]. ELM first maps the input data from d-dimensional space into the L-dimensional hidden layer random feature space (also called ELM feature mapping) and then through ELM learning, the system achieves the output results. ELM can achieve better generalization performance than the other conventional learning algorithms at an extremely fast learning speed. Moreover, ELM is less sensitive to user-specified parameters and can be deployed faster and more conveniently [56], [57].

1) ELM Feature Mapping: The output function of the ELM network structure for generalized SLFNs is the following:

$$f(\mathbf{x}) = \sum_{i=1}^{L} \beta_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})\beta, \tag{1}$$

where $\beta = [\beta_1, \cdots, \beta_L]^T$ denotes the output weights' vector between the hidden layer and the output layer with $m \geq 1$ output nodes, while $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \cdots, h_L(\mathbf{x})]$ is the output vector of the hidden layer, which is called ELM *nonlinear feature mapping*. Different activation functions can be used in different hidden neurons [58]. Especially in real applications $h_i(\mathbf{x})$ can be written as follows:

$$h_i(\mathbf{x}) = G(\mathbf{a}_i, b_i, \mathbf{x}), \quad \mathbf{a}_i \in \mathbf{R}^d, \ b_i \in R, \tag{2}$$

where $G(\mathbf{a}, b, \mathbf{x})$ denotes a nonlinear piecewise continuous function, and Table I shows the commonly used activation functions. Here, $(\mathbf{a}_i, b_i)$ expresses the jth hidden node weight vectors and biases, respectively. ELM trains a SLFN that includes two critical stages, and random feature mapping is the first stage. In this stage, by randomly initializing the hidden

TABLE I
COMMONLY USED MAPPING FUNCTIONS IN ELM

Sigmoid	$G(\mathbf{a},b,\mathbf{x}) = \frac{1}{1+\exp(-(\mathbf{a}\cdot\mathbf{x}+b))}$
Hyperbolic tangent function	$G(\mathbf{a},b,\mathbf{x}) = \frac{1-\exp(-(\mathbf{a}\cdot\mathbf{x}+b))}{1+\exp(-(\mathbf{a}\cdot\mathbf{x}+b))}$
Gaussian function	$G(\mathbf{a},b,\mathbf{x}) = \exp(-b\|\|\mathbf{x}-\mathbf{a}\|\|)$
Multiquadric function	$G(\mathbf{a},b,\mathbf{x}) = (\|\|\mathbf{x}-\mathbf{a}\|\|+b^2)^{1/2}$
Hard limit function	$G(\mathbf{a},b,\mathbf{x}) = \begin{cases} 1, & if\ \mathbf{a}\cdot\mathbf{x}+b\le 0 \\ 0, & others \end{cases}$
Cosine function/Fourier basis	$G(\mathbf{a},b,\mathbf{x}) = \cos(\mathbf{a}\cdot\mathbf{x}+b)$

layer, **h(x)** maps the data from the d-dimensional input space into the L-dimensional hidden layer random feature space (which is also called the ELM *feature space*) [59]. Therefore, **h(x)** denotes a random feature mapping in essence, which is also called ELM *feature mapping*. ELM learning is the second stage which we will discuss next.

2) ELM Learning: In contrast to traditional feedforward neural network learning algorithms, without needing to adjust the hidden neural, the goal of ELM theory is not only to reach the smallest training error but also to achieve the smallest norm of the output weights [56], [55], [60], [61]. That goal can be written as follows:

$$\text{Minimize}: \ ||\beta||_\mu^{\sigma_1} + \lambda||\mathbf{H}\beta - \mathbf{T}||_\nu^{\sigma_2}, \tag{3}$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, and $\mu, \nu = 0, \frac{1}{2}, 1, \cdots, +\infty$. λ is a parameter that controls the trade-off between these two terms. $\mathbf{H}$ denotes the hidden layer output matrix, which can be denoted as follows:

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \vdots \\ \mathbf{h}(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} h(\mathbf{x}_1) & \cdots & h_L(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ h_1(\mathbf{x}_N) & \cdots & h_L(\mathbf{x}_N) \end{bmatrix}, \tag{4}$$

and Equation (5) expresses the training data target matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1{}^T \\ \vdots \\ \mathbf{t}_N{}^T \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{N1} & \cdots & t_{Nm} \end{bmatrix}. \tag{5}$$

There are many efficient methods for computing the output weights β, such as orthogonal projection methods, singular value decomposition (SVD), and iterative methods [59], while according to [56], [57], the optimization solution for ELM is $\sigma_1 = \sigma_2 = \mu = \nu = 2$, which has been proven to be more stable and have better generalization performance. Therefore, β can be written as follows:

$$\beta = \begin{cases} \mathbf{H}^T(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T}, & if\ N \le L \\ (\frac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T}, & if\ N > L \end{cases} \tag{6}$$

Theorem 2.1: Universal approximation capability [19]: For any nonconstant piecewise continuous function that is used as the activation function, if the parameters of the hidden neurons are tuned, then the function can make the SLFNs approximate any target continuous function $f(x)$. Then, according to any continuous distribution probability, the function sequence $\{h_i(\mathbf{x})\}_{i=1}^L$ can be randomly generated, and it has the universal approximation capability, which means that $\lim_{L\to\infty}||\sum_{i=1}^L \beta_i h_i(\mathbf{x}) - f(\mathbf{x})|| = 0$ holds with probability of one with appropriate output weights β.

Theorem 2.2: Classification capability [56]: For any nonconstant piecewise continuous function that is used as the activation function, if the parameters of the hidden neurons are tuned, the function could make the SLFNs approximate any target continuous function $f(\mathbf{x})$, and then, with the random hidden layer mapping **h(x)**, SLFNs can separate arbitrary disjoint regions of any shapes.

Therefore, ELM not only has universal approximation but also possesses classification. From the description above, the process of ELM can be described as follows. First, ELM randomly assigns hidden neuron parameters $(\mathbf{w}_i, b_i)$. Then, it calculates the hidden layer output matrix $\mathbf{H}$. Finally, we can calculate the output weight vector β.

Huang *et al.* [62] also proved that the resulting solution was stable, and the system had better performance when a positive value $1/\lambda$ was added to the diagonal of $\mathbf{H}^T\mathbf{H}$ or $\mathbf{H}\mathbf{H}^T$ in the calculation of the output weights β based on the ridge regression theory. When we use ELM to address large-scale dataset, it is easy to find $N \gg L$. Therefore, we can easily compute $\mathbf{H}^T\mathbf{H}$, because its size is much smaller than that of $\mathbf{H}\mathbf{H}^T$. The output weights β can be written as in Equation (7):

$$\beta = \left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T}. \tag{7}$$

Then, we can obtain the ELM output function:

$$f(\mathbf{x}) = \mathbf{h}(\mathbf{x})\beta = \mathbf{h}(\mathbf{x})\left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T}. \tag{8}$$

We use $\mathbf{U}$ to denote $\mathbf{H}^T\mathbf{H}$ and $\mathbf{V}$ to express $\mathbf{H}^T\mathbf{T}$, and thus, Equation (8) can be described as Equation (9):

$$f(\mathbf{x}) = h(\mathbf{x})\beta = h(\mathbf{x})\left(\frac{\mathbf{I}}{\lambda} + \mathbf{U}\right)^{-1}\mathbf{V}. \tag{9}$$

IV. ARCHITECTURE OF THE CNN2ELM MODEL

In this section, we present the design of our hybrid structure in detail. Fig. 1 presents the architecture of the CNN2ELM. As shown, our network includes three stages: feature fusion and enhancement, age grouping, and age decision. For race or gender grouping, the neural network mainly includes two stages: feature extraction and race or gender classification. The feature extraction stage contains a convolutional layer, contrast normalization layer, and max pooling layer. We also provide correlative parameters, such as the number of each filter type, size of each feature map, kernel size of each filter, and stride of each sliding window. For example, the first convolutional layer consists of 96 filters, its feature map size is 56×56, its kernel size is 7, and the stride of the sliding window is 4. For age grouping and age determination, we present the design of each part in the following sections.

