Efficient Group-n Encoding and Decoding for Facial Age Estimation

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

Abstract—Different ages are closely related especially among the adjacent ages because aging is a slow and extremely non-stationary process with much randomness. To explore the relationship between the real age and its adjacent ages, an age group-n encoding (AGEn) method is proposed in this paper. In our model, adjacent ages are grouped into the same group and each age corresponds to n groups. The ages grouped into the same group would be regarded as an independent class in the training stage. On this basis, the original age estimation problem can be transformed into a series of binary classification sub-problems. And a deep Convolutional Neural Networks (CNN) with multiple classifiers is designed to cope with such sub-problems. Later, a Local Age Decoding (LAD) strategy is further presented to accelerate the prediction process, which locally decodes the estimated age value from ordinal classifiers. Besides, to alleviate the imbalance data learning problem of each classifier, a penalty factor is inserted into the unified objective function to favor the minority class. To compare with state-of-the-art methods, we evaluate the proposed method on FG-NET, MORPH II, CACD and Chalearn LAP 2015 databases and it achieves the best performance.

Index Terms—Age estimation, deep learning, convolutional neural network, age grouping, data imbalance

1 Introduction

Human age estimation makes an important component in face attribute analysis [1], which has many applications in real-world, such as business intelligence, human computer interaction (HCI) and visual surveillance [2], [3], [4], [5]. However, human age is still hard to estimate precisely from a single face image even though the problem has been extensively studied for many years.

Facial aging process is filled with randomness and is not stationary for everyone. The randomness exists in many aspects, such as different diets, living or working environment, and most importantly, the various genes. All of those factors can more or less affect human aging and further leads to aging differences in the appearance. In real world, people at the same age may look differently, appearing slightly older or younger comparing to each other. On the other hand, faces from close ages look similar [6] because of the slow and gradual aging process. Sometimes it is hard to judge which one is older or younger between two faces from close ages. So, there is a strong correlation between age classes especially for adjacent ages.

Most previous methods estimated age by casting it as a classification problem [5], [7], [8], [9] or regression problem [10], [11], [12], [13], [14]. For age classification, each age class is assumed to be independent to one another, which ignores the relationship between different classes. In contrast, regression problem treats age as continuous value and employs regression methods to predict age based on extracted features, such as Partial Least Squares (PLS) [15], Canonical Correlation Analysis (CCA) [16], Support Vector Regression (SVR) [17]. However, those methods do not involve any aging information, either.

Due to the aging randomness in the aging process, there is an ambiguous mapping rather than exact mapping between face and its real age. This is particularly evident in senior people. We may say that a man looks like in his late thirties but can never be sure about his exact age just from his appearance. Thus, assigning each face with a single age label seems difficult because of the strong correlation among age classes especially among the adjacent classes. Furthermore, training with several adjacent ages together for age estimation may be more helpful than treating each age as an independent class.

Inspired by this, we group the face images within a specific age range and then regard each age group as an independent class in the training stage. Our age grouping method is inspired by [18] but with crucial differences. Unlike [18],

- *Z. Tan is with the Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Room 1402, Intelligent Building, 95 Zhongguancun Donglu, Haidian District, Beijing 100190, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: tanzichang2016@ia.ac.cn.*
- *J. Wan, Z. Lei, and S.Z. Li are with the Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Room 1402, Intelligent Building, 95 Zhongguancun Donglu, Haidian District, Beijing 100190, China. E-mail: {jun.wan, zlei, szli}@nlpr.ia.ac.cn.*
- *R. Zhi is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China. E-mail: zhirc@ustb.edu.cn.*
- *G. Guo is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506. E-mail: guodong.guo@mail.wvu.edu.*

Manuscript received 28 Feb. 2017; revised 4 Sept. 2017; accepted 17 Nov. 2017. Date of publication 3 Dec. 2017; date of current version 10 Oct. 2018. (Corresponding author: Jun Wan.)
Recommended for acceptance by S. Escalera, X. Baró, I. Guyon, H. J. Escalante, G. Tzimiropoulos, M. Valstar, M. Pantic, J. Cohn, and T. Kanade.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2779808

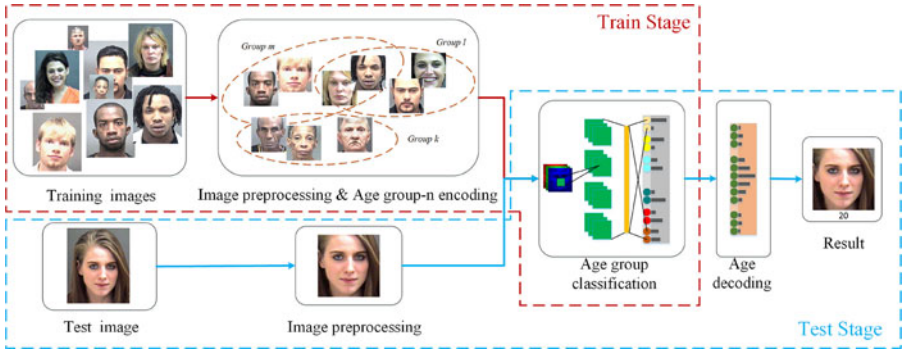


Fig. 1. The pipeline of our framework for age estimation. It consists of two stages: training stage and testing stage. In the training stage, the training images with different scales are first processed by face detection, alignment and cropping. All the images are aligned according to the point of the center of two eyes and the upper lip. Then all the training images are grouped by the age group-n encoding strategy, where the images from adjacent ages would grouped into the same group. After that, the training images are used to train the CNNs. In the testing stage, the test image is first processed in the same way as the training stage used. Then, the processed image is input into the trained CNN network, and age group classification is employed to obtain the probabilities of each group. Finally, the predicted age is obtained by decoding the group classification results.

which needs to group ages for many times and each time they divide ages into non-overlapped groups, our age grouping method conducts age division only one time, where all ages are divided into overlapped age groups. We carefully design the grouping strategy to encode ages into age groups, which ensures that each age corresponds to an unique age group set. Based on this, the exact age can be recovered by decoding the group classification results according to a certain mapping relation between the age and age groups. Therefore, our method could be implemented in a single network rather than an ensemble of networks [18] or cascaded networks [19].

Using the novel grouping method, we can transform the age estimation problem into a series of binary classification problems, where each classifier determines whether the face image belongs to the corresponding group or not. The CNN with multiple output layers is also employed in our approach. Unlike [20], [21], our method aims to explore the relationship between the adjacent ages based on age group classification, while the approaches of [20], [21] mainly exploit the relative order relation among age labels. Besides, each classifier of the network in [20], [21] acts as a comparator to determine whether or not the age of the input face is greater than a value, while each classifier of our network aims to distinguish images within each age group.

For each binary classifier, the number of training images belonging to the corresponding group is far less than the others (imbalanced data learning). This is because we group images only within a small age range. A viable solution to the imbalance data problem is to modify the algorithm via cost-sensitive learning [22], [23]. In this paper, we modify our training algorithm by employing a penalty factor to shift the bias of the classifier to favor the minority class, which increases the contributions of the minority class in the learning stage.

The proposed age estimation framework is shown in Fig. 1, and the source codes and models are available at the website.[1] The main contributions of our work include:

1) A novel age grouping strategy called Age Group-n Encoding (AGEn) is proposed, where the adjacent ages are grouped into the same group and each age corresponds to n groups. Moreover, unlike employing an ensemble of multiple networks to obtain the exact age due to grouping ages for many times [18], only a single network (see Fig. 1) is used to make the prediction with our age division.
2) To accelerate the predicting process, a Local Age Decoding (LAD) strategy is proposed to obtain the predicted age by locally decoding the outputs of the binary classifiers.
3) Inspired by previous works [22], [23], we extend the cost-sensitive learning strategy used in traditional methods (i.e., Cost-Sensitive Dataspace Weighting with Adaptive Boosting [23], Cost-Sensitive Decision Trees [23]) into our designed objective function of the proposed CNN framework for age estimation, which is effective to deal with the imbalanced data problem caused by age grouping.
4) Our method achieves the state-of-the-art results on multiple datasets, including FG-NET [24], MORPH II [25], CACD [26] and Chalearn LAP 2015 databases [27].

2 Related Work

Human Age estimation has been studied extensively for over 20 years. The earliest work of age estimation was possibly reported by Kwon et al. [29] in the 1990s, which judged the age range of face images with hand-crafted features, such as baby, young adult and senior adult. However, only dozens of face images were analyzed in their work. At that time, the lack of a large-scale age dataset also hindered the development of age estimation technology. With the joint efforts of many scholars from all over the world, large age datasets such as FG-NET [30], MORPH II [25] and CACD [26] databases are available for the community, which are also the most popular age datasets nowadays.

