Deep and Ordinal Ensemble Learning for Human Age Estimation From Facial Images

Jiu-Cheng Xie and Chi-Man Pun, *Senior Member, IEEE*

***Abstract*— Some recent work treats age estimation as an ordinal ranking task and decomposes it into multiple binary classifications. However, a theoretical defect lies in this type of methods: the ignorance of possible contradictions in individual ranking results. In this paper, we partially embrace the decomposition idea and propose the Deep and Ordinal Ensemble Learning with Two Groups Classification ($\text{DOEL}_{\text{2groups}}$) for age prediction. An important advantage of our approach is that it theoretically allows the prediction even when the contradictory cases occur. The proposed method is characterized by a deep and ordinal ensemble and a two-stage aggregation strategy. Specifically, we first set up the ensemble based on Convolutional Neural Network (CNN) techniques, while the ordinal relationship is implicitly constructed among its base learners. Each base learner will classify the target face into one of two specific age groups. After achieving probability predictions of different age groups, then we make aggregation by transforming them into counting value distributions of whole age classes and getting the final age estimation from their votes. Moreover, to further improve the estimation performance, we suggest to regard the age class at the boundary of original two age groups as another age group and this modified version is named the Deep and Ordinal Ensemble Learning with Three Groups Classification ($\text{DOEL}_{\text{3groups}}$). Effectiveness of this new grouping scheme is validated in theory and practice. Finally, we evaluate the proposed two ensemble methods on controlled and wild aging databases, and both of them produce competitive results. Note that the $\text{DOEL}_{\text{3groups}}$ shows the state-of-the-art performance in most cases.**

***Index Terms*— Human age estimation, ensemble learning, ordinal regression, convolutional neural network.**

I. Introduction

PREDICTING human ages from corresponding facial images is a popular yet challenging topic in computer vision. The popularity of this topic may be because of its wide potential applications, such as intelligent advertising, human-computer interaction, effective filtering in criminal investigation, etc., [1], [2]. The challenge of this task should be partially owed to uncertainty in face aging processes, which manifests as personalized conditions even subjects are under the same age. The uncertainty are caused by two types of factors, intrinsic and extrinsic, respectively. Some instances of the former type include gender, race and states of health, and some examples of the latter type contain living environments and habits. However, the largest challenge may be the certainty in age progressions, which shows that our faces mainly change in bone structures from the baby period to the adolescent period but in skin textures after related subjects step into the adulthood phase. One of the key points in simulating this general process is appropriate modelling.

Early work tends to regard age estimation as a classification problem or a regression problem. However, compared with continuous and infinite variables in regression, age labels are discrete and finite. On the other hand, differing from independent categories in standard classification, age classes have a natural order. Thus, age estimation should be an ordinal regression (also called ordinal classification) problem essentially. Even so, compared with other ordinal regression tasks, such as credit rating [3] and information retrieval [4], age estimation is particular. The particularity lies in that the considered age range is relatively wide, usually from age 0 to 100. Unfortunately, constructing a big database with exact age labels is difficult in reality, where the bigness means each age class has enough facial image instances. These facts determine that we can not directly adopt those methods which perform well in similar kind of tasks for age prediction.

Methods of [5]–[7] all employ a common solution for age estimation, though their implement details are different. Specifically, it correlates each age label with a rank and learns a binary classifier for each rank. Every classifier is expected to predict whether the rank of the target face is lager than a specific ranking value associated with current classifier or not. Each classifier is learned on a particular database which is a relabelled version of the original entire training data. Final age prediction is obtained by summing up the results from those classifiers. An important advantage of this kind of methods is, it both encodes the ordinal information among age classes and alleviates the shortage of training data. However, a theoretical defect also exists in them. To be clear, the sum operation used by them is reasonable only when achieved results from those classifiers are consistent. Nevertheless, if the consistency is broken, for instance, the face is classified to be smaller than rank $k-1$ but larger than rank $k+1$ by two different classifiers, then this fusion strategy becomes illogical in theory. Unfortunately, the desired consistency conditions can not be satisfied all the time in practice.

Manuscript received August 7, 2019; revised December 5, 2019 and January 2, 2020; accepted January 3, 2020. Date of publication January 13, 2020; date of current version February 4, 2020. This work was partly supported by the University of Macau under Grants: MYRG2018-00035-FST and MYRG2019-00086-FST, and the Science and Technology Development Fund, Macau SAR (File no. 041/2017/A1 and 0019/2019/A). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vitomir Štruc. *(Corresponding author: Chi-Man Pun.)*

The authors are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: cmpun@um.edu.mo).

Digital Object Identifier 10.1109/TIFS.2020.2965298

We observe that the solution mentioned above is partially similar to a specific branch of ensemble methods. In general, approaches in this branch first build various independent and parallel base learners for individual predictions and then employ a particular aggregation strategy, such as averaging or voting, to combine these results. Two representative ensemble methods belonging to this branch include Bagging [8] and Random Forest [9]. Considering the solution of [5]–[7] learns a set of binary classifiers and takes the sum operation to aggregate outcomes from them, thus it seems to share some similarities with the particular type of ensemble methods talked above. However, they still have some differences, a noticeable one of which is that the conventional ensemble methods have no restrictive requirements when aggregating results of base learners. Inspired by this, we try to adapt the previous solution from the ensemble learning perspective and hope the modified version is able to bypass the consistency requirement during the fusion of individual predictions.

In this paper, we first propose an ensemble method for human age estimation, the Deep and Ordinal Ensemble Learning with Two Groups Classification ($\text{DOEL}_{2\text{groups}}$), which can be regarded as a deep modification of the special type of previous methods introduced in [5]–[7]. To be more specific, an ordinal ensemble is set up first for individual age-group predictions, the size of which is jointly decided by the considered age range and the property of base learners within it. Here the property refers to number of groups classified by a base learner. For each base learner, whole classes are divided into two non-overlapped groups, but age groups of adjacent learners are overlapped. Through this way, the concerned ordinal information is implicitly embedded into the ensemble. Each base learner will output a probability distribution of the target face belonging to specific two age groups, and outcomes from all learners can be integrated into a probability distribution matrix. In order to fuse these individual predictions, then a specially designed plurality voting strategy is applied to them. According to this strategy, the achieved probability distribution matrix is converted to a counting distribution vector, each element in which describes the possibility of the instance under a particular age class. The final estimative age is the index of the age class with the highest counting value. In practical implementation, we use a single Convolutional Neural Network (CNN) with multiple output layers to construct the ensemble, so that age estimation can be carried out in an end-to-end mode. To sum up, when contrasted with the original version, the proposed one inherits the same advantages of the former while solving its theoretical defects in fusing individual estimations.