A. The Design of Our Ensemble Structure

1) Convolutional Layer: In the convolutional layer, convolutions that are performed between the previous layer and a series of filters extract features from the input feature

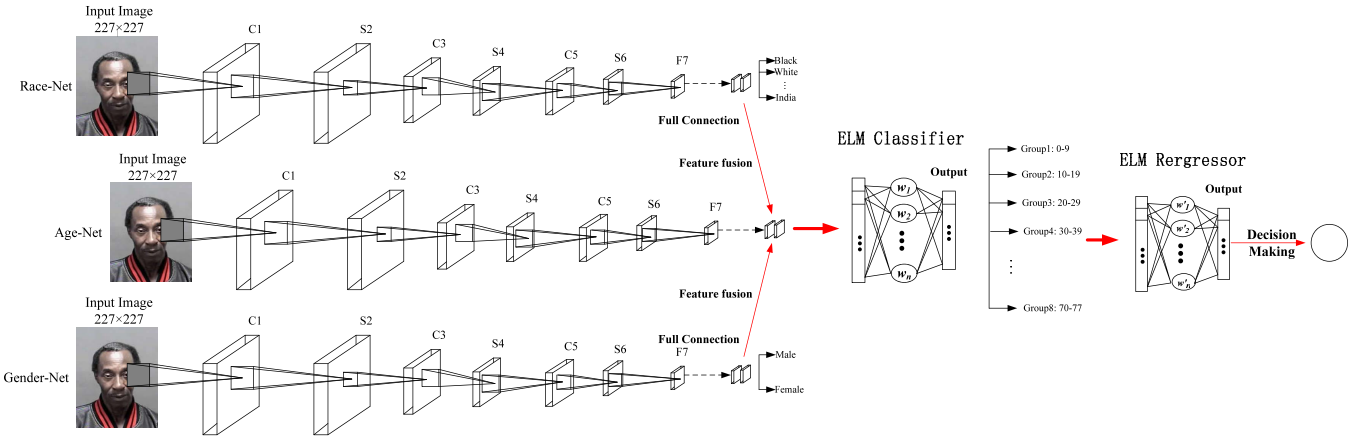


Fig. 1. Full schematic diagram of our network architecture. Race-Net, Age-Net, and Gender-Net are pretrained on the ImageNet dataset and then fine-tuned on the IMDB-WIKI dataset before fine-tuning on the training MORPH-II, Adience benchmark, and LAP-2016 datasets. During the validation and test processes, Race-Net, Age-Net, and Gender-Net are used to extract features from the same image. By fusing the features in the full connection layer, more discriminative and robust features are obtained. These features are classified into one of eight age groups, based on which the other ELM regressor performs the final age estimation.

maps [63], [64]. Next, the outputs of the convolutions add an additive bias and an element-wise nonlinear activation function is then applied to the front results. Without a loss of generality, we use the ReLU function as the nonlinear function in our experiment. In general, η_{ij}^{mn} denotes the value of a unit at position (m, n) in the jth feature map in the ith layer, and it can be expressed as Equation (10):

$$\eta_{ij}^{mn} = \sigma\left(b_{ij} + \sum_{\delta}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ij\,\delta}^{pq}\eta_{(i-1)\delta}^{(m+p)(n+q)}\right), \quad (10)$$

where b_{ij} represents the bias of this feature map, and δ indexes over the set of the feature maps in the $(i-1)$th layer, which are connected to this convolutional layer. $w_{ij\,\delta}^{pq}$ denotes the value at the position (p, q) of the kernel, which is connected to the kth feature map. The height and width of the filter kernel are P_i and Q_i, respectively.

The convolutional layer offers nonlinear mapping from the low-level representation of the images to a high-level semantic understanding. For convenience, during the later computations, Equation (10) can be simply denoted as follows:

$$\eta_j = \sigma\left(\sum w_{ij} \bigotimes \eta_{(i-1)}\right), \quad (11)$$

where $\bigotimes$ expresses the convolutional operation and w_{ij}, which is randomly initialized at first and then trained with a **BP** neural network [65], denotes the value of the ith layer in the jth feature map. $\eta_{(i-1)}$ is the outputs of the $(i-1)$ layer, and η_j is defined as the outputs of the jth feature map in the convolutional layer. Different sizes of input feature maps have various effects on the accuracy of the classification. A large of a feature map size indicates good features that are learned by the convolutional operations with the high cost of the computations, and a small size reduces the computation cost while degrading the accuracy of the classification. By a comprehensively considering the factors mentioned above and after many experiments, we set the size of the input feature map as 227 × 227, which is showed in Fig. 1.

2) Contrast Normalization Layer: The goal of the local contrast normalization layer is not only to enhance local competitions between one neuron and its neighbors but also to force the features of different feature maps in the same spatial location to be computed, which is motivated by computational neuroscience [65], [66]. To achieve the target, two normalization operations, *i.e.*, subtractive and divisive, are performed. η_{mnk} denotes the value of an unit at position (m, n) in the kth feature map. We have

$$z_{mnk} = \eta_{mnk} - \sum_{p=-\frac{P_i-1}{2}}^{\frac{P_i-1}{2}}\sum_{q=-\frac{Q_i-1}{2}}^{\frac{Q_i-1}{2}}\sum_{j=1}^{J_i}\varepsilon_{pq}\eta_{(m+p)(n+q)j}, \quad (12)$$

where ε_{pq} is a normalized Gaussian filter with the size of 7×7 at the first stage and 5×5 at the second stage. z_{mnk} not only represents the input of the divisive normalization operations but also denotes the output of the subtractive normalization operations. Equation (13) expresses the operator of the divisive normalization:

$$\eta_{mnk} = \frac{z_{mnk}}{\max\left(M, M(m, n)\right)}, \quad (13)$$

where

$$M(m,\ n) = \sqrt{\sum_{p=-\frac{P_i-1}{2}}^{\frac{P_i-1}{2}}\sum_{q=-\frac{Q_i-1}{2}}^{\frac{Q_i-1}{2}}\sum_{j=1}^{J_i}\varepsilon_{pq}\eta_{(m+p)(n+q)j}^2}, \quad (14)$$

and

$$M = \left(\sum_{m=1}^{s1}\sum_{n=1}^{s2} M(m,\ n)\right)/(s1 \times s2). \quad (15)$$

During the contrast normalization operations above, the Gaussian filter ε_{pq} is calculated with zero-padded edges, meaning that the size of the output of the contrast normalization operations is the same as its input.

TABLE II
THE DATASET SELECTED FROM **MORPH-II** FOR OUR EXPERIMENT

	Female	Male	Female and Male
Black	5,757	36,809	42,560
White	2,601	7,999	10,600
White and Black	8,358	44,802	53,160

3) Max Pooling Layer: Generally speaking, the purpose of the pooling strategy is to transform the joint feature representation into a novel, more useful representation that maintains crucial information while discarding irrelevant details. Each feature map in the subsampling layer is achieved by max pooling operations that are carried out on the corresponding feature map in convolutional layers. Equation (16) is the value of a unit at position (m, n) in the jth feature map in the ith layer or subsampling layer after the max pooling operation:

$$\eta_{ij}^{mn} = \max\{\eta_{(i-1)j}^{mn},\ \eta_{(i-1)j}^{(m+1)(n+1)}, \ldots, \eta_{(i-1)j}^{(m+P_i)(n+Q_i)}\}. \quad (16)$$

The max pooling operation generates position invariance over large local regions and downsamples the input feature maps. The number of feature maps in the subsampling layer is 96, the size of the filter is 3, and the stride of the sliding window is 2. The aim of the max pooling action is to detect the maximum response of the generated feature maps while reducing the resolution of the feature map. Moreover, the pooling operation also offers built-in invariance to small shifts and distortions. The procedures of the other convolutional layers and subsampling layers are the same as those of the layers mentioned above, except with different kernel sizes and strides.