With the development of facial analysis technology, researchers started to predict the exact age rather than simply estimate the coarse age range from face images. Also, a large number of methods have been proposed for age estimation, such as Active Appearance Models (AAM) [31], AGing patErn Subspace (AGES) [7], [32], age manifold [10], [33], [34],

1. http://www.cbsr.ia.ac.cn/users/zctan/projects/AgeEncodingDecoding/main.htm

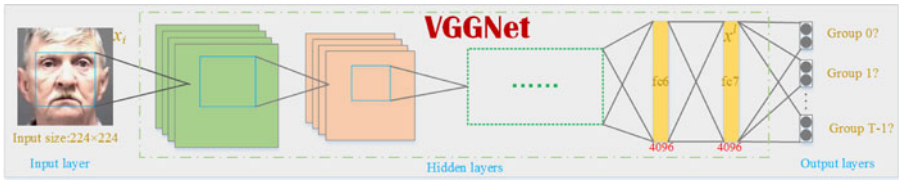


Fig. 2. The architecture of the proposed network. Our network is based on the VGG-16 network [28] and we adopt the BGR face image as the input with the size of 224×224. The CNN network consists of two fully connected layers and the later one produces a feature vector for age group classification. After that, the network branches out T output layers, where each layer is employed as an binary classifier that judges whether the input image belongs to the corresponding age group or not. Moreover, all the convolutional layers are followed by ReLU non-linearity.

and methods with local features [8], [35], [36]. Particularly, Biologically Inspired Features (BIF) [8] has the most outstanding ability in age estimation among those local features. After features extracted by local image descriptors, classification or regression methods would be employed to obtain the predicted age, such as BIF+SVM [8], BIF+SVR [8], BIF+CCA [12]. More recently, Geng et al. [6], [37] allowed each face image labeled with a label distribution rather than a single age label, where both the real age and its adjacent ages would contribute to the learning. The work in [38], [39] also integrate the idea of label distribution into deep learning framework and achieve promising performance.

Recently, deep learning has gained a lot of success on age estimation. Yi et al. [14] deployed many parallel CNNs with multi-scale face images for age estimation. Malli et al. [18] estimated apparent ages with age grouping to account for multiple labels per image. However, this work needs an ensemble of models to further predict the exact age, seeming relatively tedious. Antipov et al. [40] developed a children-specialized deep learning method for apparent age estimation, and achieved the best performance at Chalearn Looking At People (LAP) challenge 2016. Niu et al. [21] casted age estimation as an ordinal regression problem with a multiple outputs CNN, which achieved the state-of-the-art result on MORPH II database. Zhu et al. [19] first used age group classifier to acquire the coarse age range of face images with CNN, and then multiple local age estimators were employed to predict the exact age. Liu et al. [41] exploited a general-to-special transfer learning scheme for age estimation based on GoogleNet [42]. Rothe et al. [9] proposed a Deep EXpectation (DEX) method for apparent age estimation based on VGG-16 architecture [28] and won the first place at Chalearn LAP challenge 2015. However, DEX only conducts the refinement that fuses all ages information in the prediction phase but neglects the correlation between different ages in the training stage.

In this work, the correlation between adjacent ages would be explored through grouping and training the adjacent ages together. Different from previous grouping-based methods, which estimate the age for a facial image through an ensemble of models or cascaded structures, the proposed method estimates age from facial images with a single network based on well-designed group-n encoding and decoding processes. To our best knowledge, it is the first work to conduct age estimation with a single network based on age group classification.

3 Our Method

The pipeline of our method for age estimation is shown in Fig. 1, and our method mainly consists of fine-grained age grouping, age group classification and age decoding. The specific algorithm is given in Algorithm 1.

Algorithm 1. The Algorithm of the Proposed Method

Input: The training data $D = \{x_i, y_i\}_{i=0}^{N-1}$, and the test data $D' = \{x'_i\}_{i=0}^{M-1}$.

Output: The predictions $\{y'_i\}_{i=0}^{M-1}$ for the test data.

1: conduct age grouping for training data D with AGEn, and obtain the group labels $\left\{\{g_i^t\}_{t=0}^{T-1}\right\}_{i=0}^{N-1}$, the age group index $\mathcal{C}_a$ for each age a and the age set $\mathcal{S}_t$ for each group t.
2: train MO-CNN with $\{x_i, \{g_i^t\}_{t=0}^{T-1}\}_{i=0}^{N-1}$ for searching the optimal network parameters $\{\mathcal{W}, W\}$.
3: **for** $i = 0, 1, \ldots, M-1$ **do**
4: input the face image x'_i into MO-CNN
5: obtain $\{\{p(g_i^t = m | x'_i, \mathcal{W}, W)\}_{m=0}^{1}\}_{t=0}^{T-1}$
6: $m \leftarrow \arg\max_t p(g_i^t = m | x'_i, \mathcal{W}, W)$
7: **for** $a \in \mathcal{S}_m$ **do**
8: compute $\mathcal{P}(a | x'_i, \mathcal{W}, W)$ according to Eq. (7)
9: **end forend**
10: $y'_i \leftarrow \arg\max_{a \in \mathcal{S}_m} \mathcal{P}(a | x'_i, \mathcal{W}, W)$
11: **end forend**
12: **return** The predictions $\{y'_i\}_{i=0}^{M-1}$

3.1 Fine-Grained Age Grouping

Unlike previous age grouping methods where each age corresponds to one group, we introduce a novel age grouping method called Age Group-n Encoding for age estimation, where each face image is assigned to n groups. The grouping rules are given below:

1. Given the age set $\mathcal{Y} = \{l_0, l_1 \ldots, l_K\}$, we can group ages into T $(T = K + n)$ groups. Note that l_0 and l_K are the minimum and maximum ages, respectively, and $l_0 < l_1 < \cdots < l_K$.
2. For age l_i, it is assigned to group $i, i+1, \ldots, i+n-1$, where each age corresponds to n groups. Each group includes at least one age but at most n ages.

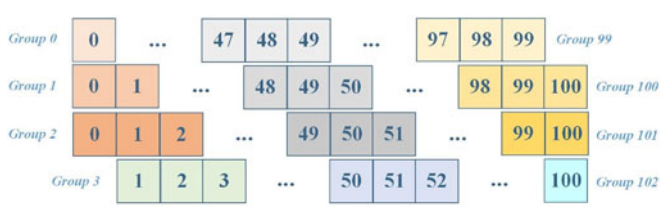


Fig. 3. Example of grouping results with Age Group-3 Encoding for age set $\{0, 1 \ldots, 100\}$. There are 103 groups in total and each age corresponds to 3 groups.

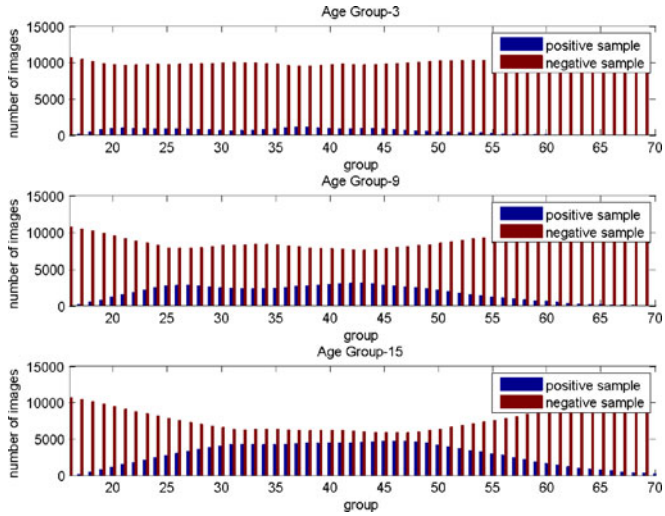


Fig. 4. The distribution of positive and negative samples for each age group on MORPH II training set with AGE3, AGE9 and AGE15. When grouped by AGE3, the distribution is extremely uneven and negative samples is many times larger than positive samples. The number of positive samples of middle groups would increase as n rises, but the imbalance is still serious in the marginal groups.

Fig. 3 gives a grouping example when $K = 100$ and $n = 3$ for age set $\{0, 1, \ldots, 100\}$. According to our grouping rules, each age is encoded into a unique group set, which is essential for the prediction stage that is a decoding process from group to age. In order to facilitate later parts of the paper, $\mathcal{C}_a = \{c_0, c_1, \ldots, c_{n-1}\}$ are used to denote the indices of the groups that age a belongs to. We also let $\mathcal{S}_t$ represents those ages that the tth group includes. For example, as shown in Fig. 3, $\mathcal{C}_1$=$\{1, 2, 3\}$ indicates that age 1 is assigned to group 1, 2 and 3 and $\mathcal{S}_3$=$\{1, 2, 3\}$ denotes that group 3 consists of ages 1, 2 and 3.