Furthermore, we suggest splitting the whole age range into three age groups for each base learner in the ensemble, and the derived new approach is named the Deep and Ordinal Ensemble Learning with Three Groups Classification ($\text{DOEL}_{3\text{groups}}$). As observed in experimenting, for base learners in the $\text{DOEL}_{2\text{groups}}$, they often make false judgments when the age label of input face is at the boundary of two age groups. Therefore, we think about treating the particular age class near the border between the original two groups as another age group. We hope this more refined grouping scheme will improve the classification accuracy of each base learner and then further boost final estimation performance by the total ensemble. Construction of the ensemble and followed aggregation of individual estimations in this method are both similar to corresponding operations in the $\text{DOEL}_{2\text{groups}}$. Note that the previous work [10] also advises a three grouping scheme, but it adopts hand-crafted features and applies different fusion strategy on base learners. Finally, effectiveness of this three-group version is validated with theoretical analyses and experimental evidence.

In summary, main contributions of this work include:

1) We propose two ensemble methods under the CNN framework for human age estimation, both of which consist of an ordinal ensemble and a special aggregation strategy. Because of building the prediction system from the ensemble learning perspective, our methods possess better interpretability in theory and robustness in performance.
2) Concerning the construction of the ordinal ensemble, the superiority of a ternary grouping arrangement over a binary one is proved through theoretical deduction and experimental validations.
3) Our approaches both achieve good prediction results on controlled and wild age databases. Notably, the one with a ternary grouping design obtains the state-of-the-art performance in most circumstances.

The rest of this paper is organized as follows. First, Section II reviews related work. Then, Section III elaborates on the proposed two age estimation methods as well as related comparison analyses. After that, Section IV reports comprehensive experiments for validation and evaluation. Finally, we draw conclusions in Section V.

II. Related Work

In this section, we first introduce three different kinds of ages for estimation. Then conventional methods and modern ones based on neural networks are briefly reviewed. Finally, we categorize existing approaches considering the ordinal relationship among age classes and summarize their respective characteristics.

A. Different Kinds of Ages

The first study on age estimation goes back to 1994, Kwon and da Vitoria Lobo [11] presented a classification method to judge whether the target face belongs to a baby, a young adult, or a senior. The judgment was based on craniofacial changes in feature position ratios and skin texture analysis. As you can see, that paper only focused on age-group classification rather than exact age prediction. In fact, the first mention of chronological age estimation may be in 2002, in a research on simulating facial aging process by Lanitis *et al.* [12]. Note that the authors obtained the prediction in a regression way. Apart from the age group and the chronological age, another related concept is the apparent age, which is decided by human perception. In 2015 and 2016, many creative and effective methods, such as [13]–[18],

came out because of the holding of two consecutive ChaLearn LAP challenges on apparent age estimation.

B. Conventional Methods

Early age estimation systems can be typically split into two separate and successive phases, aging features extraction and age prediction. To represent age information in a face, common used descriptors include general ones and special designed ones. Some instances of the former include the Active Appearance Models (AAM) [19], the Local Binary Patterns (LBP) [20] and the Gabor feature [21], and of the latter contain the AGing pattErn Subspace (AGES) [22], the age manifold [23] and the Biologically Inspired Features (BIF) [24]. After obtaining age features, the rest is to estimate human age. Most works dealt with these features using classification methods such as the Support Vector Machine (SVM) [24], [25] and the fuzzy Liner Discriminant Analysis (LDA) [21], or regression methods like the Support Vector Regression (SVR) [24], the Semidefinite Programming [26] and the Kernel Partial Least Squares (KPLS) [27]. In addition, a few researchers proposed to combine these two types of methods so that the combinations could possess merits of both sides [23], [25]. For more detailed review of early studies on this topic, please refer to [28].

C. CNN Based Methods

During the past few years, the Convolutional Neural Network (CNN) technology has gained a huge success in computer vision, including human age estimation from facial images. There are two crucial benefits of this type of technology. One is that it enables powerful and automatic feature representation, and the other is that it allows features extraction and classification within a single CNN framework. Owing to these two advantages, typical two stages in age estimation can be executed by an unified model. More importantly, age prediction results given by CNN based approaches, even without extra optimization, are comparable to those produced by excellent methods belonging to the conventional type [29]. To further improve the estimation performance, researchers made their efforts in several directions. For example, to reduce the uncertainties caused by internal and external factors, Wan *et al.* [30] proposed a cascaded structure based on CNN to achieve gender, race and age attributes consecutively. Yoo *et al.* [31] and Xie and Pun [32] both employed a conditional mutitask learning strategy, which could take those influential factors into consideration in an efficient way. Liu *et al.* [33] first made prediction in several age groups and then fused derived results. All of them follow the advice suggested by [34], that is performing age estimation under the same condition can achieve smaller predictive errors. The second direction is semiautomatic age feature extraction and fusion, while the belief behind this is automatically extracted features by CNN may be not discriminative enough. In this direction, Chen *et al.* [35] combined different features of specified sub-regions and global regions, whereas Taheri and Toygar [36] exploited multistage features from different layers of a CNN. Despite the strong capability of CNNs, the size of parameters and the amount of related computation in them are usually enormous, which calls for powerful hardware equipments. To make age estimation possible on mobile devices, Gao *et al.* [37] and Yang *et al.* [38] respectively proposed a compact model for age prediction, while both of them produced low estimation errors. Zhang *et al.* [39] argued that, when the compromise between estimation accuracy and model compactness is considered, the standard convolution is enough for small size facial images, whereas a deep-wise separation is unnecessary.

D. Consideration of Ordinal Relationship

In addition, another popular direction for further boosting age estimation performance is to take ordinal relationships among adjacent age labels into account. Existing prediction approaches working in this direction can be roughly classified into four categories. Methods in the first category used cost-sensitive classification since the costs of misclassifications can be forced to be different depending on the distance between real and estimative age class [32], [40]. In the second category, they utilized pairwise constraints to map instances to a real line in order [41], or exploited different relations for different age gaps [42], or learned the general concept of "old and young" [43]. Approaches belonging to the third category employed the Label Distribution Learning (LDL) technique, under which each facial image was equipped with a label distribution rather than a single age label. The label distribution covered a certain number of age labels, while elements in the distribution represented the degree of each label describes the related face [37], [44], [45]. Methods of the forth category decomposed the ordinal regression problem into numerous binary classification ones. To handle these decomposed subtasks, several independent models [5], [7] or a single model with multiple output [6], [46] were employed. A noticeable difference is that methods in the former two types only leverage the instances with the same exact age label during model training process, whereas approaches belonging to the latter two types also use additional samples with adjacent age labels. Since it is hard to collect enough facial images for each age class, methods falling in the first and second categories summed above should be more suitable for age group prediction rather than chronological or apparent age estimation.