B. Implementation Details

For the experiments, Race-Net, Age-Net, and Gender-Net are initialized with the weights obtained from training on the ImageNet dataset. Then they are further pretrained on the IMDB-WIKI dataset. These processes are the same as in [22]. Finally, the three models are fine-tuned using the MORPH-II, Adience benchmark, and LAP-2016 datasets. Because the fine-tuning processes using the latter three datasets are nearly the same, we present only the fine-tuning process of Race-Net, Age-Net, and Gender-Net on MORPH-II.

1) Fine-Tuning of CNN2ELM: Race-Net is fine-tuned using **MORPH-II** [51] with the race issue. **MORPH-II** has approximately 55,000 face images, of which approximately 77% of the images are Black faces, 19% are White faces, and the rest are faces of mainly Hispanics, Asians, and Indians [3]. Due to the unbalanced distribution of race and because a small-scale image dataset may bias the results, we utilize only the images of Black and White in our work. Table II presents the details of the dataset selected for our experiment.

We use the selected image dataset to train Race-Net. Race-Net models race estimation as an end-to-end deep classification problem, and Fig. 2 shows the process of Race-Net. The selected dataset is divided into 4:1:1, which means that 35,440 images are selected randomly as the training samples, 8,860 as the validation samples, and 8,860 as the testing samples.

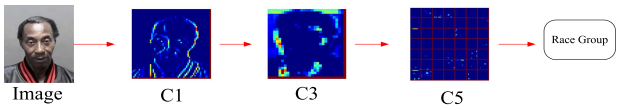


Fig. 2. The process of Race-Net. The color indicates the output of convolutional layers.

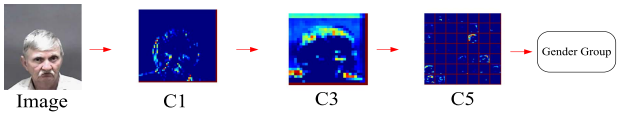


Fig. 3. The process of Gender-Net. The color indicates the output of convolutional layers.

For Race-Net, we set the base_lr as 0.01, gamma as 0.1, and momentum as 0.9. The weight_decay is set as 0.0005. The batch size is set as 50, and the total iterations is 50K. Training Race-Net requires nearly ten hours, while classifying a single image into a race takes approximately 600ms.

For Gender-Net, the selected dataset and settings of the network and parameters are the same as for Race-Net. Fig. 3 shows the process of Gender-Net. If face images include *e.g.*, much makeup, noise, obstructive lighting, and ambiguity, which increase the difficulty of estimating race or gender characteristics, subsystem (*i.e.*, "Race-Net + Age-Net + ELM classifier + ELM regressor (RAN)" or "Gender-Net + Age-Net + ELM classifier + ELM regressor (GAN)",) of CNN2ELM is invoked.

The training process for Age-Net is the same as that for Race-Net and Gender-Net when the target is human age. According to [3], for age estimation, crossing race and gender can cause a significant error. Thus, we first group the human face images into different race or gender groups. The goal of race grouping is to classify images into different race groups according to each person's race. Here, we select only the white and black racial datasets to train our proposed models (due to the unbalanced racial distribution) and train Race-Net via random 6-fold cross-validation. During the validation and test stages, Race-Net, Age-Net, and Gender-Net are used to extract features from the same image. By fusing these features, a more discriminative and enhanced feature set is achieved.

2) Face Attributes Classifier: We use the cross-entropy loss function as the Race-Net, Age-Net, and Gender-Net classifier, and the loss is:

$$L_i^k(\theta) = (y_i^k \log(p_i^k) + (1 - y_i^k)(1 - \log(p_i^k))). \quad (17)$$

where θ is the parameter of the network and p_i^k denotes the probability of the k_th attribute produced by our proposed network. We use y_i^k to denote the ground-truth of the k_th attribute.

The learning target is as follows:

$$\min_{\theta} \sum_{i=1}^{N_2} \sum_{k=1}^{N_1} L_i^k(\theta). \quad (18)$$

where N_1 and N_2 denote the number of the attributes and training examples, respectively.

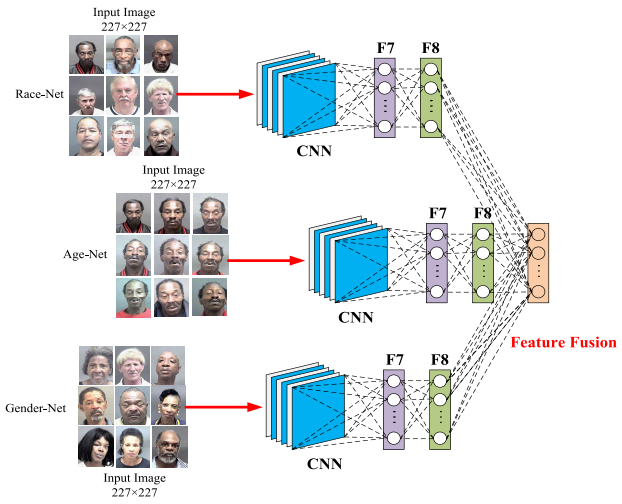


Fig. 4. The process of feature fusion. During the validation and test stage, Race-Net, Age-Net, and Gender-Net are used to extract features from the same image. By fusing these features, a more discriminative and enhanced feature set is achieved.

3) Feature Vector Fusion: When input features have more information, classifiers will achieve a better result. For example, a doctor may make a better diagnosis if he/she knows more information (such as body temperature, blood pressure, and blood glucose) about a patient. Therefore, we aggregate these features according to the ensemble principle and find that the new aggregated features are more discriminatory and robust, being conducive to classification. In other words, the method enhances the features and a better result may be achieved.

The purpose of our feature vector fusion is to obtain an enhanced feature vector that is beneficial to the ELM classifiers. Fig. 4 shows the process of feature fusion. Because classifiers learn more correlative and useful information, they can achieve a better generalization performance. We use the well-tuned CNN models to extract features from the input test dataset and fuse the feature vectors with the average combination rules. At this time, the batch size is 124 and we assume that the entire test batch set is $\aleph$. In each circulation, let $\phi(\mathbf{x}) \in R^n$ denote the feature vectors of the $F8$ layer and $\phi(\mathbf{x})_{race}$, $\phi(\mathbf{x})_{gender}$, and $\phi(\mathbf{x})_{age}$ denote the corresponding feature vectors. $\mathbf{x}$ denotes the input features, and ϕ denotes a series of operations, such as convolutions, ReLU, subsampling, or LRN. Fusion is often based on fixed combination rules such as product and average [67]; average is adopted in our fusion process. Algorithm 1 presents the feature fusion process.

4) ELM Classification: We use the ELM classifier and regressor for age estimation because of its fast learning speed and high accuracy. According to [28], the radial basis function (RBF) showed a superior performance over linear and polynomial kernels; and thus, we use it to calculate β through the original features. During the validation and test stage, the extracted features are first fused into more discriminative and robust set, after which ELM classifies these features into one of the whole age groups. Afterward, the final age decision is made by ELM regressors. **MORPH-II** divides the images into 7 groups: (10-19), (20-29), (30-39), (40-49), (50-59), (60-69), (70-77). When the race or gender properties are not easily grouped, our proposed substructure is used. For instance, if the race characteristic of the dataset cannot be assessed due to makeup, the gender property can be easy classified, so our proposed subsystem "GAN" will be used.