3.2 Age Group Classification

The network architecture of age group classification, called Multiple Outputs CNN (MO-CNN), is illustrated in Fig. 2. The network includes multiple output layers, where each output layer corresponds to a binary classification task that judges whether the input sample belongs to the age group or not. Assuming we have a training set with N samples, where each sample is attached with a chronological label and T age group labels where $T = K + n$ when the Age Group-n Encoding strategy is employed. Each sample is represented as $\left\{x_i, y_i, \{g_i^t\}_{i=0}^{T-1}\right\}$, where $x_i \in R^d$ is the ith sample, $y_i \in \mathcal{Y}$ represents the age label for x_i and $g_i^t \in \mathcal{G} = \{0, 1\}$ is the age group label indicating whether the ith sample belongs to age group t or not. If x_i belongs to age group t, $g_i^t = 1$; otherwise, $g_i^t = 0$. As shown in Fig. 2, the network extracts high level feature x_i^l through a sequence of non-linear mappings with a set of parameters $\mathcal{W} = \{\mathcal{W}_i\}_{i=0}^{l}$, where $\mathcal{W}_i$ represents the weights of layer i. With shared representation x_i^l, we conduct the group classifications via multiple binary classifiers with the parameters $W = \{W_t\}_{t=0}^{T-1}$, where W_t denotes the weights of tth classifier. Thus, the parameters of the whole network can be denoted as $\{\mathcal{W}, W\}$.

For each classifier, the cross-entropy loss is used as the loss function, thus the objective function of the tth classifier can be written as

$$J_t = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{m=0}^{1} 1\{g_i^t = m\}\log\left(p(g_i^t = m|x_i, \mathcal{W}, W)\right), \quad (1)$$

where $p(g_i^t = m|x_i, \mathcal{W}, W) = \frac{\exp\{(W_t^m)^T x_i^l\}}{\sum_j \exp\{(W_t^j)^T x_i^l\}}$ is softmax function and W_t^j denotes the jth column of the parameter matrix W_t for tth task.

However, the data distribution is extremely unbalanced for each classifier, and training unevenly could jeopardize the whole model. Each sample in a binary classifier has two states, belonging to the group (a positive sample) or not belonging to the group (a negative sample). As shown in Fig. 4, the number of positive samples is much less than the negative samples. To alleviate the imbalanced data learning problem, we impose penalty factors to penalize positive and negative samples at different degrees for each task. The penalty coefficients are represented as $\rho = \{\rho_t^0, \rho_t^1\}_{t=0}^{T-1}$, where ρ_t^0 is the penalty coefficient for negative samples and ρ_t^1 for positive samples. Thus the objective function of tth task is

$$J_t = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{m=0}^{1} 1\{g_i^t = m\}\rho_t^m\log\left(p(g_i^t = m|x_i, \mathcal{W}, W)\right). \quad (2)$$

Therefore, we can balance the contribution of positive and negative samples via adjusting the magnitude of the penalty coefficients.

We have T binary classification tasks all together and each task corresponds to an output layer. Let α_t denotes the importance level of the t-task, and the objective function of the whole CNN can then be written as

$$J = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{t=0}^{T-1}\sum_{m=0}^{1}\Big(\alpha_t 1\{g_i^t = m\}\rho_t^m \cdot \log\left(p(g_i^t = m|x_i, \mathcal{W}, W)\right)\Big). \quad (3)$$

In the training process, we apply the stochastic gradient descent (SGD) [43] to search the suitable parameters $\{\mathcal{W}, W\}$ for our MO-CNN.

3.3 Age Decoding

We elaborate a delicate CNN with multiple binary classifiers to determine which groups a face image belongs to. However, we can only acquire an ambiguous age range using the classification framework. Since only an ambiguous age range can be acquired using the classification framework, a decoding stage is further developed to obtain the exact age considering the specific mapping relation between ages and age groups. Detailed age decoding stage is explained below.

The objective function, Eq. (3), can be rewritten as

$$J = -\frac{1}{N}\log\Big(\prod_{i=0}^{N-1}\prod_{t=0}^{T-1}\big(\sum_{m=0}^{1} 1\{g_i^t = m\} \cdot p(g_i^t = m|x_i, \mathcal{W}, W)^{\rho_t^m}\big)^{\alpha_t}\Big). \quad (4)$$

Removing the negative logarithm and average factor terms of Eq. (4), our learning procedure is actually to maximize the following equation

$$p(G|X,\mathcal{W},W)=\prod_{i=0}^{N-1}\prod_{t=0}^{T-1}\Big(\sum_{m=0}^{1}1\{g_i^t=m\}\cdot p\big(g_i^t=m|x_i,\mathcal{W},W\big)^{\rho_t^m}\Big)^{\alpha_t}, \tag{5}$$

where $X=\{x_i\}_{i=0}^{N-1}$ and $G=\Big\{\{g_i^t\}_{t=0}^{T-1}\Big\}_{i=0}^{N-1}$ are the whole dataset and the corresponding group labels, respectively.

In Section 3.1, we use an index set $\mathcal{C}_a$ to represent the groups that the face images with age a belong to. And Eq. (5) can be rewritten as following with the index set $\mathcal{C}_a$

$$p(G|X,\mathcal{W},W)=\prod_{i=0}^{N-1}\Big(\prod_{t\in\mathcal{C}_{y_i}}p\big(g_i^t=1|x_i,\mathcal{W},W\big)^{\alpha_t\rho_t^1}\cdot\prod_{t\in\bar{\mathcal{C}}_{y_i}}p\big(g_i^t=0|x_i,\mathcal{W},W\big)^{\alpha_t\rho_t^0}\Big). \tag{6}$$

Note that $\mathcal{C}_{y_i}$ represents those groups that the face image with age y_i belongs to, and $\bar{\mathcal{C}}_{y_i}$ is the complementary set of $\mathcal{C}_{y_i}$. It is assumed that the samples are independent to each other. Therefore, we can define the probability of a face image belongs to age a as following:

$$\mathcal{P}(a|x_i,\mathcal{W},W)=\frac{1}{Z}\prod_{t\in\mathcal{C}_a}p\big(g_i^t=1|x_i,\mathcal{W},W\big)^{\alpha_t\rho_t^1}\cdot\prod_{t\in\mathcal{C}_a}p\big(g_i^t=0|x_i,\mathcal{W},W\big)^{\alpha_t\rho_t^0}, \tag{7}$$

where Z is the normalization factor that makes sure $\sum_{a\in\mathcal{Y}}\mathcal{P}(a|x_i,\mathcal{W},W)=1$. In the training stage, our learning procedure aims to make the probability $\mathcal{P}(a|x_i,\mathcal{W},W)$ reach its maximum when a equals to its real age label y_i. Therefore, the predicting age y_i' for image x_i is

$$y_i'=\arg\max_{a\in\mathcal{Y}}\mathcal{P}(a|x_i,\mathcal{W},W). \tag{8}$$

Our age decoding method is to find the maximal probability of $\mathcal{P}(a|x_i,\mathcal{W},W)$ for the whole age set $\mathcal{Y}$ and take the corresponding age as the final estimated age. This is called Global Age Decoding (GAD). However, it also leads to an enormous computational burden because it conducts computation for all ages and then finds the maximum as its corresponding age. Actually, we can get the coarse age range from the age group classification results, and then use the Local Age Decoding to recover the exact age to reduce the computational complexity. Assume that group m is the group with the maximal probability $p\big(g_i^t=1|x_i,\mathcal{W},W\big)$ for image x_i, which shows that the images x_i is most likely to belong to group m. Thus LAD only compares the probabilities for the ages in $\mathcal{S}_m$, and it can be written as

$$y_i'=\arg\max_{a\in\mathcal{S}_m}\mathcal{P}(a|x_i,\mathcal{W},W). \tag{9}$$

We have made comparisons between GAD and LAD in Section 5.3, which shows that the LAD is more efficient.

4 Experiments

In this section, we first introduce the databases and explain some training details about our experiments. Then we present the experimental results.

TABLE 1
Summary of the Databases Used in Our Experiments

Database	Images	Age range
MORPH	55244	16 - 77
80-20 protocol	5493	
Train (80% images)	4395	
Test (20% images)	1098	
S1-S2-S3 protocol	55244	
S1	10634	
S2	10634	
S3	33976	
FG-NET	1002	0 - 69
Train	990(avg.)	
Test	12(avg.)	
CACD	162941	14 - 62
Train(1800 celebs)	144792	
Val(80 celebs)	7585	
Test(120 celebs)	10564	
Chalearn LAP 2015	4691	3 - 85
Train	2476	
Validation	1136	
Test	1079	
Chalearn LAP 2016	7591	1 - 89
Train	4113	
Validation	1500	
Test	1978	
IMDB-WIKI	523051	0 - 100
Train	297163	
Val	10000	

The table contains the age range information, and the number of images of the corresponding database and its split. The non-face images (e.g., the tattoo images in MORPH database) are removed in our experiments, thus those images are not counted in this table.