In this work, we also transform the complex age estimation task into several simple prediction subtasks, which is partially similar to the fourth kind of methods mentioned above. However, differing from them, our approaches are designed from the ensemble learning perspective, and they possess more solid interpretability in theory. Recently, Shen *et al.* [47] proposed the Deep Regression Forests (DRFs) for age estimation, which is also an ensemble method. Nevertheless, they ignored the ordinal relationships among neighboring age classes, whereas we implicitly encode these information when building our ensembles.

III. Methodology

A. Problem Formulation

For a facial image representation $\boldsymbol{x}_i \in \mathcal{X}$, its chronological age is $y_i \in \mathcal{Y}$, where $\mathcal{Y} = \{0, 1, 2, \ldots, K\}$. Mathematically,

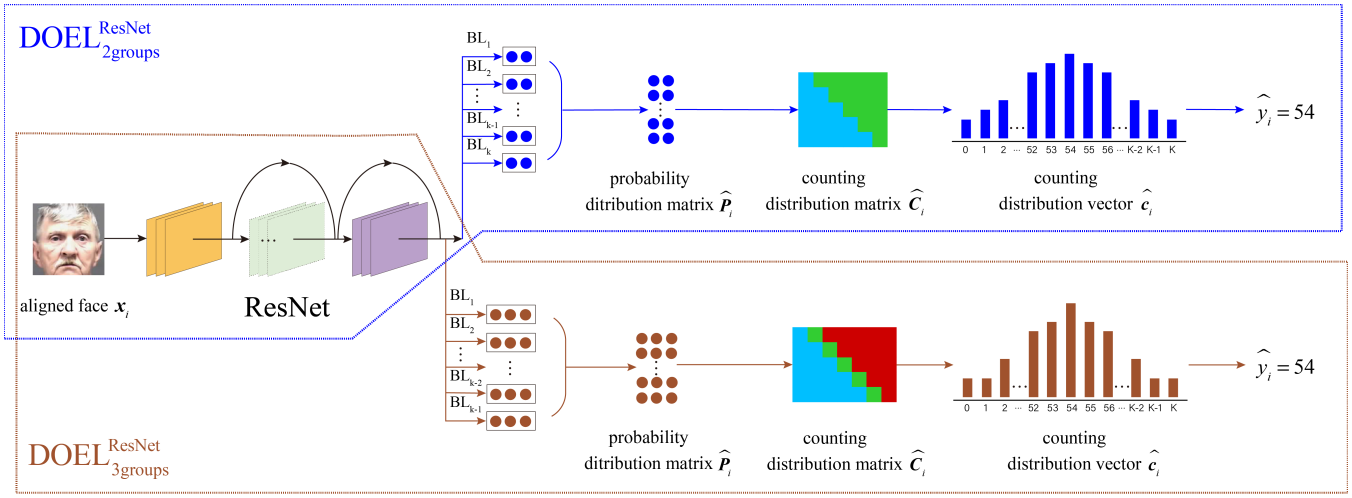

Fig. 1. Illustration of two independent age estimation systems based on the proposed $\text{DOEL}_{2\text{groups}}$ and $\text{DOEL}_{3\text{groups}}$, respectively. Note that since these two systems have the same frontal part, only one is drawn here for simplicity.

age estimation is to make the prediction model find an appropriate mapping from input space to output space, which can be denoted by $f(\cdot): \mathcal{X} \rightarrow \mathcal{Y}$. The model is trained on a sample set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=0}^{N-1}$ iteratively until the predictive age $\hat{y}_i$ for target face $\boldsymbol{x}_i$ is close to its truth age y_i enough. In this paper, we propose two age estimation methods, the $\text{DOEL}_{2\text{groups}}$ and the $\text{DOEL}_{3\text{groups}}$, both of which are based on ensemble learning idea.

B. $DOEL_{2groups}$

First, an ordinal ensemble represented by $E = \{BL_1, BL_2, \ldots, BL_K\}$ is built to decompose the age estimation into several subtasks. This ensemble consists of K different base learners, each of which correlated with a specific age class. Concerning the kth $(1 \leq k \leq K)$ base learner BL_k, the whole age range is divided into two age groups: those age classes before the label k are assigned to the group zero $G_{k,0} = \{a_0, a_1, \ldots, a_{k-1}\}$, and remained ones are allocated to the group one $G_{k,1} = \{a_k, a_{k+1}, \ldots, a_K\}$, where a_m $(0 \leq m \leq K)$ represents the mth age class. Following this special grouping rule, the ordinal relationships are established among adjacent base learners. When given a target face $\boldsymbol{x}_i$, the learner is expected to predict the probabilities of the face belonging to corresponding two age groups, which are denoted by $\hat{P}(G_{k,0}|\boldsymbol{x}_i, BL_k)$ and $\hat{P}(G_{k,1}|\boldsymbol{x}_i, BL_k)$. For simple expression, above two predictive probabilities are referred to as $\hat{P}_i^{k,0}$ and $\hat{P}_i^{k,1}$ in later descriptions. Apparently, they should satisfy following two constraints: $\hat{P}_i^{k,0}, \hat{P}_i^{k,1} \in [0, 1]$ and $\hat{P}_i^{k,0} + \hat{P}_i^{k,1} = 1$. A two dimensional vector, $\hat{\boldsymbol{p}}_i^k = (\hat{P}_i^{k,0}, \hat{P}_i^{k,1})$, is used to stand for the estimative probability distribution. When taking all base learners into consideration, we fuse their outcomes into a probability distribution matrix, $\hat{\boldsymbol{P}}_i = (\hat{\boldsymbol{p}}_i^1; \hat{\boldsymbol{p}}_i^2; \ldots; \hat{\boldsymbol{p}}_i^K)$.