Algorithm 1 Feature Vector Fusion

Input:
The entire test batch set $\aleph$;
Test feature vectors: $\phi(\mathbf{x})_{race}$, $\phi(\mathbf{x})_{gender}$, and $\phi(\mathbf{x})_{age}$.
Output:
Enhanced feature vectors.
1: Start with the empty feature vector set $\phi_{average}$;
2: **for** i in $len(\aleph)/124$:
3: $\phi = average\ (\phi(\mathbf{x})_{race}, \phi(\mathbf{x})_{gender}, \phi(\mathbf{x})_{age})$;
4: $\phi_{average}.append(\phi)$;
5: **end for**
6: $\phi_{average} = \{\phi_1, \phi_2, \ldots, \phi_{len(\aleph)/124}\}$;
7: Return the enhanced feature vectors.

For each age group, our ELM classifier learns the whole binary classification models and β is calculated by a random 6-fold cross validation within the training set (race or gender feature vectors). Through this process, we optimize the ELM classifier model and obtain the final classifier.

5) ELM Regression: During the training stage of the ELM classifier, its outputs are the input of the ELM regressor. We also verify the regressor through the ELM classifier validation results.

C. Process of Our CNN2ELM

Undoubtedly, our ensemble structure must tune the parameters of convolutional structures from the learning process during the training stage before invoking the ELM. Race-Net, Age-Net, and Gender-Net are pretrained on the large ImageNet dataset and then fine-tuned using IMDB-WIKI. Afterward, we further pretrain the three networks on the MORPH-II, Adience benchmark, and LAP-2016 datasets. For every 1000 iterations, we verify the accuracy of the structure; *e.g.*, whether it has tuned the parameters and extracted discriminative features. The ELM classifier is invoked during the validation and testing stages. At that time, we first fuse features and compute the hidden layer weights of the ELM classifier. Afterward, the output of the ELM classifier during the training stage is treated as the input of the ELM regressor, and β is obtained. We use the CNN2ELM to estimate human age during the testing stage. The steps are summarized as follows:

Step 1: Race-Net, Age-Net, and Gender-Net are pretrained on ImageNet and then fine-tuned on IMDB-WIKI. The three networks are further pretrained on the MORPH-II, Adience benchmark, and LAP-2016 datasets.

Step 2: Race-Net, Age-Net, and Gender-Net are used to extract features from the same image during the validation and test stages, and a more discriminative and robust set is achieved via the fusion method.

Step 3: The hidden layer weight β matrixes of the ELM classifier are calculated based on these fusion features during the validation stage, and the ELM classifies test images into one of the age groups.

Step 4: The hidden layer weight β matrixes of the ELM regression are calculated based on the output of the ELM classifier, and the human face age is estimated.

During the experiments, although CNN2ELM generates better results compared with other algorithms when evaluating the same problems, we find that race and gender misclassifications still exist. For the ELM classifier and regressor, we find that our structure requires more memory because of caching the hidden layer weights and calculating the Moore-Penrose generalized inverse matrix $\mathbf{H}^{\dagger}$. Note that our ensemble structure estimates human facial age by taking full consideration of race, gender, and the age groups, which improves the accuracy of age estimation. In general, improving the number of hidden layer nodes of the ELM classifier or ELM regressor can improve the classification accuracy, but if the number of these exceed a specific scope (nearly 4,500 in our experiments), then the accuracy will be degraded because of the higher commutation cost, need for more memory, information losing, and aggravated overfitting.

V. Experiments

In this section, we use the MORPH-II, Adience benchmark, and LAP-2016 datasets to verify the performance of RAGN. Our proposed structure is implemented using the publicly available *cuda-convnet* [12] and Caffe [68] codes. The entire network in this paper is trained using an NVIDIA Tesla P100. First, we resize the input image to 256 $\times$ 256 pixels, and then, a 224 $\times$ 224 crop is selected from the center of the image or the four corners from the entire processed image. We also adopt different dropout measures to limit the risk of overfitting. For ELM classification and ELM regression, β is set as $\{10^{-5}, 10^{-4}, \ldots, 10^{2}, 10^{3}\}$ and L is set as $\{1800, 2000, \ldots, 4000\}$. To verify the advantage of CNN2ELM, an unconstrained dataset is also used to test its performance. Each experiment is conducted four times and we obtain the average of the relevant results.

A. Adience Benchmark

For the unconstrained dataset, we used the recently released Adience benchmark [43], [69] to test CNN2ELM. To this end, the benchmark of face photos is composed of images created from smart-phone devices. Because these images were uploaded to Flickr without prior manual filtering, they are highly unconstrained, meaning that they are representative of the challenges of real-world applications. Therefore, the images include variations in appearance, noise, pose, lighting and more, meaning that the photos are used without careful preparation or posing. It consisits of 19,487 images, in which, 8,192 are male and the rest are female. We obtained the public dataset from the Computer Vision Lab at the Open University of Israel (OUI) [70].

B. LAP-2016 Dataset

The ChaLearn Looking at People 2016 - Apparent Age Estimation challenge dataset [34] consists of 7,591 images that were labelled by several human annotators. The mean μ and the standard deviation σ are provided for each label sample. The dataset is divided as follows: 4,113 images for training, 1,500 for validation, and 1978 for testing.

The structure extracts the features from the fixed size images in the following way. First, the faces in the input images are detected using a DPM detector [71]. Because the faces in the LAP datasets are in unconstrained poses, we rotate the input images in the interval of $[-60°, 60°]$ by 5° as in [72], and by $-90°$, 90° and 180°. The highest detection score and rotation angle are achieved by the face box. Second, the face box size is enlarged by 40% in both width and height, and the face image is cropped. Finally, the image is reduced to 256 $\times$ 256 pixels, and then either a 224 $\times$ 224 crop is selected from the centre of the image or the four corners are selected from the entire processed image.

C. Evaluation Criteria

1) Mean Absolute Error (MAE): Averaging the absolute deviation of each sample's label from its corresponding estimated value is an effective way to measure the accuracy of the ELM regressor. In general, the mean absolute error (MAE) for a testing dataset can be described as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|, \tag{19}$$

where x_i denotes the true label, y_i is the predicted value, and N expresses the number of testing samples. The cumulative score (CS) is defined as:

$$CS(L) = (n_{e\leq L}/N) \times 100\% \tag{20}$$

where $n_{e\leq L}$ expresses the number of test images whose absolute error e of the age estimation is not larger than L years.

2) Normal Score (ϵ): Because the LAP-2016 is labelled by several annotators, the performance of an age estimation system might be measured more accurately by considering the variance of the annotations for each sample. Therefore, by fitting a normal distribution with mean μ and standard deviation σ of the annotations for each sample, the ϵ-score is calculated as follows:

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{21}$$

Therefore, the average ϵ-score for a dataset can range between 0 (best case) and 1(worst case).