4.1 Databases

For real age estimation, we evaluate the proposed method on FG-NET [24], Morph II [25] and CACD [26] databases, under both the controlled and uncontrolled environments. We also evaluate the performance of the proposed method for apparent age estimation on Chalearn LAP datasets [27], [44]. IMDB-WIKI database [27], [44] is also introduced to pretrain our network when evaluating our model on FG-NET, Morph and Chalearn LAP datasets. A summary of those databases is given in Table 1, including age range information, the size of each database and its corresponding spits. Fig. 4 shows some exemplar images of each database. Here, we take a brief introduction on those databases and the test protocols.

FG-NET. The FG-NET dataset contains 1,002 color or grayscale face images of 82 subjects. Those images are taken in a totally uncontrolled environment with large variations of lighting, poses, and expressions. When evaluating on this dataset, we take leave-one person-out (LOPO) cross validation strategy according to the setup of [5], [33], [45], [46], and the averaging performance over the 82 splits is reported.

MORPH II. This database is probably the largest database with precise age labeling and ethnicities. The database includes about 55 thousand face images and age ranges from 16 to 77 years. In our experiments, we employ two typical protocols for evaluation on MORPH dataset:

- According to the test protocol[2] provided by Yi et al. [14], the MORPH dataset would be split into three

2. http://www.cbsr.ia.ac.cn/users/dyi/agr.html

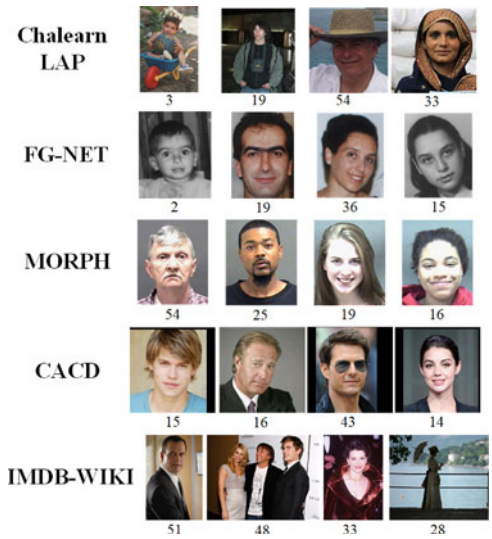


Fig. 5. Sample images from Chalearn LAP, FG-NET, MORPH, CACD and IMDB-WIKI databases. The value below the image is its corresponding age label. FG-NET database includes some old photos (gray image) as shown in the second row. The face images of Chalearn LAP and MORPH databases are taken from the ordinary people, while the images of CACD and IMDB-WIKI databases are from the celebrities. And this difference can be easily found from the figure. Additionally, the CACD database contains some noise. For example, the second image of this databases was wrong labeled. For IMDB-WIKI, it contains more noise, such as a image contains more than one face (see the second image of IMDB-WIKI database) or no face (see the last image of IMDB-WIKI database).

non-overlapped subsets S1, S2, S3 obeying the constructing rules that are detailed in the website provided above. All experiments are repeated twice: 1) training with S1 and testing with S2+S3. 2) training with S2 and testing with S1+S3. Table 1 shows the number of images in each subset. It can be found that, in either way, the number of training images is about a quarter of testing images. For simplicity, we call this test protocol as *S1-S2-S3 protocol*.

- Following the experimental setting in[21], [45], [46], [47], a subset of 5,493 images was used, where the images are selected from Caucasian descent to reduce the cross-race influence. We also randomly split the whole dataset into two non-overlapped parts: 80 percent images for training and 20 percent images for testing. The number of images for training and testing sets are also given in Table 1. In this way, the number of testing images is a quarter of training images. And we call this protocol as *80-20 protocol* for convenience.

CACD. The Cross-Age Celebrity Dataset (CACD) is the largest public cross-age database, which is collected from the Internet Movie DataBase (IMDB). This database, collected from search engines using celebrity name and year (2004-2013) as keywords, contains more than 160 thousand images from 2,000 celebrities. However, the database contains much noise because the age was simply estimated by query year and birth year of that celebrity. We split the database into three subsets: 1,800 noisy celebrities for training, where the number of images is big enough but the age labeling is less precise; 80 cleaned celebrities for validation and 120 cleaned celebrities for testing, where the images are manually checked and the noise images are removed.

Chalearn LAP. The Chalearn LAP challenge is the first competition for apparent age estimation, and it offers images labeled by at least 10 users and then the average age is used as the final annotation. Moreover, the dataset offers the standard deviation for each age label. For the first edition of Chalearn LAP challenge (2015) [27], the organizers collected 4,691 images and all images were split into three subsets: 2,476 images for training, 1,136 images for validation and 1,079 images for testing. For the second edition of Chalearn LAP challenge (2016) [44], the dataset has been extended to 7,591 images, where 4,113 images for training, 1,500 for validation and 1978 for testing. In addition to increasing the number of images, most ages in the dataset are not integers and the standard deviation covers a larger range. Some sample images are given in Fig. 5.

IMDB-WIKI. IMDB-WIKI [5], [9], which contains 523,051 images in total, is the largest dataset for age estimation as far as we know, where the images are crawled from celebrities in IMDb[3] and Wikipedia.[4] However, this dataset contains much noise. The age label is just calculated based on the date of birth of the corresponding celebrity and the year when the photo was taken, thus the accuracy of the age annotations cannot be guaranteed when wrong timestamp occurs or the image comes from a wrong celebrity. Additionally, tiny faces, multiple faces or non-face problems also occur in the dataset as shown in Fig. 5. Even though this dataset is not suitable for evaluation, it is still a good dataset for pretraining for that the majority of the annotations are correct. To use the dataset effectively, we select about 300 thousand images according to the settings in [5], where all non-face images and part of images with multiple faces are removed. What's more, as shown in Table 1, the selected images are randomly divided into two parts: 10,000 images for validation and the rest for training.

4.2 Preprocessing and Experimental Setting

Face Alignment. Face alignment is helpful for age estimation. First, all images are processed by a face detector [48] and a few non-face images would be removed, for example, tattoo images in Morph II database. Then, the active shape models (ASM) [49] are used to detect facial landmarks and all faces would be aligned according to the eyes center and the upper lip. After that, all images are cropped into the size of 224×224 and then fed into the network. Some aligned images are shown in Fig. 9.

Data Augmentation. When evaluating on FG-NET, MORPH II and Chalearn LAP databases, the training images are extremely insufficient. For example, less than five thousand images are used for training when evaluation is taken on Morph dataset with 80-20 protocol. The training set of Chalearn LAP 2015 dataset contains no more than three thousand images, which is even more inadequate. Therefore, increasing training samples is necessary to improve the performance. Usually, there are two ways to expand the training data. One is to enrich the training set with other datasets. For example, we usually pretrain the network from other larger datasets to improve its performance. Another way is to add the virtual image samples.

3. www.imdb.com
4. en.wikipedia.org

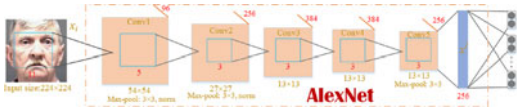


Fig. 6. The network that we used for parameters searching. The network is based on the AlexNet [43], and the last layer also be replaced with multiple binary classifiers. More details of the convolution and pooling layers are shown in the figure.

The first one is a well-known technology and we mainly introduce the method that is used to increase virtual images in our experiments. Here, we augment training images with flipping, rotating by $\pm 5^0$ and $\pm 10^0$, and adding Gaussian white noise with variance of 0.001, 0.005, 0.01, 0.015 and 0.02. The total number of images was increased by 36 times after augmentation. However, data augmentation is only conducted for FG-NET, MORPH II and Chalearn LAP datasets since it is not necessary for CACD database.

Experimental Setting. We train the deep network with a weight decay of 0.0005 and a momentum of 0.9. The learning rate starts from 0.001 and reduced by a factor of 10 along with the number of iterations increases. We set $\alpha_t = 0.1$ for all tasks. AGE7 grouping strategy is taken when experimenting on Chalearn 2016 dataset and AGE9 is taken for the others. Moreover, we set ρ_t^1 to 1 for the experiments on CACD dataset and set ρ_t^1 to 2 for the rest experiments. More details of the setting of AGEn and parameters of balance strategy can be found in Section 4.4. Our algorithm is implemented within the caffe framework [50] on TITAN X GPU. And for all experiments the VGG-16 network was initialized with the weights from training on ImageNet dataset first. For some experiments, the network would be pretrained on IMDB-WIKI dataset and we would explain it in the text.