To learn base learners in the constructed ensemble, we need to relabel original training face samples for each of them. Concerning the kth learner, the target face $\boldsymbol{x}_i$ is relabelled with a ground truth distribution $\boldsymbol{p}_i^k = (P_i^{k,0}, P_i^{k,1})$. Here, severer constraints are applied to the ground truth probability distribution. Specifically, it stipulates that $P_i^{k,0}, P_i^{k,1} \in \{0, 1\}$. Then, according to the age grouping rule established above, the ground truth probability distribution is labelled as

$$\boldsymbol{p}_i^k = \begin{cases} (0, 1), & \text{if} \quad 1 \leq k \leq y_i \\ (1, 0), & \text{if} \quad y_i + 1 \leq k \leq K \end{cases}. \tag{1}$$

In total, each face in the original training set are relabelled for K times and $\acute{S} = \left\{(\boldsymbol{x}_i, \boldsymbol{p}_i^1, \boldsymbol{p}_i^2, \ldots, \boldsymbol{p}_i^K, y_i)\right\}_{i=0}^{N-1}$ is used to represent the new database for convenience. The ordinal ensemble is constructed with the help of a deep CNN. To be more specific, a stack of convolutional layers are adopted for facial feature extraction and a fully-connected (FC) layer with softmax is employed for the two-age-group classification. The combination of these two components constitutes a base learner. Furthermore, all base learners are made to share a common feature extraction module so as to reduce the total computation costs. An intuitive demonstration of this ensemble is given in the blue dotted box of Fig. 1. Note that although we employ the popular ResNet here, other effective network structures are alternative. The cross entropy loss is used to supervise the ensemble training process. For the kth base learner, its individual loss is

$$L_k = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{1} P_i^{k,j} \log(\hat{P}_i^{k,j}). \tag{2}$$

When all base learners of the ensemble are taken into consideration, the total loss becomes

$$L = \sum_{k=1}^{K} L_k = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=0}^{N-1} \sum_{j=0}^{1} P_i^{k,j} \log(\hat{P}_i^{k,j}). \tag{3}$$

After the acquisition of the matrix $\hat{\boldsymbol{P}}_i$, it comes to aggregate different probability distributions for the final age prediction. Here we propose a special aggregation strategy which consists of two phases. In the first phase, the obtained probability distribution matrix $\hat{\boldsymbol{P}}_i$ is transformed into a counting distribution

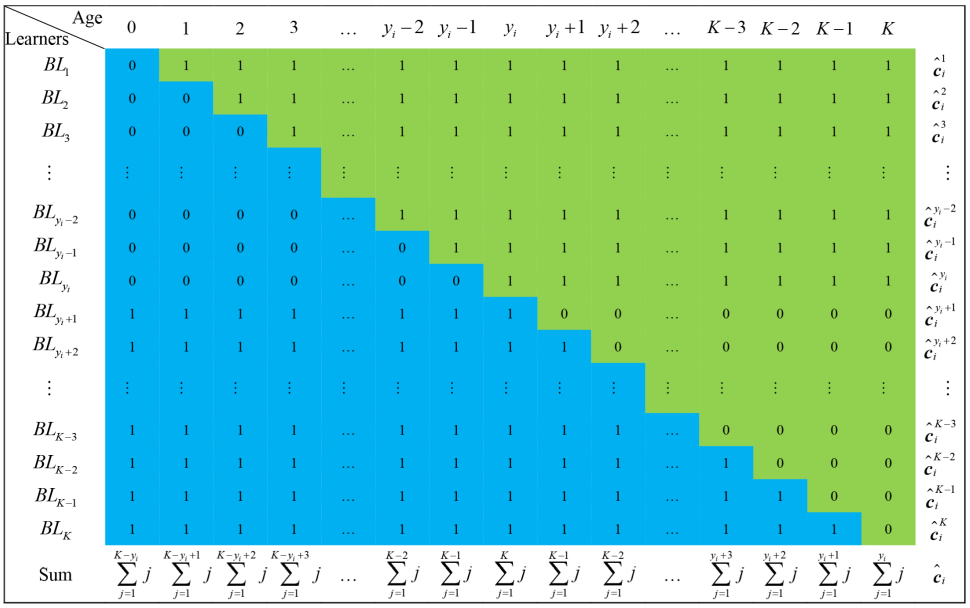

Fig. 2. An application example of the proposed aggregation strategy in the $\text{DOEL}_{2\text{groups}}$. The colored rectangle is the transformed counting distribution matrix. For visual discrimination, age classes belonging to the group zero are marked with the deep sky blue and remained ones belonging to the group one are marked with the green. The bottom row is the final counting distribution vector.

matrix $\hat{\boldsymbol{C}}_i$. To be more particular, for every base learner, related transformation is carried out according to the following rule: assigning counting value 1 to each class belonging to the group with a higher prediction probability, but value 0 to remained age classes. For the kth base learner, its applied transformation rule can be expressed in mathematics as

$$\begin{cases} \hat{C}_i^{k,m} = 1 \text{ and } \hat{C}_i^{k,n} = 0, & \text{if } \ max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,0} \\ \hat{C}_i^{k,n} = 1 \text{ and } \hat{C}_i^{k,m} = 0, & \text{if } \ max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,1}, \end{cases} \tag{4}$$

where

$$\forall m \in \{0, 1, \ldots, k-1\} \text{ and } \forall n \in \{k, k+1, \ldots, K\}. \tag{5}$$

Reasons for taking above operations are two-fold: (1) Distributing grouping results to individual age classes converts heterogeneous predictions into homogeneous ones; (2) Discrete counting values are more robust than continuous probability values. For convenience, a $K+1$ dimensional vector, $\hat{\boldsymbol{c}}_i^k = (\hat{C}_i^{k,0}, \hat{C}_i^{k,1}, \ldots, \hat{C}_i^{k,K})$, is used to denote the transformed counting distribution by the current learner. Furthermore, counting distributions from all learners are collected and combined into a counting distribution matrix, $\hat{\boldsymbol{C}}_i = (\hat{\boldsymbol{c}}_i^1; \hat{\boldsymbol{c}}_i^2; \ldots; \hat{\boldsymbol{c}}_i^K)$. Then in the second phase, deduction of the final age is based on this new matrix. Attributing to the homogeneousness with $\hat{\boldsymbol{C}}_i$, thus we can apply sum operations on elements of each column in this generated matrix. For the mth $(0 \leq m \leq K)$ column, the summed result is denoted by $\hat{C}_i^m$. After this operation, the final counting distribution is acquired, which is

$$\hat{\boldsymbol{c}}_i = (\hat{C}_i^0, \hat{C}_i^1, \ldots, \hat{C}_i^K). \tag{6}$$

This distribution reveals the voting results from the whole ensemble about possibilities of the target face belonging to different age classes. Obviously, according to the plurality voting principle, the final prediction age should be the index of the age class with the largest counting value, which can be written as

$$\hat{y}_i = index(max(\hat{\boldsymbol{c}}_i)). \tag{7}$$

An application example of this aggregation strategy is shown in Fig. 2.

We call this ensemble method the Deep and Ordinal Ensemble Learning with Two Groups Classification ($\text{DOEL}_{2\text{groups}}$) for age estimation.