D. Age Estimation

1) Age Estimation Based on the MORPH-II Database: In this section, we use MORPH-II to verify the performance of RAN, GAN, and RAGN and compare our structure with CNN-ELM [28], which includes identical convolutional layers. At the same time, to verify that race and gender characteristics play an important role in age estimation, we present the results of "Race-Net + ELM classifier + ELM regressor (RN)", "Gender-Net + ELM classifier + ELM regressor (GN)", and "Race-Net + Gender-Net + ELM classifier + ELM regressor (RGN)". We train our structure model using a mini-batch

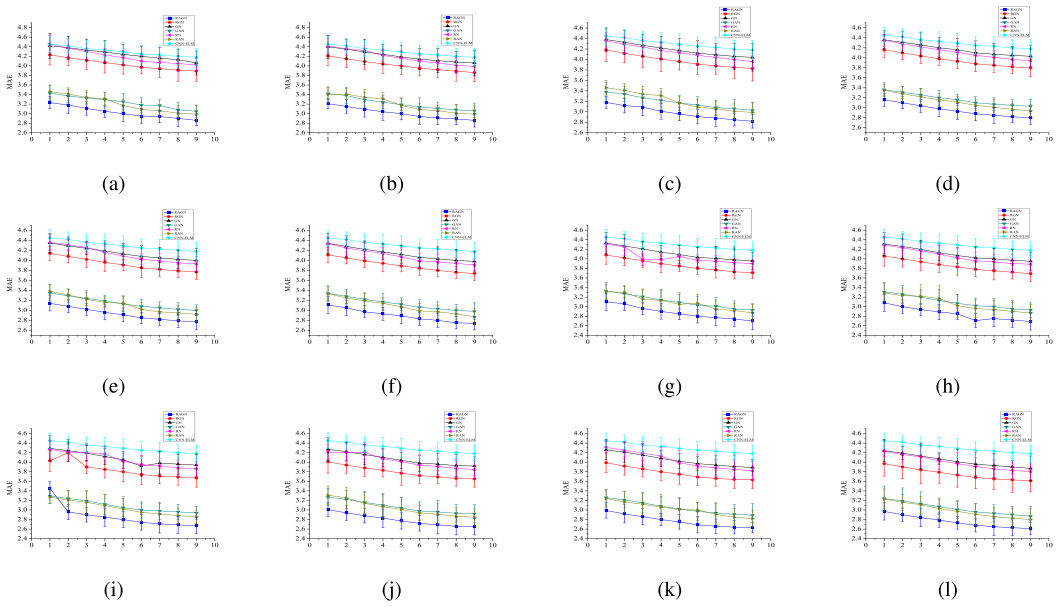


Fig. 5. Different MAEs of age estimation under different β and L values. β representation: $1 \rightarrow 10^{-5}$, $2 \rightarrow 10^{-4}$, $3 \rightarrow 10^{-3}$, $4 \rightarrow 10^{-2}$, $5 \rightarrow 10^{-1}$, $6 \rightarrow 10^{0}$, $7 \rightarrow 10^{1}$, $8 \rightarrow 10^{2}$, $9 \rightarrow 10^{3}$. (a) $L = 1800$. (b) $L = 2000$. (c) $L = 2200$. (d) $L = 2400$. (e) $L = 2600$. (f) $L = 2800$. (g) $L = 3000$. (h) $L = 3200$. (i) $L = 3400$. (j) $L = 3600$. (k) $L = 3800$. (l) $L = 4000$. Listed are the MAE $\pm$ standard error.

stochastic gradient descent of 0.7. During the fine-tuning of parameters in our hybrid structure, the learning rate at the beginning is set as 10^{-3} and is decreased to 10^{-4} after 20K iterations. After 45K iterations, the rate is set as 10^{-5}.

Fig. 5 shows the different MAEs of age estimation under different β and L values. As shown, the MAEs are affected by these conditions, and we find that the CNN-ELM, RN, GN, RAN, GAN, and RAGN seem to have convergence characteristics similar to that of the ELM. The performance of these structures tends to be quite stable under different hidden nodes when β is prefixed. However, when β increases, the MAEs decrease rapidly until converging at a better performance. Age-Net plays an important role in RAN and GAN. By fusing the features from Race-Net, Gender-Net, and Age-Net, higher discrimination features are achieved, and better results are obtained. The MAEs of RAN and GAN are much better than those of RN and GN. At the same time, the race and gender characteristics are nonnegligible factors in age estimation. With these characteristics, RAGN obtains the best performance. More importantly, in our experiments, we find that when β is in the range of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$, the MAEs quickly converge to better values, while they decrease slightly when β is in the range of $\{10^{1}, 10^{2}, 10^{3}\}$.

Although β and L have different influences on age estimation performance, the MAEs are reduced as β and L increase. At the same time, we can observe from these tables that the MAEs decrease slightly as β increases and that the performances of our proposed structures are better than those of RN, GN, and the CNN-ELM. For the same β and L, the MAEs of RAGN have the best performance, and RAN performs better than GAN.

The robustness and effectiveness of our proposed ensemble CNN2ELM structures are analysed in terms of the MAEs and CSs, and we compare the performance of RAGN with that of

TABLE III
MAEs OF DIFFERENT AGE ESTIMATION ALGORITHMS FOR THE MORPH-II DATABASE

Method	MAE
CSOHR [48]	3.82
DEX [22]	2.68
Best from [73]	2.78
Best from [74]	3.27
CNN+ELM [28]	4.03
RAGN [ours]	**2.61**

CSOHR [48], DEX [22], Hu *et al.* [73], and Niu *et al.* [74]. According to Fig. 5, β and L ultimately reach 10^{3} and 4000, respectively.

Table III shows that the listed age estimation methods acquired different MAEs and that our proposed RAGN achieved the best results. Although CSOHR exploits relative-order information among the age labels and aggregates a series of binary classification results to obtain the age rank, it does not consider the race and gender properties. DEX serves as an effective method for age estimation, the results of which may be attributed to the models it learns from large datasets, *e.g.*, from being pretrained on ImageNet and fine-tuned using IMDB-WIKI and MORPH-II. The method proposed in [73] achieves competitive results, as a novel learning scheme is employed to take advantage of the weakly labelled dataset obtained using the DCNN and a sufficient training dataset is obtained. Reference [74] proposed an end-to-end deep learning method that addresses the ordinal regression problem by transforming the problem into a series of binary classification subproblems. The method fully utilizes the correlation between these tasks but does not exploit the information (*e.g.*, gender and race properties) of human face images. The method in [28]

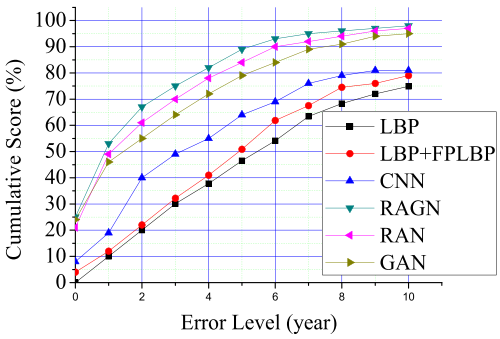


Fig. 6. The cumulative scores (CS) of age estimation.

fully utilizes the advantages of the CNN and ELM but does not consider the gender and race properties of human face images.

There are three main reasons for the satisfactory performance of our age estimation model. First, our Gender-Net, Age-Net and Race-Net are pretrained on ImageNet and fine-tuned on the IMDB-WIKI dataset before being fine-tuned on the MORPH-II dataset. Second, according to the above analysis, age estimation is significantly affected by race and gender. For different targets (race, gender, and age), different kinds of features are extracted by Age-Net, Gender-Net, and Race-Net, and more robust and discriminative features are achieved by fusing these features. Third, as stated in our motivation, age estimation based on a narrow age range may lead to better results. Age grouping is done using the ELM classifier, and our proposed structure makes the most of the enhancement features by classifying them into one group before performing age estimation. This process improves the performance of age estimation and gives our proposed structure lower MAEs than those of the compared algorithms.

CS is presented to show the performance of age estimation in Fig. 6. It is apparent that our proposed RAN, GNA, and RAGN methods outperform other state-of-the-art algorithms by a significant margin. Furthermore, we observe Fig. 6 that our proposed RGAN structure can reach a high accuracy, verifying that using race and gender properties can obtain a better result of age estimation.