4.3 Evaluation Metrics

For real age estimation, the Mean Absolute Error (MAE) and Cumulative Score (CS) are usually used as evaluation metrics. MAE indicates the mean absolute error between the predicted result and the ground truth for testing set, and it is calculated as

$$MAE = \frac{1}{m}\sum_{i=0}^{m-1}\left|y_i' - y_i\right|, \tag{10}$$

where y_i' denotes the predicting age for ith image and m is the number of testing face images. MAE is the most frequently used evaluation metric, and obviously, lower MAE result means a better performance. CS(n) is computed as follows:

$$CS(n) = \frac{m_{e\leq n}}{m}, \tag{11}$$

where $m_{e\leq n}$ represents the total number of test images whose absolute error between the predicting results and the ground truth is not greater than n years. Obviously, the higher the CS(n), the better performance it gets.

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

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{12}$$

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. Of course, lower ϵ-error means a better performance and it reaches to 0 when the perfect prediction is achieved.

4.4 Parameters Discussion

As shown in Fig. 11, the distributions of MORPH II and CACD databases differ greatly. We believe that the optimal parameters of the model is closely related to training data distributions. In this section, we find appropriate age grouping range n of age grouping strategy and penalty coefficient ρ of data balance strategy via conducting the experiments on validation set with a variety of n and ρ. For the penalty coefficient $\rho = \{\rho_t^0, \rho_t^1\}_{t=0}^{T-1}$. We assume that it is the same for all tasks, so it can be written as $\rho = \{\rho^0, \rho^1\}$. However, it would take a lot of effort for $\rho = \{\rho^0, \rho^1\}$. For the sake of simplicity, we set $\rho^0 = 1$ and only change the value of ρ^1 in the parameter searching process.

The CACD and Chalearn LAP 2015 & 2016 datasets have offered validation sets. Thus, we directly evaluate the model on their validation set to choose the appropriate parameters. However, since validation set is not offered in MORPH II dataset, we randomly select 2,000 images from its training set as validation set. These images will, therefore, not be used for training in the parameter searching process. Random selection also ensures that the distribution of training data remains unchanged. Since training with VGG-16 network consumes a lot of time, we conduct the experiments with a shallower network, basd on AlexNet [43], which is shown in Fig. 6.

The results on validation set are shown in Table 2. From the results, we adopt AGE9 strategy and $\rho^1 = 2, \rho^1 = 1, \rho^1 = 2$

TABLE 2
MAE Results with a Variety of n and ρ^1 on Validation Set

ρ^1 \ n	(a)				(b)				(c)				(d)				(e)			
	5	7	9	11	5	7	9	11	5	7	9	11	5	7	9	11	5	7	9	11
1	3.61	3.45	3.41	3.32	3.44	3.39	3.22	3.27	5.52	5.33	**5.23**	5.32	4.97	4.97	4.87	4.90	5.41	5.15	4.98	5.01
2	3.38	3.30	**3.21**	3.30	3.26	3.18	**3.17**	**3.17**	5.26	5.34	5.43	5.33	4.93	4.89	**4.86**	4.88	5.08	**4.94**	4.99	5.04
3	3.26	3.28	3.23	3.38	3.20	3.22	3.25	3.25	5.25	5.41	5.54	5.50	4.96	5.04	5.03	**4.86**	5.15	4.97	5.02	5.08
4	3.63	3.32	3.34	3.45	3.18	3.19	**3.17**	3.32	5.45	5.50	5.63	5.49	4.99	4.96	4.91	4.99	5.08	5.14	5.06	5.15

(a) Results on the validation set of S1 on MORPH II. (b) Results on the validation set of S2 on MORPH II. (c) Results on the validation set of CACD database. (d) Results on the validation set of Chalearn LAP 2015 database. (e) Results on the validation set of Chalearn LAP 2016 database.

TABLE 3
The Comparisons Between the Proposed Method and Other State-of-the-Art Methods on MORPH II Database with the S1-S2-S3 Protocol

Method	Train Set	Test Set	MAE	Avg. MAE
Ours(IMDB-WIKI)	S1	S2+S3	2.82	**2.70**
	S2	S1+S3	2.58	
Ours	S1	S2+S3	3.04	**2.86**
	S2	S1+S3	2.68	
Soft softmax [38] (IMDB-WIKI)	S1	S2+S3	3.14	3.03
	S2	S1+S3	2.92	
Soft softmax [38]	S1	S2+S3	3.24	3.14
	S2	S1+S3	3.03	
Multi-scale CNN [14]	S1	S2+S3	3.72	3.63
	S2	S1+S3	3.54	
BIF+KCCA [12]	S1	S2+S3	4.00	3.98
	S2	S1+S3	3.95	
BIF+KPLS [11]	S1	S2+S3	4.07	4.04
	S2	S1+S3	4.01	
BIF+rCCA [12]	S1	S2+S3	4.43	4.42
	S2	S1+S3	4.40	
BIF+PLS [11]	S1	S2+S3	4.58	4.56
	S2	S1+S3	4.54	
CNN [51]	S1	S2+S3	4.64	4.60
	S2	S1+S3	4.55	
BIF+KSVM [12]	S1	S2+S3	4.89	4.91
	S2	S1+S3	4.92	
BIF+LSVM [12]	S1	S2+S3	5.06	5.09
	S2	S1+S3	5.12	
BIF+CCA [12]	S1	S2+S3	5.39	5.37
	S2	S1+S3	5.35	

for Morph II, CACD and Chalearn LAP 2015 datasets, respectively. Moreover, AGE7 and $\rho^1 = 2$ are adopted for Chalearn LAP 2016 dataset. For FG-NET database, it contains too few images and all images would be used to evaluate with LOPO strategy. Therefore, we use $n = 9, \rho^1 = 2$ for FG-NET by our experience because those two parameters perform well in most cases.

We can see that AGE9 is a relatively stable grouping strategy and the model could achieve promising results with AGE9 strategy on most validation sets. For grouping range n, when n is smaller, the relationship between the adjacent ages cannot be explored thoroughly and the imbalanced data problem between the images belonging to group or not is more serious because each group includes fewer images. When n is bigger, the images within the group shows greater diversity, which would be harmful to the model. Thus, AGE9 strategy performs well maybe because $n = 9$ is an appropriate grouping value which achieves a good tradeoff between the above two aspects.

4.5 Comparisons

4.5.1 Real Age estimation

In this section we conduct comprehensive evaluations of the proposed method on Morph, FG-NET and CACD datasets for real age estimation.

Results on MORPH II with S1-S2-S3 Protocol. The proposed method achieves an average MAE of 2.86 without pretraining on any additional age dataset. It reduces the MAE by 0.17 compared with the previous state-of-the-art result reported in [38] (see Table 3). To the best of our knowledge, it is the first report with MAE below 3 years under this protocol. The pretraining on IMDB-WIKI dataset further improve the performance, which achieves a MAE of 2.70 years. The CS results are shown in Fig. 8, and our method achieves the best performance.

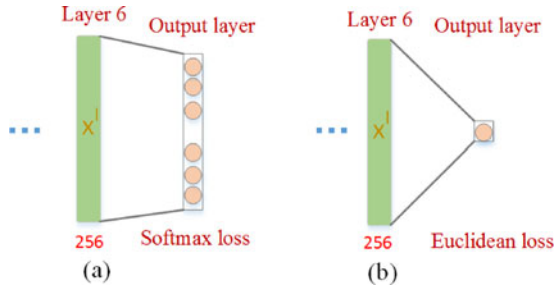


Fig. 7. (a) and (b) show the last two layers of the network of DEX and VGG+euclidean, respectively. The architecture of the lower layers of DEX and VGG+euclidean are the same to our network's.

Results on MORPH II with 80-20 Protocol. Usually, age estimation can be treated as a classification or regression problem. We take two baseline methods of age classification and regression for comparison in this protocol. For age classification, each age is regarded as an independent class. We take Deep EXpection [5], [9] as the baseline method for age classification. DEX is one of the most popular methods for age estimation, which won the first prize of the ChaLearn Looking At People ICCV 2015 challenge [60]. For age regression, we take the classic regression method as the baseline for comparison where the euclidean loss is employed as the loss function. For a fair comparison, the network architecture of DEX and regression-based method are the same to our MO-CNN except the output layer, which are shown in Fig. 7.

From Table 4, our method achieves the state-of-the-art performance with the MAE of 2.93 when directly finetuning on Morph dataset. As far as we know, it is also the first work that reduces the MAE to under 3 years without finetuning on additional age dataset. To further improve the performance, the network is first finetuned on the IMDB-WIKI dataset before finetuning on the Morph dataset, and the proposed method achieves a MAE of 2.52 years, which reduces the state-of-the-art performance by 0.18 years. Besides, the CS comparisons with the state-of-the-art methods are shown in Fig. 8, again our approach also shows its superiority.