C. $DOEL_{3groups}$

The critical difference between the last proposed approach and current one is the number of age groups for each base learner: the former leverages two whereas the latter uses three. Thus, we call our second ensemble method the Deep and Ordinal Ensemble Learning with Three Groups Classification ($\text{DOEL}_{3\text{groups}}$) for age prediction.

First, an ordinal ensemble composed of $k-1$ different base learners is set up, and it is denoted by $E = \{BL_1, BL_2, \ldots, BL_{K-1}\}$. Each learner of this ensemble is correlated with three particular and ordered age groups. To be more specific, when given an face instance $\boldsymbol{x}_i$, the kth $(1 \leq k \leq K-1)$ learner BL_k is expected to classify it into one of the following age groups: the group zero $G_{k,0} = \{a_0, a_1, \ldots, a_{k-1}\}$, the group one $G_{k,1} = \{a_k\}$ or the group two $G_{k,2} = \{a_{k+1}, a_{k+2}, \ldots, a_K\}$. For convenience, a three dimensional vector, $\hat{\boldsymbol{p}}_i^k = (\hat{P}_i^{k,0}, \hat{P}_i^{k,1}, \hat{P}_i^{k,2})$, is used to represent the outcome probability distribution by current learner. Moreover, results from all base learners are combined into a probability distribution matrix, $\hat{\boldsymbol{P}}_i = (\hat{\boldsymbol{p}}_i^1; \hat{\boldsymbol{p}}_i^2; \ldots; \hat{\boldsymbol{p}}_i^{K-1})$.

To learning a model that can output the desired $\hat{\boldsymbol{P}}_i$, each face instance for training is relabelled with a particular probability

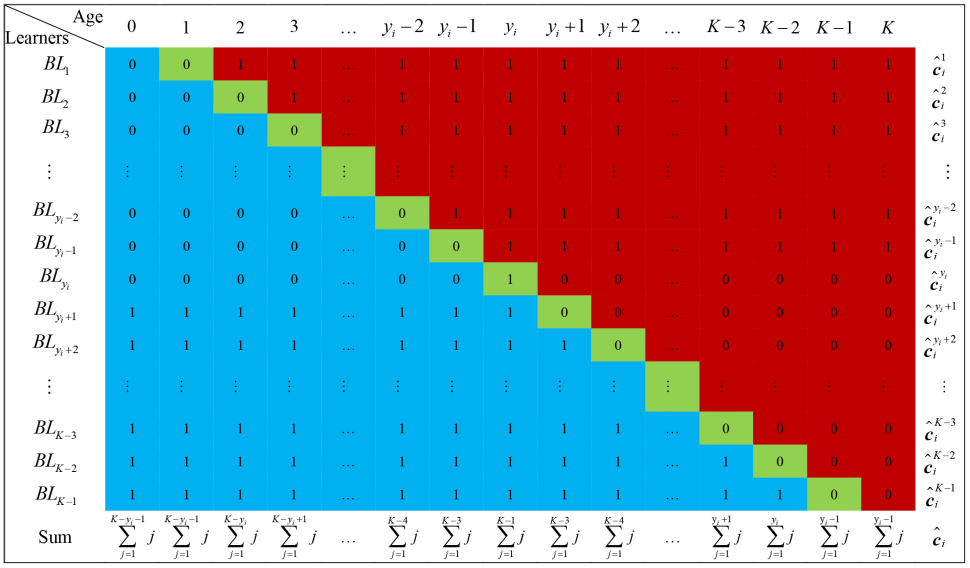

Fig. 3. An application example of the proposed aggregation strategy in the $\text{DOEL}_{3\text{groups}}$. The colored rectangle is the converted counting distribution matrix. For visual discrimination, age classes belonging to the group zero are marked with the deep sky blue, belonging to the group one are marked with the green and belonging to the group two are marked with the red, respectively. The bottom row is the final counting distribution vector.

distribution for every specific base learner. For the learner BL_k, we use $\boldsymbol{p}_i^k = (P_i^{k,0}, P_i^{k,1}, P_i^{k,2})$ to represent its ground truth probability distribution. The relabelling operations are implemented under following two constraints,

$$\begin{cases} P_i^{k,0} < P_i^{k,1} < P_i^{k,2}, & \text{if} \quad 1 \le k \le y_i - 1 \\ P_i^{k,0} < P_i^{k,1} \text{and } P_i^{k,2} < P_i^{k,1}, & \text{if} \quad k = y_i \\ P_i^{k,0} > P_i^{k,1} > P_i^{k,2}, & \text{if} \quad y_i + 1 \le k \le K - 1, \end{cases} \tag{8}$$

and

$$P_i^{k,0} + P_i^{k,1} + P_i^{k,2} = 1. \tag{9}$$

Empirically, the desired distribution is labelled as follows,

$$\boldsymbol{p}_i^k = \begin{cases} (0, 0.2, 0.8), & \text{if} \quad 1 \le k \le y_i - 1 \\ (0.1, 0.8, 0.1), & \text{if} \quad k = y_i \\ (0.8, 0.2, 0), & \text{if} \quad y_i + 1 \le k \le K - 1. \end{cases} \tag{10}$$

Consequently, it produces a new training set $\acute{S} = \{(\boldsymbol{x}_i, \boldsymbol{p}_i^1, \boldsymbol{p}_i^2, \ldots, \boldsymbol{p}_i^{K-1}, y_i)\}_{i=0}^{N-1}$. To construct this ordinal ensemble, a model structure similar to the last method $\text{DOEL}_{2\text{groups}}$ is employed and the diagram of it is shown in the sienna dotted box of Fig. 1. Note that since each base learner of current ensemble divides all age classes into three groups rather than two in the last one, then accordingly, each FC layer has three output neurons here rather than two. The Kullback-Leibler divergence, a measure of how one probability distribution is different from a second, is employed as the individual loss function for each base learner, It can be expressed as

$$L_k = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{2} P_i^{k,j} \ln \left(\frac{P_i^{k,j}}{\hat{P}_i^{k,j}} \right). \tag{11}$$

Furthermore, the total loss for the whole ensemble should be

$$L = \sum_{k=1}^{K-1} L_k = \frac{1}{N} \sum_{k=1}^{K-1} \sum_{i=0}^{N-1} \sum_{j=0}^{2} P_i^{k,j} \ln \left(\frac{P_i^{k,j}}{\hat{P}_i^{k,j}} \right). \tag{12}$$