2) Age Estimation Based on the Adience Benchmark: To verify the performance of our proposed system, we use an unconstrained dataset to test its performance. Because the images do not have prior manual filtering, it is difficult to distinguish their racial categories. Without a loss of generality, we present our results for age classification while mixing our structure with a dropout layer between the convolutional and classification layers. It is clear that the dropout structure can limit the risk of overfitting. We set the dropout ratio to 0.5 (50% probability to set the output value of a neural as 0). Each experiment was performed more than ten times, after which we obtained the corresponding averages. Table IV presents the accuracy of RAGN on the Adience Benchmark under different L or β.

According to Table IV, as L or β increase, the accuracy increases. Moreover, the performance of RAGN is more sensitive to β than to L. The accuracies increase rapidly when β changes in the range of $\{10^{-5}, 10^{-4}, \ldots, 10^{0}\}$, while the accuracies change slightly when β is 10^{1}, 10^{2}, or 10^{3}. In Table V, we also compare the performance of our proposed method with that of some of the latest methods when β and L are 10^{3} and 4000, respectively. Our proposed RAGN obtains the highest accuracy among all compared methods. The method in [43] was the first method able to predict age based on the Adience benchmark using a simple CNN; however, it ignores the gender and race properties in human face images. According to our analysis, DEX serves as an effective method for age estimation, the results of which may be attributed to the models it learns from large datasets, but it still does not consider the gender and race characteristics. The method in [28] combines the CNN with the ELM to estimate age but does not fully consider the gender and race properties of human face images. Our CNN2ELM not only fully utilizes the CNN and ELM but also exploits the gender and race characteristics in human face images. These experimental results show that age estimation is affected by gender and that a better result is obtained if the gender or race properties of testing images are known before performing age estimation.

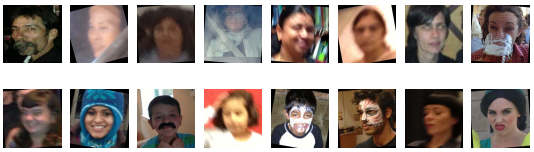


Fig. 7. Age misclassification.

Undoubtedly, our proposed algorithm did result in some cases of misclassification. Some misclassification results are shown in Fig. 7. The subjects in the top row (who were relatively older in age) were mistakenly classified as being younger, while the opposite occurred for the subjects in the bottom row. The unconstrain face images used in our experiments are the main source of this misclassification, and we can learn from Fig. 7 that most notable mistakes are caused by blurring or low resolution and by the use of heavy makeup.

3) Age Estimation Based on the LAP-2016 Dataset: In this section, we show the results of the RAGN system. Our CNN2ELM is pretrained on ImageNet and then fine-tuned on IMDB-WIKI and MORPH-II. In Table VI, we present the classification accuracy and recall for the 8 overlapping age groups. We show the performance of the entire system on the validation set of the LAP-2016 dataset in the final row.

According to Table VI, the system achieves lower classification accuracy for people in the age range of 20 to 40. During the regression process, the younger age groups yield smaller MAEs, as the age group variance increases as age progresses, which makes the apparent age estimation task harder. At the same time, their ϵ-scores behaved in an almost opposite manner because younger subjects are usually annotated with less variance.

Some validation sample estimation results are shown in Figs. 8 and 9. Fig. 8 displays the invariance of CNN features to common difficulties, such as pose, occlusions, and blur. Fig. 9

TABLE IV

ACCURACY OF RAGN ON ADIENCE BENCHMARK UNDER DIFFERENT CONDITIONS. THE VALUES OF THE MEAN ACCURACY ± STANDARD ERROR OVER ALL AGE CATEGORIES ARE LISTED. THE BEST RESULTS ARE IN BOLD

	1800		2000		2200		2400	
β	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM
10^{-5}	0.4916±0.0431	0.3379±0.0508	0.5043±0.0441	0.3491±0.0513	0.5157±0.0438	0.3597±0.0524	0.5289±0.0419	0.3690±0.0535
10^{-4}	0.5043±0.0397	0.3458±0.0513	0.5147±0.0437	0.3587±0.0521	0.5248±0.0441	0.3681±0.0527	0.5386±0.0424	0.3784±0.0527
10^{-3}	0.5137±0.0415	0.3547±0.0497	0.5184±0.0408	0.3661±0.0499	0.5309±0.0429	0.3762±0.0532	0.5462±0.0415	0.3879±0.0531
10^{-2}	0.5211±0.0426	0.3629±0.0517	0.5243±0.0415	0.3752±0.0517	0.5399±0.0435	0.3845±0.0529	0.5541±0.0427	0.3982±0.0523
10^{-1}	0.5284±0.0418	0.3719±0.0489	0.5363±0.0423	0.3839±0.0518	0.5481±0.0421	0.3929±0.0531	0.5624±0.0416	0.4067±0.0519
10^{0}	0.5367±0.0409	0.3792±0.0523	0.5434±0.0419	0.3915±0.0509	0.5562±0.0436	0.4037±0.0526	0.5694±0.0409	0.4158±0.0526
10^{1}	0.5413±0.0411	0.3862±0.0527	0.5503±0.0435	0.4038±0.0524	0.5609±0.0436	0.4097±0.0519	0.5729±0.0421	0.4213±0.0511
10^{2}	0.5459±0.0424	0.3953±0.0513	0.5582±0.0427	0.4083±0.0493	0.5653±0.0427	0.4127±0.0534	0.5788±0.0413	0.4270±0.0536
10^{3}	0.5501±0.0421	0.4037±0.0522	0.5631±0.0413	0.4118±0.0506	0.5723±0.0432	0.4203±0.0528	0.5835±0.0428	0.4347±0.051
	2600		2800		3000		3200	
β	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM
10^{-5}	0.5385±0.0417	0.3789±0.0574	0.5496±0.0421	0.3881±0.0556	0.5586±0.0455	0.3975±0.0563	0.5664±0.0473	0.4058±0.0587
10^{-4}	0.5461±0.0421	0.3867±0.0569	0.5579±0.0417	0.3975±0.0564	0.5674±0.0457	0.4089±0.0559	0.5762±0.0465	0.4171±0.0581
10^{-3}	0.5547±0.0416	0.3959±0.0572	0.5664±0.0424	0.4058±0.0563	0.5769±0.0463	0.4202±0.0567	0.5854±0.0464	0.4284±0.0573
10^{-2}	0.5624±0.0413	0.4052±0.0577	0.5742±0.0413	0.4163±0.0557	0.5842±0.0451	0.4315±0.0565	0.5942±0.0471	0.4397±0.0579
10^{-1}	0.5699±0.0425	0.4149±0.0571	0.5803±0.0419	0.4247±0.0569	0.5924±0.0467	0.4427±0.0571	0.6031±0.0469	0.4506±0.0582
10^{0}	0.5773±0.0419	0.4231±0.0563	0.5879±0.0433	0.4339±0.0571	0.6009±0.0465	0.4537±0.0569	0.6119±0.0467	0.4615±0.0577
10^{1}	0.5823±0.0422	0.4296±0.0566	0.5927±0.0415	0.4397±0.0582	0.6057±0.0471	0.4582±0.0573	0.6172±0.0466	0.4658±0.0584
10^{2}	0.5872±0.0426	0.4361±0.0561	0.5961±0.0414	0.4451±0.0571	0.6097±0.0468	0.4627±0.0577	0.6203± 0.0474	0.4701±0.0576
10^{3}	0.5911±0.0427	0.4428±0.0578	0.6007±0.0426	0.4539±0.0569	0.6124±0.0459	0.4673±0.0566	0.6229±0.0461	0.4751±0.0583
	3400		3600		3800		4000	
β	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM	RAGN	CNN-ELM
10^{-5}	0.5711±0.0467	0.4131±0.0581	0.5789±0.0481	0.4197±0.0592	0.5842±0.0497	0.4249±0.0611	0.5891±0.0515	0.4291±0.0624
10^{-4}	0.5799±0.0471	0.4249±0.0587	0.5892±0.0487	0.4301±0.0599	0.5962±0.0504	0.4379±0.0604	0.6002±0.0517	0.4423±0.0619
10^{-3}	0.5894±0.0488	0.4374±0.0573	0.5981±0.0492	0.4397±0.0601	0.6049±0.0503	0.4501±0.0609	0.6133±0.0509	0.4559±0.0622
10^{-2}	0.5994±0.0469	0.4487±0.0569	0.6072±0.0485	0.4489±0.0604	0.6157±0.0491	0.4627±0.0607	0.6251±0.0519	0.4681±0.0625
10^{-1}	0.6089±0.0465	0.4603±0.0591	0.6159±0.0479	0.4592±0.0597	0.6271±0.0506	0.4753±0.0615	0.6361±0.0521	0.4813±0.0631
10^{0}	0.6189±0.0482	0.4721±0.0588	0.6263±0.0496	0.4691±0.0593	0.6386±0.0511	0.4881±0.0613	0.6492±0.0513	0.4945±0.0629
10^{1}	0.6241±0.0477	0.4769±0.0579	0.6308±0.0488	0.4775±0.0595	0.6425±0.0509	0.4936±0.0599	0.6536±0.0511	0.5011±0.0624
. 10^{2}	0.6289±0.0484	0.4816±0.0582	0.6367±0.0491	0.4867±0.0604	0.6467±0.0513	0.4987±0.0606	0.6599±0.0523	0.5073±0.0633
10^{3}	0.6329±0.0481	0.4872±0.0589	0.6409±0.0483	0.4953 ±0.0591	0.6518±0.0505	0.5048±0.0614	**0.6649±0.0508**	**0.5127±0.0604**