Results on FG-NET. Due to FG-NET dataset contains only 1,002 images, we first pretrain our network on IMDB-WIKI datset and then finetune on FG-NET. Two baseline methods have also been added for comparisons. As shown in Table 4, our method achieves the state-of-the-art performance on FG-NET database with an average MAE of 2.96. This improves the previous state-of-the-art result by 0.13. The CS comparisons are shown in Fig. 8, and the proposed method also performs better than other methods.

Results on CACD. Only few works conduct evaluation on CACD database because of its noise. Here, we compare the result with two baseline methods, which are VGG+euclidean regression and DEX. The comparisons are shown in Table 7. Our method achieves the best performance with the lowest MAE of 4.68 years. When CS is taken as the criteria, our

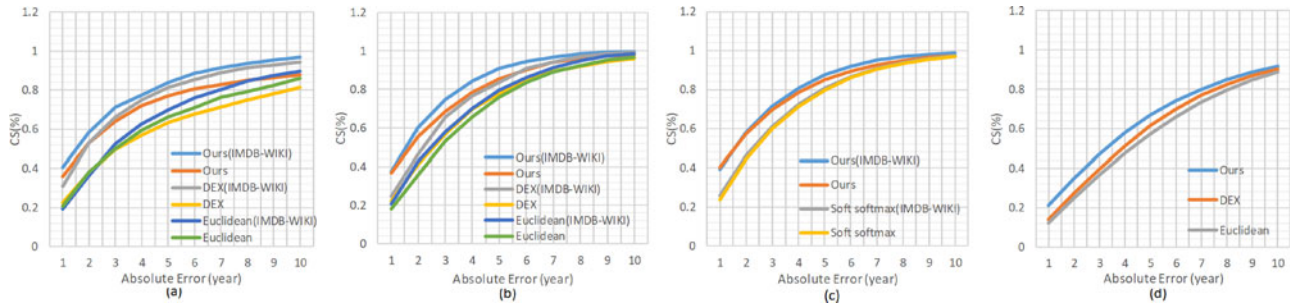


Fig. 8. (a) CS comparisons on FG-NET. (b) CS comparisons on MORPH II with 80-20 protocol when training with 80 percent images and testing with 20 percent images. (c) CS comparisons on MORPH II with S1-S2-S3 protocol. The experiments are repeated twice: 1) Training with S2 and testing with S1+S3; 2) training with S1 and testing with S2+S3, and the average CS performance is reported. (d) CS comparisons on CACD when training with 1,800 celebrities and testing with 120 celebrities.

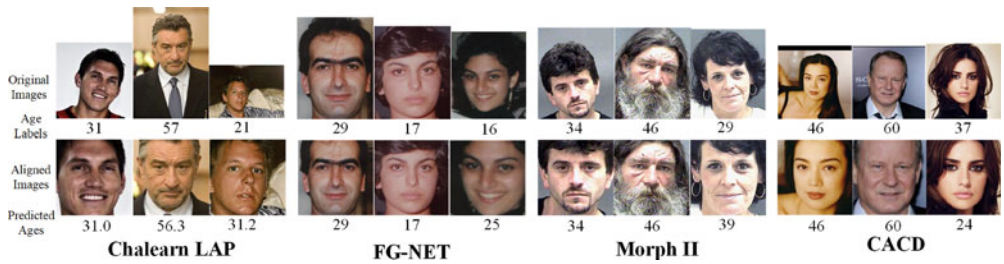


Fig. 9. The original and aligned images of Chalearn LAP, FG-NET, Morph and CACD databases. The predicted ages of both good and bad estimation are given in the figure. Note that the predicted age on Chalearn LAP dataset is not an integer due to the averaging of the predictions of the augmented testing images and an ensemble of networks.

method also performs much better than other methods as shown in Fig. 8. This indicates that our method is capable of estimating age from face images in the wild. Note that we do not finetune our network on IMDB-WIKI dataset because some images from IMDB-WIKI and CACD are duplicated.

4.5.2 Apparent Age Estimation

In this section, the evaluation on Chalearn LAP dataset will be presented.

Results on Chalearn LAP 2015. As a competition dataset of apparent age estimation, Chalearn LAP dataset is more special than other public datasets. Following the tricks used in [5], [9], [41], we finetune our network on both training and validation sets after finetuning on a large additional age dataset, e.g., IMDB-WIKI dataset. In the test phase, each image is flipped, and then rotated by 0^o, $\pm 5^o$, thus each image would be tested by 6 times and then averaging those predictions. Note that for all results except for Chalearn LAP dataset in this paper are based on a single test image. To further improve the performance, an ensemble of 8 networks is employed and we take the average of the predictions as the final estimated age. But the ensemble technology is only taken when evaluating on the test set of Chalearn LAP dataset. We also report the performance on the validation set with only finetuning on training set.

The experimental results are shown in Table 5. The proposed method achieves a better performance than other teams with a final ϵ-error of 0.263547. For validation set, our method also achieves a lower MAE and ϵ-error based on a single network. Due to many tricks we have employed in this evaluation, more training details is presented in Section 5.1.

Results on Chalearn LAP 2016. Different from Chalearn LAP 2015, most ages in Chalearn LAP 2016 dataset are not integers. If we train the network with rounding ages, much information would be sacrificed. To reduce information lossing, we follow the work [40], [41] to encode each age label y and its corresponding deviation σ into a label distribution. The distribution is a set of possibilities representing the description degrees of their corresponding labels, which is defined as follows:

$$P_{l_i} = \frac{1}{Z_{y,\sigma}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(l_i-y)^2}{2\sigma^2}}, i = 0, \ldots, K, \tag{13}$$

where $Z_{y,\sigma}$ is the normlization factor related to age label y and its deviation σ. We generate a random age label for each image according to its label distribution and regard the random age label as the ground truth label in the training

TABLE 4
The Results on Morph II Database with 80-20 Protocol and FG-NET Database

Method	Morph II	FG-NET
Human workers [52]	6.30	4.70
AGES [7]	8.83	6.77
MTWGP [53]	6.28	4.83
CA-SVR [46]	5.88	4.67
OHRank [45]	5.69	4.85
DLA [47]	4.77	4.26
VGG+SVR [54]	3.45	–
VGG+euclidean	3.49	4.77
VGG+euclidean (IMDB-WIKI)	3.15	4.30
DEX [5]	3.25	4.63
DEX (IMDB-WIKI)[5]	2.68	3.09
Ours	**2.93**	**4.34**
Ours (IMDB-WIKI)	**2.52**	**2.96**

Our method achieves the state-of-the-art performance on both databases.

TABLE 5
Comparisons with the State-of-the-Art Methods on the Chalearn LAP 2015 Dataset

Rank	Team	Validation Set[1]		Test Set[2]		Pretrain Set	Network	Num. of Networks
		MAE↓	ϵ-error↓	MAE↓	ϵ-error↓			
–	Ours	**3.21**	**0.28**	**2.94**	**0.263547**	IMDB-WIKI	VGG-16	8
1	CVL ETHZ [5], [9]	3.25	0.28	–	0.264975	IMDB-WIKI	VGG-16	20
2	ICT-VIPL [41]	3.33	0.29	–	0.270685	FG-NET,Morph, CACD, et al.	GoogleNet	8
3	WVU CVL [19]	–	0.31	–	0.294835	FG-NET, Morph, CACD, et al.	GoogleNet	5
4	SEU NJU [55]	–	0.34	–	0.305763	FG-NET, Morph, Adience[56], et al.	GoogleNet	6
	human reference	–	–	–	0.34	–	–	–
5	UMD	–	–	–	0.373352	–	–	–
6	Enjuto	–	–	–	0.374390	–	–	–
7	Sungbin Choi	–	–	–	0.420554	–	–	–
8	Lab219A	–	–	–	0.499181	–	–	–
9	Bogazici	–	–	–	0.524055	–	–	–
10	Notts CVLab	–	–	–	0.594248	–	–	–

[1] *The performance on evaluation set is tested based on a single network.*
[2] *The performance on test set is evaluated by a ensemble of multiple networks, where the number of networks used is shown in the last column of the table. The proposed method achieves the state-of-the-art performance. (↓: the smaller the better).*

TABLE 6
Comparisons with the State-of-the-Art Methods on the Chalearn LAP 2016 Dataset

Rank	Team	Test Set[2]		Pretrain Set	Network	Num. of Networks
		MAE↓	ϵ-error↓			
–	Ours	**3.82**	**0.3100**	IMDB-WIKI	VGG-16	1
1	OrangeLabs[40]	–	0.2411	cleaned IMDB-WIKI, a private children dataset	VGG-16	14
2	palm seu [57]	–	0.3214	IMDB-WIKI	VGG-16	4
3	cmp+ETH [58]	–	0.3361	IMDB-WIKI	VGG-16	10
4	WYU CVL	–	0.3405	–	–	–
5	ITU SiMiT [18]	–	0.3668	IMDB-WIKI	VGG-16	3
6	Bogazici [59]	–	0.3740	–	VGG-16	8
7	MIPAL SNU	–	0.4569	–	–	–
8	DeepAge	–	0.4573	–	–	–

(↓: the smaller the better).

stage. Other experimental settings are the same to Chalearn LAP 2015's.