Then a two-stage aggregation strategy similar to the one used in the last method is employed to deal with the obtained probability distribution matrix $\hat{\boldsymbol{P}}_i$. In the first stage, $\hat{\boldsymbol{P}}_i$ is converted into a counting distribution matrix $\hat{\boldsymbol{C}}_i$. The conversions are implemented as follows: assigning counting value 1 to each class belonging to the age group with the highest prediction probability, but value 0 to remained age classes. Mathematically, for the kth base learner BL_k, this transformation rule can be written as

$$\begin{cases} \hat{C}_i^{k,m} = 1 \text{ and } \hat{C}_i^{k,k} = \hat{C}_i^{k,n} = 0 & \text{if} \quad max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,0} \\ \hat{C}_i^{k,k} = 1 \text{ and } \hat{C}_i^{k,m} = \hat{C}_i^{k,n} = 0 & \text{if} \quad max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,1} \\ \hat{C}_i^{k,n} = 1 \text{ and } \hat{C}_i^{k,m} = \hat{C}_i^{k,k} = 0 & \text{if} \quad max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,2}, \end{cases} \tag{13}$$

where

$$\forall m \in \{0, 1, \ldots, k-1\} \text{ and } \forall n \in \{k+1, k+2, \ldots, K\}. \tag{14}$$

A $K+1$ dimensional vector, $\hat{\boldsymbol{c}}_i^k = (\hat{C}_i^{k,0}, \hat{C}_i^{k,1}, \ldots, \hat{C}_i^{k,K})$, is used to represent the predictive counting distribution from BL_k. Moreover, predictions by $K-1$ different base learners are integrated into a counting distribution matrix, $\hat{\boldsymbol{C}}_i = (\hat{\boldsymbol{c}}_i^1; \hat{\boldsymbol{c}}_i^2; \ldots; \hat{\boldsymbol{c}}_i^{K-1})$. Then in the second stage, operations similar to those employed in the same phase of the last method are made on $\hat{\boldsymbol{C}}_i$. In short, each column's elements of $\hat{\boldsymbol{C}}_i$ are summed up and that generates the final counting distribution $\hat{\boldsymbol{c}}_i$. Clearly, the index of the largest summed counting value should be taken as the estimative age. Please refer to (6) and (7) for mathematical expressions. In addition, a demonstration of this strategy is given in Fig. 3.

D. Comparison Analyses

Authors of [6], [7] propose an ordinal ranking method which decomposes the age estimation into multiple binary classifications. For convenience, this method is called the Ordinal Ranking with Two Groups Classification ($\text{OR}_{2\text{groups}}$) in the

reset of this paper. From another perspective, these binary classifications can be regarded as base learners in an ensemble. Thus, for a intuitive comparison with our $\text{DOEL}_{2\text{groups}}$, next we will briefly review the $\text{OR}_{2\text{groups}}$ from the ensemble learning angle.

To encode the ordinal relationship, the $\text{OR}_{2\text{groups}}$ associates different age classes with ordered ranks and the correspondence between them is one to one. For the kth ($1 \le k \le K$) base learner BL_k, it is expected to judge whether the target face $\boldsymbol{x}_i$ is smaller than the rank k or not. Considering the corresponding relations between age classes and ranks, the learner essentially predicts whether the target face is younger than age k or not. In other words, it judges the face $\boldsymbol{x}_i$ should belong to the age group zero $G_{k,0} = \{a_0, a_1, \ldots, a_{k-1}\}$ or the age group one $G_{k,1} = \{a_k, a_{k+1}, \ldots, a_K\}$. The judgment rule is based on the output probability distribution $\hat{\boldsymbol{p}}_i^k = (\hat{P}_i^{k,0}, \hat{P}_i^{k,1})$, which is produced by the current learner. Mathematically, this rule can be written as

$$\hat{r}_i^k = \begin{cases} 0, & \text{if } max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,0} \\ 1, & \text{if } max(\hat{\boldsymbol{p}}_i^k) = \hat{P}_i^{k,1} \end{cases}, \tag{15}$$

where $\hat{r}_i^k = 0$ means the face is smaller than the rank k, otherwise the opposite. To trained the ensemble, specific training data is constructed for each base learner. For the kth learner, its ground truth probability distribution $\boldsymbol{p}_i^k = (P_i^{k,0}, P_i^{k,1})$ is labelled as follows,

$$\boldsymbol{p}_i^k = \begin{cases} (0,1), & \text{if } 1 \le k \le y_i \\ (1,0), & \text{if } y_i + 1 \le k \le K. \end{cases} \tag{16}$$

The model structure and training process of the $\text{OR}_{2\text{groups}}$ are similar to those with our $\text{DOEL}_{2\text{groups}}$. Here, for a fair comparison between them, the same configuration is adopted for these two methods. The biggest difference among them lies in the way they use to obtain the final age prediction. For the $\text{OR}_{2\text{groups}}$, in view of the correspondence between age classes and ranks, it sums all ranking results as the final age, which can be denoted by

$$\hat{y}_i = \sum_{k=1}^{K} \hat{r}_i^k. \tag{17}$$

The rationality of predicting human age according to (17) calls for a consistent precondition, which also had been mentioned by authors of [6]. To be more specific, for all predictions by K base learners, supposing the jth rank is the last rank that satisfies $\hat{\boldsymbol{r}}_i^j = 1$. Then ideal classification results from those base learners should be as follows, $\hat{r}_i^m = 1$ and $\hat{r}_i^n = 0$, where $\forall m \in \{1, 2, \ldots, j\}$ and $\forall n \in \{j+1, j+2, \ldots, K\}$. However, it is hard to ensure that the desired consistency is always satisfied in actual situations. Strictly speaking, if the consistent condition is broken, although we could make age estimation following (17), this processing mode is hard to be explained theoretically. On the contrary, proposed aggregation strategies in the $\text{DOEL}_{2\text{groups}}$ and the $\text{DOEL}_{3\text{groups}}$ may provide a more reasonable solution. The common idea of our aggregations is based on probability statistics substantially. Thus, no matter each base learner makes a right or a wrong judgement, decisions from all of them should be considered and serve for the final age estimation.

Next, error expectations of the $\text{OR}_{2\text{groups}}$ and two proposed ensemble methods on age estimation are calculated for comparison. To facilitate quantitative analyses, an assumption is made that the amount of training data are enough to make base learners produce unbiased estimations. Since base learners in the $\text{OR}_{2\text{groups}}$ and the $\text{DOEL}_{2\text{groups}}$ both works on binary classification tasks, the probability of making a right or a wrong decision by each learner should be the same in theory, that is $1/2$. On the other hand, learners in the $\text{DOEL}_{3\text{groups}}$ make ternary classifications. As a result, the theoretical probability of a correct or a incorrect decision by each of them should be $1/3$ and $2/3$, respectively.