TABLE V

CLASSIFICATION ACCURACY, RECALL AND REGRESSION PERFORMANCE FOR VALIDATION SET WITH DIFFERENT AGE GROUPS

Method	Average Accuracy
Best from [43]	0.5071 ± 0.051
CNN+ELM [28]	0.5127 ± 0.0497
DEX [22]	0.64 ± 0.042
RAGN [ours]	**0.6649 ± 0.0508**

TABLE VI

CLASSIFICATION ACCURACY, RECALL AND REGRESSION PERFORMANCE FOR VALIDATION SET WITH DIFFERENT AGE GROUPS. N DENOTES THE NUMBER OF SAMPLES. LISTED ARE THE MEAN ACCURACY ± STANDARD ERROR AND MAE ± STANDARD ERROR

Group	N_tr	N_val	Acc.	Rec.	ϵ	MAE
0-15	860	152	0.9613±0.0786	0.8018	0.4389	2.39±0.0467
10-25	2366	436	0.8634±0.0669	0.6937	0.3015	2.81±0.0647
15-30	3686	662	0.8722±0.0591	0.8981	0.3024	3.09±0.0571
20-35	4072	705	0.8307±0.0717	0.9143	0.3208	3.41±0.0347
30-40	1764	311	0.8273±0.0901	0.4115	0.3323	3.72±0.0463
35-50	1568	288	0.8814±0.0814	0.5071	0.3347	4.13±0.0293
45-60	976	184	0.9413±0.0902	0.5129	0.2914	3.76±0.0399
55-∞	554	106	0.9713±0.0621	0.6275	0.2783	4.23±0.0735
Overrall	8032	1462	-	-	0.3250	3.67

presents failed age estimations due to many possible reasons, such as insufficient number of samples to model, alignment errors, and face misdetection.

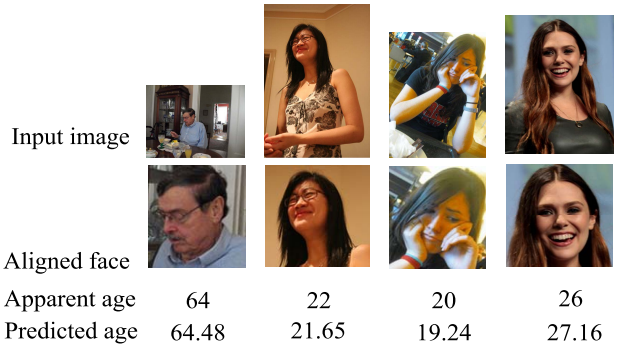


Fig. 8. Examples of satisfactory estimations from the validation set.

Our CNN2ELM model obtains 0.325 ϵ-score in the development phase and 0.3679 ϵ-score in the test phase of the challenge. Table VII shows the final results of the challenge.

E. Ablation Study

1) Investigation on Feature Enhancement Schemes: As shown in Fig. 1, the baseline deep model is the CNN, and three CNNs with different targets (race, gender, and age) are used in our structure. As discussed earlier, age estimation is easily affected by race and gender characteristics, and Age-Net,

Fig. 9. Examples of poor estimations from the validation set.

TABLE VII

CHALEARN LOOKING AT PEOPLE 2016 APPARENT AGE ESTIMATION CHALLENGE FINAL RESULTS

Position	Team	Test error
1	OrangeLabs	0.2411
2	palm_seu	0.3214
3	cmp+ETH	0.3361
4	WYU_CVL	0.3405
5	ITU_SiMiT	0.3668
6	RAGN (Ours)	0.3679
7	Bogazici	0.3740
8	MIPAL_SNU	0.4569
9	DeepAge	0.4573

TABLE VIII

ABLATION STUDY FEATURE ENHANCEMENT SCHEMES ON DIFFERENT DATASETS. LISTED ARE THE FINAL AVERAGE MAES FOR MORPH-II, THE FINAL AVERAGE PREDICTION ACCURACY FOR THE ADIENCE BENCHMARK, AND THE FINAL ϵ-SCORE FOR THE LAP-2016 DATASET

Datasets	RAN	GAN	RAGN	Age-Net(CNN-ELM)
MORPH-II	2.8	2.87	2.61	4.03
Adience Benchmark	0.6138	0.6014	0.6649	0.5127
LAP-2016 Dataset	0.4175	0.4391	0.3679	0.6143

Race-Net, and Gender-Net are used to extract different kinds of features. By fusing these features, more discriminative feature sets are achieved. The three networks are investigated in Table VIII. We find that by fusing the features obtained using Race-Net, Age-Net, and Gender-Net, our RAGN achieves the best performance on the MORPH-II, Adience benchmark, and LAP-2016 datasets. At the same time, RAN or GAN fuses features from only Race-Net and Age-Net or Gender-Net and Age-Net, achieving a better performance than that of the CNN-ELM but failing short compared to that of RAGN.

2) Investigation on Baseline CNN and ELM Models: Combining the CNN and ELM for age estimation is one key technique employed in our work. The CNN is used to extract features, and then the ELM classifies these into different age groups. Age is estimated using the ELM regressor model. In our extensive experiments, the age predicted using just the CNN on MORPH-II was 4.47, while that of the CNN (with an identical layer for feature extraction) + ELM (for age estimation) was 4.03; the result of our RAGN was 2.61. At the same time, the accuracy of the CNN used to estimate age on the Adience Benchmark was 0.5037, while the accuracy of the CNN+ELM was 0.5127, and that of our RAGN was 0.5349. We can conclude that the gain of the CNN for age estimation is nearly more than 80%, while that of the ELM is nearly 20%.