We find that the performance of methods [18], [40], [57], [58] varies on validation set and test set, for example, OrangeLabs's method didn't achieve the best performance on validation set but outperformed other methods by a large margin on test set. Therefore, we only conduct the evaluation on test set for a consistent comparison. The comparisons are reported in Table 6. Our method achieves the performance on test set with epsilon error of 0.3100 based on a single network, which is the second best result only next to OrangeLabs's [40]. OrangeLabs's method could achieve better performance mainly due to the following reasons: first, they pretrained their network on a cleaned IMDB-WIKI dataset that was arranged and annotated by 26 persons lasting for a few days; second, they manually collected a private dataset with a considerable quantity of images of children, and they have trained 3 separate models for estimating apparent ages of children using the children dataset; third, they used an ensemble of multiple models to boost the performance.

4.6 Computation Time Analysis

We train an age group classification network treating adjacent ages as an independent class. Then a decoding process (LAD or GAD) is used to obtain the probability of each age. In this section, we mainly analyze the accuracy and computational efficiency between the GAD and LAD methods. The comparative experiments are conducted on MORPH II database with CPU, and we only compare the time consumed in the decoding phase. In decoding, there are only two terms changed between the probability of a and $a+1$ accordingly to Eq. (7). To avoid decrease in performance due to rounding error in the continuous calculation process, $\mathcal{P}(a), \mathcal{P}(a+1), \ldots$ are not calculated sequentially. Instead, we compute $\mathcal{P}(a)$ for each a in the whole age set $\mathcal{Y}$ (GAD) or age group set $\mathcal{S}_m$ (LAD) with the maximal classification probability. We find that LAD could spend less time while gets the same performance as with GAD. As shown in Table 8, LAD only needs 4.6 ms to analyze one face image while GAD needs 51.7 ms to do so, which decreases the decoding speed by about 10 times. The visualization of the age probabilities with both LAD and GAD is shown in Fig. 10 (a randomly selected sample from test set). The decoding results are virtually with no difference.

5 Discussion

5.1 Exploring Training Details

Many tricks have been employed when evaluating on Chalearn LAP dataset, e.g., pretraining, data augmentation, a ensemble of networks. In this section a step-by-step investigation is conducted to explore the contributions of each trick.

TABLE 7
The Comparisons on CACD Dataset

Method	Train Set	Test Set	Avg. MAE
Ours	1,800 celebs	120 celebs	**4.68**
DEX [5]	1,800 celebs	120 celebs	4.79
VGG+euclidean	1,800 celebs	120 celebs	5.08

TABLE 8
The Comparisons Between GAD and LAD

Methods	GAD	LAD
Avg. MAE	2.52004	2.52004
Time (per image)	51.7 ms	4.6 ms

The experiment is conducted on Morph II dataset with 80-20 protocol.

As shown in Table 9, the pretraining on IMDB-WIKI dataset seems very helpful, which can reduce the ϵ-error from 0.3709 to 0.2789. This significant improvement shows that the IMDB-WIKI dataset is still useful even though it contains much noise. Also, our data augmentation on training set also makes a great contribution. The ϵ-error is dropped by about 0.012 with the training data augmentation. It is worth noting that the proposed method achieves an ϵ-error of 0.2669 with a single network, which is very close to the best result of Chalearn LAP competition [5], [9].

5.2 Detailed Comparison with DEX

To compare with DEX method thoroughly, we re-implement DEX with the same experimental settings where both face alignment and data augmentation are used. The network of re-implemented DEX method is the same to ours except the last layer as shown in Fig. 7. We conduct the comparisons on FG-NET, Morph II and CACD datasets. When experimenting on FG-NET and Moroh II datasets, the networks are first pretrained on IMDB-WIKI datatset. As shown in the Table 11, our method could still perform better than DEX method on those datasets when adopting the same experimental settings. Furthermore, we also implement our method with the same training settings as DEX's. Besides having selected part of images with small noise of IMDB-WIKI dataset for pretraining, Rothe et al. [5] have also equalized the age distribution of the selected images to improve the model generalization capability. However, they didn't make the list of pretraining images public to the community. Therefore, for a fair comparison, we didn't pretrain the model on IMDB-WIKI dataset when conducting experiments with the same training settings as DEX's. The comparisons on FG-NET, Morph and CACD dataset are shown in Table 11, our method achieves a better performance. No matter which training settings are employed, our method shows superiority to DEX.

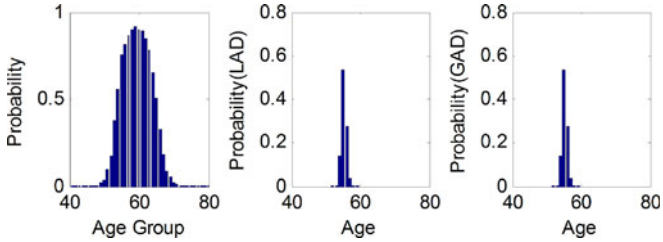


Fig. 10. The visualization of age and age group probabilities.

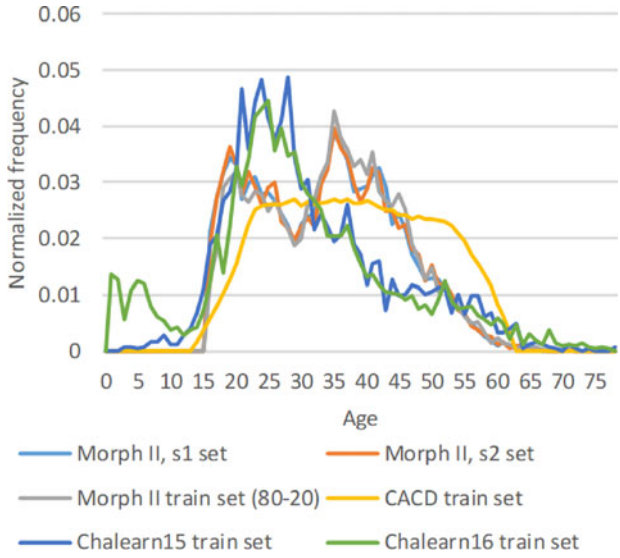


Fig. 11. The distribution of training sets.

5.3 Ablation Study

In this section, we conduct the ablation analysis on the grouping and decoding components of the proposed method. We train the network with multiple classifiers without the grouping component, where each classifier is used to determine the input image belonging to the corresponding age or not. By removing the grouping stage, the predicted age could be directly obtained via maximum probability of the classifiers. So the decoding stage is also dropped in this way. To make a fair comparison, we also conduct experiments with a variety of ρ^1 to find an appropriate value. As shown in Table 10, the minimum MAE can only reach 2.70 when grouping and decoding components are dropped. This means that the performance (or deviation of

TABLE 9
Some Training Details of Our Method on Chalearn Dataset

Crop size of training images	Data augmentation on training Set	IMDB-WIKI pretraining	Data augmentation on testing Set	Num. of networks	MAE	ϵ-error
224×224	No	No	No	1	4.64	0.4027
224×224	Yes	No	No	1	4.30	0.3709
224×224	Yes	Yes	No	1	3.08	0.2789
224×224	Yes	Yes	Yes	1	2.97	0.2669
224×224	Yes	Yes	Yes	8	**2.94**	**0.2635**

All results are finetuning on both training and validation sets, and then testing on test set.

TABLE 10
The Comparisons Between with and without Grouping and Decoding Components on Morph II Dataset Under 80-20 Protocal

	With grouping and decoding	Without grouping and decoding										
ρ^1	–	1	2	3	4	5	6	7	8	9	10	11
MAE	2.52[1]	2.89	2.82	2.79	2.78	2.74	2.73	2.70	2.75	2.71	2.71	2.76

[1] *The result is got by using AGE9 and ρ^1 of 2.*

TABLE 11
The Comparisons Between Our Method and DEX on FG-NET, Morph II and CACD Datasets

	Our training settings			DEX's training settings		
Method	FG-NET	Morph II	CACD	FG-NET	Morph II	CACD
Ours	2.96	2.52	4.68	4.30	3.01	4.73
DEX	3.01	2.66	4.75	4.63	3.25	4.79

Note that here we adopt 80-20 protocol when evaluating on Morph II dataset.

age estimation) of model without grouping and decoding components is 0.18 years less than that with those components. When grouping and decoding components are dropped, each age would be regarded as a single age group and the relationship between adjacent ages can't be explored either. All those result in a decrease in performance. From this perspective, the grouping and decoding components are of critical importance to our method.