First, let's consider a simple situation that, among decisions made by the whole base learners of an ensemble, no more than one of them is wrong. Given a facial instance $\boldsymbol{x}_i$ aged y_i, for the base learner BL_k ($1 \le k \le K$) in the $\text{OR}_{2\text{groups}}$ or the $\text{DOEL}_{2\text{groups}}$, a two-dimensional vector, $\hat{\boldsymbol{p}}_k = (\hat{P}_k^0, \hat{P}_k^1)$, is used to represent a collection of probabilities of two cases. $\hat{P}_k^0$ is the probability of the first case that current learner classifies the target face into the age group $G_{k,0}$, and $\hat{P}_k^1$ is the probability of the second case that the face is classified into the age group $G_{k,1}$, while other base learners all provide correct classifications in both cases. According to the deduction in the last paragraph, it is easy to get $\hat{\boldsymbol{p}}_k = (1/2^K, 1/2^K)$. However, because of different aggregation strategies employed by the $\text{OR}_{2\text{groups}}$ and the $\text{DOEL}_{2\text{groups}}$, estimation errors of the discussed two cases are different for them. Here $\hat{\boldsymbol{ae}}_k = (\hat{AE}_k^0, \hat{AE}_k^1)$ is used to denote correlated error results. Specifically, for the method $\text{OR}_{2\text{groups}}$, no matter which base learner in it makes the wrong judgement, the absolute error between the final prediction age and the ground truth age is always one year. For convenience, possible error terms are recorded as follows,

$$\hat{\boldsymbol{ae}}_k = \begin{cases} (1,0), & \text{if } 1 \le k \le y_i \\ (0,1), & \text{if } y_i + 1 \le k \le K. \end{cases} \tag{18}$$

Furthermore, the error expectation (EE) of the $\text{OR}_{2\text{groups}}$ can be calculated,

$$EE_1 = \sum_{k=1}^{K} \hat{\boldsymbol{p}}_k \hat{\boldsymbol{ae}}_k^{\mathrm{T}} = \frac{K}{2^K}. \tag{19}$$

On the contrary, for the $\text{DOEL}_{2\text{groups}}$, different learners making a wrong decision will lead to different error conditions. These conditions include

$$\hat{\boldsymbol{ae}}_k = \begin{cases} (0,0), & \text{if } 1 \le k \le y_i - 2 \text{ or } y_i + 2 \le k \le K \\ (2,0), & \text{if } k = y_i - 1 \\ (1,0), & \text{if } k = y_i \\ (0,1), & \text{if } k = y_i + 1. \end{cases} \tag{20}$$

Then the error expectation of the $\text{DOEL}_{2\text{groups}}$ can be computed by

$$EE_2 = \sum_{k=1}^{K} \hat{\boldsymbol{p}}_k \hat{\boldsymbol{ae}}_k^{\mathrm{T}} = \frac{4}{2^K}. \tag{21}$$

When it comes to the method $\text{DOEL}_{3\text{groups}}$, a three-dimensional vector, $\hat{\boldsymbol{p}}_k = (\hat{P}_K^0, \hat{P}_k^1, \hat{P}_k^2)$, is used to stand for probabilities of three possible cases with the base learner BL_k $(1 \leq k \leq K-1)$. Elements in this vector are probabilities of the target face is categorized into $G_{k,0}$, $G_{k,1}$ or $G_{k,2}$ respectively. In theory, $\hat{\boldsymbol{p}}_k = (1/3^{K-1}, 1/3^{K-1}, 1/3^{K-1})$. Accordingly, a three-dimensional vector $\hat{\boldsymbol{ae}}_k = (\hat{AE}_k^0, \hat{AE}_k^1, \hat{AE}_k^2)$ is used to denote error terms about three cases. Different circumstances cover

$$\hat{\boldsymbol{ae}}_k = \begin{cases} (0,0,0), & \text{if } \ 1 \leq k \leq y_i - 2 \text{ or } y_i + 1 \leq k \leq K-1 \\ (0,1,0), & \text{if } \ k = y_i - 1 \\ (1,0,0), & \text{if } \ k = y_i. \end{cases} \tag{22}$$

Then, the error expectation of the $\text{DOEL}_{3\text{groups}}$ is able to be obtained, which is

$$EE_3 = \sum_{k=1}^{K-1} \hat{\boldsymbol{p}}_k \hat{\boldsymbol{ae}}_k^{\mathrm{T}} = \frac{2}{3^{K-1}}. \tag{23}$$

Since the considered largest human age K is usually a big value (e.g., K = 100), it is easy to reach the conclusion that $EE_3 < EE_2 < EE_1$.

Then another more complex situation that at least two base learners in an ensemble misjudge is taken into consideration. In fact, there are too many possible combinations of false base learners under this situation and it is infeasible to enumerate all of them here. Fortunately, an overview of error expectations of compared three methods is obtained after a large amount of theoretical deduction. Generally speaking, for these three age estimation approaches, their respective error expectations almost increase with the number of false base learners. Nevertheless, some exceptions occur with the $\text{OR}_{2\text{groups}}$. To be more specific, for that method, if indexes of wrong learners lie in two sides of the ground truth age class, then the estimation error will be reduced. In particular, when the number of those wrong base learners on each side are the same, the error decreases to zero. On the other hand, for the proposed two ensemble methods, the $\text{DOEL}_{2\text{groups}}$ and the $\text{DOEL}_{3\text{groups}}$, their individual error expectations are also partially correlated with the distance between the indexes of false learners and the ground truth age. When the distance is long, then their estimation errors tend to be small, otherwise the opposite. A similar trend can be observed in the previous situation of existing no more than one wrong base learner, referring to (20) and (22). More importantly, the $\text{DOEL}_{3\text{groups}}$ always produces lower error expectation than the $\text{DOEL}_{2\text{groups}}$ when they have the same number of mistaken base learners and this superiority is verified by comparative experiments in Section IV-E.

From above detailed comparison analyses, it can be concluded that the proposed two methods own better interpretability in theory and robustness in performance than the conventional approach $\text{OR}_{2\text{groups}}$.

IV. Experiments

A. Databases

To evaluate the performance of the proposed two methods, we performed experiments on four representative databases for chronological or apparent age estimation. These datasets are different in shooting environments or the size of contained facial images.

MORPH II contains about 55,000 facial images taken under a controlled environment from more than 13,000 subjects [48]. A problem of this set is the distributions of gender and race are both uneven. To alleviate the unevenness, we followed the operations employed by the previous work [49] to split all faces into three non-overlapped subsets S1, S2 and S3. Experiments on them were repeated twice:(a) training on S1 and testing on S2 + S3, (b) training on S2 and testing on S1 + S3. Then the average of them was recorded.