VI. CONCLUSION AND FUTURE WORK

In this paper, an ensemble learning framework called CNN2ELM is proposed for age estimation. The main innovations are the features enhancement, age grouping, and combination of a CNN with an ELM classifier and ELM regressor. Extensive experiments conducted on MORPH-II, the Adience benchmark, and the LAP-2016 dataset demonstrate the effectiveness of our proposed system. The experimental results show that age estimation is easily affected by gender and race and that fusing the features extracted from Age-Net, Race-Net, and Gender-Net before performing age estimation improves the performance. In the future, we will improve the ensemble systems and use them to process images that have varied facial poses, such as turned or tilted faces.

REFERENCES

[1] X. Liu *et al.*, "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 258–266.

[2] D. Yi, Z. Lei, and S. Z. Li, *Age Estimation by Multi-scale Convolutional Network*. Cham, Switzerland: Springer, 2015, pp. 144–158. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16811-1_10

[3] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Proc. IEEE Comput. Soc. Conf. (CVPRW)*, Jun. 2010, pp. 71–78.

[4] K.-H. Liu, S. Yan, and C.-C. J. Kuo, "Age estimation via grouping and decision fusion," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2408–2423, Nov. 2015.

[5] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," *Comput. Vis. Image Understand.*, vol. 74, no. 1, pp. 1–21, 1999.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[7] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, USA, Oct. 2006, pp. 307–316.

[8] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[9] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[10] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[11] 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)*, Jun. 2009, pp. 112–119.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25. no. 2, pp. 1097–1105.

[13] 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.

[14] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Correction to 'automatic age estimation based on facial aging patterns,'" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, p. 368, Feb. 2008.

[15] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "A probabilistic fusion approach to human age prediction," in *Proc. IEEE Comput. Soc. Conf. (CVPR)*, Jun. 2008, pp. 1–6.

[16] 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)*, Apr. 2013, pp. 1–6.

[17] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR*, Jun. 2011, pp. 657–664.

[18] Y. Yang, Y. Wang, and X. Yuan, "Bidirectional extreme learning machine for regression problem and its learning effectiveness," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 9, pp. 1498–1505, Sep. 2012.

[19] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[20] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Sep. 2006.

[21] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[22] 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, doi: 10.1007/s11263-016-0940-3.

[23] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[24] X.-X. Niu and C. Y. Suen, "A novel hybrid CNN-SVM classifier for recognizing handwritten digits," *Pattern Recognit.*, vol. 45, no. 4, pp. 1318–1325, 2012.

[25] F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognit.*, vol. 48, no. 10, pp. 2983–2992, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315001582

[26] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2015.

[27] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1174–1185, Mar. 2015.

[28] F. Gürpinar, H. Kaya, H. Dibeklioglu, and A. A. Salah, "Kernel ELM and CNN based facial age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 785–791.

[29] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.

[30] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[31] H. Han, A. K. Jain, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2017.2738004.

[32] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Proc. AAAI*, 2017, pp. 4068–4074. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14749

[33] R. C. Malli, M. Aygün, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 714–721.

[34] S. Escalera *et al.*, "Chalearn looking at people and faces of the world: Face analysisworkshop and challenge 2016," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 706–713.

[35] 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. Workshop (ICCVW)*, Dec. 2015, pp. 252–257.

[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. (CVPR)*, Jun. 2015, pp. 1–9.

[37] W. B. Horng, C. P. Lee, and C. W. Chen, "Classification of age groups based on facial features," *Tamkang J. Sci. Eng.*, vol. 4, no. 4, pp. 183–192, 2001.

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

[39] A. Gunay and V. V. Nabiyev, "Automatic age classification with LBP," in *Proc. Int. Symp. Comput. Inf. Sci.*, 2008, pp. 1–4.

[40] M. A. Hajizadeh and H. Ebrahimnezhad, "Classification of age groups from facial image using histograms of oriented gradients," in *Proc. 7th Iranian Conf. Mach. Vis. Image Process.*, Nov. 2011, pp. 1–5.

[41] S. Fu, H. He, and Z. G. Hou, "Learning race from face: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2483–2509, Dec. 2014.

[42] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[43] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 34–42.

[44] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 762–767.

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

[46] 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.

[47] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[48] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 785–798, Mar. 2015.

[49] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu, and H. Lu, "Human age estimation based on locality and ordinal information," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2522–2534, Nov. 2015.

[50] A. Lanitis. (2002). *The Fg-Net Aging Database*. [Online]. Available: www-prima.inrialpes.fr/FGnet/html/benchmarks.html

[51] K. Ricanek, Jr., and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 341–345.

[52] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2547–2553.

[53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[54] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1610–1623, Oct. 2010.

[55] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.

[56] M. Duan, K. Li, X. Liao, and K. Li, "A parallel multiclassification algorithm for big data using an extreme learning machine," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2654357.

[57] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, 2017, doi: https://doi.org/10.1016/j.neucom.2017.08.062.

[58] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Netw.*, vol. 61, pp. 32–48, Jan. 2015.

[59] G.-B. Huang, Z. Bai, L. L. C. Kasun, and C. M. Vong, "Local receptive fields based extreme learning machine," *IEEE Comput. Intell. Mag.*, vol. 10, no. 2, pp. 18–29, May 2015.

[60] G.-B. Huang, "An insight into extreme learning machines: Random neurons, random features and kernels," *Cognit. Comput.*, vol. 6, no. 3, pp. 376–390, 2014.

[61] L. L. C. Kasun, Y. Yang, G.-B. Huang, and Z. Zhang, "Dimension reduction with extreme learning machine," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3906–3918, Aug. 2016.

[62] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[63] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

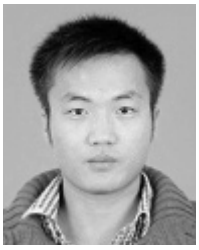
[64] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2247–2256, Aug. 2015.

[65] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2011, pp. 2809–2813.

[66] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Comput. Biol.*, vol. 4, no. 1, p. e27, 2008.

[67] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognit.*, vol. 34, no. 2, pp. 299–314, 2001.

[68] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093
[69] 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.
[70] *Open University of Israel*. Accessed: Dec. 15, 2014. [Online]. Available: http://www.openu.ac.il/home/hassner/Adience
[71] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 720–735.
[72] M. Uricár, R. Timofte, R. Rothe, J. Matas, and L. Van Gool, "Structured output SVM prediction of apparent age, gender and smile from deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun./Jul. 2016, pp. 730–738.
[73] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3087–3097, Jul. 2017.
[74] 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)*, Jun. 2016, pp. 4920–4928.



Mingxing Duan is currently pursuing the Ph.D. degree with the School of Computer Science, National University of Defense Technology, China. His research interests include big data and machine learning. He has authored several papers in international journals, such as the IEEE-TNNLS, *Neurcomputing*, CCPE, and has been a reviewer of several international journals, such as the IEEE-TNNLS, the IEEE-TMM, and the IEEE-SMC: Systems, JCSS, IJPRAI, and CCPE.

Kenli Li (SM'16) received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, China, in 2003. He was a Visiting Scholar with the University of Illinois at Urbana–Champaign from 2004 to 2005. He is currently a Full Professor of computer science and technology with Hunan University and the Deputy Director of the National Supercomputing Center, Changsha. His major research includes parallel computing, cloud computing, and Big Data computing. He has authored over 150 papers in international conferences and journals, such as the IEEE-TC, the IEEE-TPDS, and the IEEE-TSP. He is currently serving on the editorial boards of the IEEE TRANSACTIONS ON COMPUTERS and *International Journal of Pattern Recognition and Artificial Intelligence*. He is an Outstanding Member of CCF.

Keqin Li (F'14) is currently a SUNY Distinguished Professor of computer science. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU-GPU hybrid and cooperative computing, multi-core computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of things, and cyberphysical systems. He has authored over 520 journal articles, book chapters, and refereed conference papers. He has received several best paper awards. He is currently or has served on the editorial boards of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.