6 Conclusion

In this paper, we propose a deep learning solution for age estimation based on a single network to account for aging randomness. First, an age group-n encoding strategy is proposed to group ages, where adjacent ages are grouped into the same group and each group is regarded as an independent class. Then, age group classification is implemented in a CNN with multiple outputs and we recover the exact age for each face image by decoding the classification results. Moreover, we modify our algorithm to address the imbalance data learning problem. Finally, the evaluations on multiple age databases show that the proposed method achieves the state-of-the-art performance.

Acknowledgments

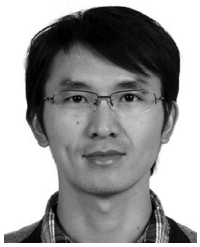
This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects ♯61502491, ♯61473291, ♯61572501, ♯61572536, ♯61673052, Science and Technology Development Fund of Macau (No.112/2014/A3, 151/2017/A, 152/2017/A), NVIDIA GPU donation program and AuthenMetric *R&D* Funds. Zichang Tan and Jun Wan contribute equally to this paper.

References

[1] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Berlin, Germany: Springer, 2005.
[2] M. Fairhurst, *Age Factors in Biometric Processing*. Stevenage, U.K.: Institution Eng. Technol., 2013.
[3] Y. Ma and G. Qian, *Intelligent Video Surveillance: Systems and Technology*. Boca Raton, FL, USA: CRC Press, 2009.
[4] C. Shan, F. Porikli, T. Xiang, and S. Gong, *Video Analytics for Business Intelligence*. Berlin, Germany: Springer, 2012.
[5] 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, 2016, doi:10.1007/s11263-016-0940-36.
[6] X. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[7] X. Geng, Z. Zhou, and K. Smithmiles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.
[8] G. Guo, G. Mu, Y. Fu, and T. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 112–119.
[9] 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*, Dec. 2015, pp. 252–257.
[10] 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.
[11] 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.*, 2011, pp. 657–664.
[12] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA versus PLS," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–6.
[13] T. Liu, Z. Lei, J. Wan, and S. Z. Li, "DFDnet: Discriminant face descriptor network for facial age estimation," in *Proc. Chin. Conf. Biometric Recognit.*, 2015, pp. 649–658.
[14] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 144–158.
[15] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chimica Acta*, vol. 185, no. 86, pp. 1–17, 1986.
[16] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[17] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters Reviews*, vol. 11, no. 10, pp. 203–224, 2007.
[18] R. C. Malli, M. Aygun, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 714–721.
[19] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 267–273.
[20] H. F. Yang, B. Y. Lin, K. Y. Chang, and C. S. Chen, "Automatic age estimation from face images via deep ranking," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 55.1–55.11.
[21] 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.*, Jun. 2016, pp. 4920–4928.
[22] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," arXiv preprint arXiv:1305.1707, 2013.
[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[24] The fg-net aging database. [Online]. Available: http://www.fgnet.rsunit.com/
[25] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 341–345.
[26] B. C. Chen, C. S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015.
[27] S. Escalera, et al., "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 1–9.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[29] N. D. V. Young HoKwon, "Age classification from facial images," *Comput. Vis. Image Understanding*, vol. 74, no. 1, pp. 1–21, 1999.
[30] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.
[31] 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.
[32] 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*, 2006, pp. 307–316.
[33] 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.
[34] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2007, pp. 1383–1386.
[35] F. Gao and H. Ai, "Face age classification on consumer images with Gabor feature and fuzzy LDA method," in *Advances in Biometrics*. Berlin, Germany: Springer, 2009, pp. 132–141.
[36] A. Günay and V. V. Nabiyev, "Automatic age classification with LBP," in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, 2008, pp. 1–4.
[37] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 4465–4470.
[38] Z. Tan, Z. Shuai, W. Jun, L. Zhen, and S. Z. Li, "Age estimation based on a single network with soft softmax of aging modeling," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 203–216.
[39] Z. Huo, et al., "Deep age distribution learning for apparent age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 17–24.
[40] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 96–104.
[41] X. Liu, et al., "AgeNet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 258–266.
[42] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[44] S. Escalera, et al., "ChaLearn looking at people and faces of the world: Face analysis workshop and challenge 2016, " in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 706–713.
[45] K. Y. Chang, C. S. Chen, and Y. P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 585–592.
[46] K. Chen, S. Gong, T. Xiang, and C. L. Chen, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2467–2474.
[47] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 534–541.
[48] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I–I.
[49] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape modelstheir training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[50] Y. Jia, et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[51] M. Yang, S. Zhu, F. Lv, and K. Yu, "Correspondence driven adaptation for human profile recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 505–512.
[52] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human versus machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.
[53] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2622–2629.
[54] 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.*, 2016, pp. 5553–5561.
[55] X. Yang, et al., "Deep label distribution learning for apparent age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 344–350.
[56] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 34–42.
[57] Z. Huo, et al., "Deep age distribution learning for apparent age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 722–729.
[58] M. Uricar, 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*, Jun. 2016, pp. 730–738.
[59] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel ELM and CNN based facial age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 785–791.
[60] X. Baro, et al., "ChaLearn looking at people 2015 challenges: Action spotting and cultural event recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 1–9.

Zichang Tan received the BE degree from the Department of Automation, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2016. He was named as an outstanding graduate of the college when he graduated. He is currently working toward the PhD degree at the Institute of Automation, Chinese Academy of Science (CASIA). His main research interests include deep learning, face attribute analysis and face recognition.



Jun Wan received the BS degree from the China University of Geosciences, Beijing, China, in 2008, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. Since January 2015, he has been an assistant professor in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science (CASIA). He received the 2012 ChaLearn One-Shot-Learning Gesture Challenge Award, sponsored by Microsoft, ICPR 2012. He also received the 2013, 2014 Best Paper Award from the Institute of Information Science, Beijing Jiaotong University. His main research interests include computer vision, machine learning, especially for gesture and action recognition, facial attribution analysis (i.e., age estimation, facial expression, gender and race classification). He has published papers in top journals, such as the *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*. He has served as the reviewer on several top journals and conferences, such as the *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*, the *Public Relations Journal*, ICPR2016, CVPR2017, ICCV2017, FG2017. He is a member of the IEEE.

Zhen Lei received the BS degree in automation from the University of Science and Technology of China, in 2005, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an associate professor. He has published more than 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015. He is a senior member of the IEEE.

Ruicong Zhi received the PhD degree in signal and information processing from Beijing Jiaotong University, in 2010. From 2008 to 2009, she visited the Sound and Image Processing Laboratory, Royal Institute of Technology (KTH) as a joint PhD. She is currently an associate professor in the School of Computer and Communication Engineering, University of Science and Technology Beijing. She has published more than 50 papers, and has six patents. She has been the recipient of more than ten awards, including the National Excellent Doctoral Dissertation Award nomination, the prize of Science and Technology of Beijing etc. Her research interests include facial and behavior analysis, emotion analysis, image processing, and pattern recognition.

Guodong Guo (M'07-SM'07) received the BE degree in automation from Tsinghua University, Beijing, China, the PhD degree in pattern recognition and intelligent control from Chinese Academy of Sciences, Beijing, China, and the PhD degree in computer science from the University of Wisconsin-Madison, Madison, Wisconsin. He is an associate professor in the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, West Virginia. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France, Ritsumeikan University, Kyoto, Japan, Microsoft Research, Beijing, China, and North Carolina Central University. He authored a book, *Face, Expression, and Iris Recognition Using Learning-based Approaches* (2008), coedited a book, *Support Vector Machines Applications* (2014), and published about 100 technical papers. His research interests include computer vision, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, Outstanding Researcher (2013-2014) at CEMR, WVU, and New Researcher of the Year (2010-2011) at CEMR, WVU. He was selected the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council (MMTC) on July 29, 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15", respectively. He is a senior member of the IEEE.

Stan Z. Li received the BEng degree from Hunan University, China, the MEng degree from National University of Defense Technology, China, and the PhD degree from Surrey University, United Kingdom. He is currently a professor and the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He was with Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor in the Nanyang Technological University, Singapore. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published more than 200 papers in international journals and conferences, and authored and edited eight books. He was an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and is acting as the editor-in-chief for the *Encyclopedia of Biometrics*. He served as a program cochair for the International Conference on Biometrics 2007 and 2009, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE fellow for his contributions to the fields of face recognition, pattern recognition and computer vision and he is a member of the IEEE Computer Society.