FG-NET includes 1,002 facial images of 82 subjects [50]. Note that images within this database are all captured under wild environments. In experiments, we took the leave-one person-out (LOPO) strategy following [32], [46] and reported the average results on all splits.

AgeDB has 16,488 faces belonging to 568 subjects [51]. These images are also shot in wild environments. For evaluation, we adopted five-fold cross-validation.

Chalearn LAP 2015 is a competition database for apparent age estimation [52]. It involves 4,691 images captured under non-controlled environments, where 2,476 images for training, 1,136 images for validation and 1,079 images for testing purposes, respectively. Each face is given to at least 10 users for predicting, then the average age and related standard deviation of these predictions are taken as its labels.

IMDB-WIKI may be the largest open age dataset which comprises 523,051 facial images. These images are crawled from the celebrities in IMDb and Wikipedia and their age labels are obtained through subtracting the year of the photo taken from the birth of the corresponding celebrity. Thus the accuracy of age information in this database can not be vouched and it is not suitable for evaluation. Here we removed those images with no face or multi-faces and used the remained about 300 thousand faces for model's pretraining in apparent age estimation experiments, following the advice of [53] and [46].

B. Evaluation Metrics

For chronological age estimation, to evaluate the performance of involved prediction methods, two widely used metrics were considered: the Mean Absolute Error (MAE) [12], [54] and the Cumulative Score (CS) [22]. Among them, MAE is used for measuring the distortion between the predicted age and the true age, and it can be computed through

$$MAE = \frac{1}{M} \sum_{i=0}^{M-1} \left| \hat{y}_i - y_i \right|, \tag{24}$$

where M is the number of testing faces. It is obvious that a lower MAE means a better performance. The other metric CS

measures the estimation accuracy under an acceptable error level and it can be calculated by

$$CS(l) = \frac{M_l}{M} \times 100\%, \tag{25}$$

where M_l is the number of faces whose estimation error is no more than l. Concerning this metric, when a specific error level l is decided, a larger value of CS indicates a better performance.

In addition, we proposed another metric, Inconsistency Ratio (IR), to quantitatively demonstrate that the consistency condition required by methods in [5]–[7] can not be satisfied all the time in practice. This metric operates in

$$IR = M_T / M, \tag{26}$$

where M_T is the happening times of broken consistency during the evaluation on testing data.

On the other hand, for apparent age estimation, we adopted the ϵ-error for quantitative measurement. This error is measured as

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{27}$$

where x is the predicted age, μ and σ are the average and the standard deviation of human labels, respectively. The reported ϵ-error was the mean over all facial images. Its value ranges between 0 and 1, of which lower numbers represent better predictions.

C. Experimental Settings

For facial feature extraction, the ResNet18 and the ResNet101 [55] network structures were employed in the proposed methods. Specifically, we removed the FC layer in the original network and used the remained convolutional layers to automatically extract aging related features. Note that other popular networks are applicable here.

No matter before the training or the testing process, first we followed the preprocessing operations in [32], which include aligning and resizing each facial image in sequence before feeding it into the estimation model. After that, we got a new version of the original face with aligned eyes and fixed image size of 256×256 pixels. In addition, regular data augmentation involving random cropping, flipping and rotating were made to original training samples. Then, the stochastic gradient descent (SGD) algorithm with a weight decay of 0.0005 and a momentum of 0.9 was adopted to train the estimation model, of which the batch size was set to 32. The initial learning rate was 0.001 and it reduced by a factor of 10 when the model was at the boundary of overfitting. The reduction was confined to three times. We realized the proposed age estimation methods within the pytorch framework and run experiments on a desktop computer with an Intel Core i7 CPU and two NVIDIA GTX1080 GPUs.[1]

D. Weighting Discussion

To learn an appropriate estimation model based on the proposed $\text{DOEL}_{3\text{groups}}$, an important setting is the ground truth weighting assignment of three age groups for each base learner. According to the constraints in (8) and (9), we set up four possible weighting combinations, which were given in Table I. In order to save the computation time, a lighter version of the proposal based on ResNet18 was used here. Experiments were conducted on MORPH II and AgeDB databases after considering their relatively large sizes and different shooting environments. Seeing from the results, we found that centralized probability distributions performed better than even ones. It is understandable since larger probability margin will more clearly guide the model to find out the true age group that target face belonging to. In addition, we also observed that results under different weight settings jointly formed a "V" shape, rather than a monotonous one. The reason behind this deserves to be further explored. Finally, we chose the weight setting which did well on both datasets.

TABLE I
COMPARISONS AMONG DIFFERENT WEIGHT SETTINGS

$\boldsymbol{p}_i^k$	MAE; cs($l=5$); CS($l=10$)	
	on MORPH II	on AgeDB
$\begin{cases}(0,0.1,0.9)\\(0.05,0.9,0.05)\\(0.9,0.1,0)\end{cases}$	2.82; 86.9%; 98.7%	5.83; 57.6%; 84.7%
$\begin{cases}(0,0.2,0.8)\\(0.1,0.8,0.1)\\(0.8,0.2,0)\end{cases}$	**2.81**; **87.0%**; **98.8%**	**5.80**; **58.2%**; **85.2%**
$\begin{cases}(0,0.3,0.7)\\(0.15,0.7,0.15)\\(0.7,0.3,0)\end{cases}$	2.83; 86.5%; 98.7%	5.84; 57.7%; 84.7%
$\begin{cases}(0,0.4,0.6)\\(0.2,0.6,0.2)\\(0.6,0.4,0)\end{cases}$	2.91; 85.6%; 98.4%	6.12; 55.6%; 83.4%

E. Comparisons Among Different Model Formulations

Although age estimation should be an ordinal regression problem essentially, previous works which claimed excellent performance employed different model formulations. As their success may owe to several aspects (e.g., well preprocessing of facial images or powerful CNN frameworks for feature extraction), we were interested about the pure effectiveness brought by different modellings. Therefore, fair and comprehensive comparisons were made among them. Specifically, three popular types of model formulations on the age estimation task were considered: the classification, the regression and the ordinal regression. For the classification type, four prediction models were built. The first one was a typical classification model which treated each age label as an independent class and the predictive age was the index of the class that got the maximum prediction probability. For convenience, this model was called the "Argmax". The second model was a little different from the last one because the human age was inferred by calculating the expectation value over the output probabilities of all possible age classes, and it was

[1] https://github.com/Xiejiu/second_age_estimation

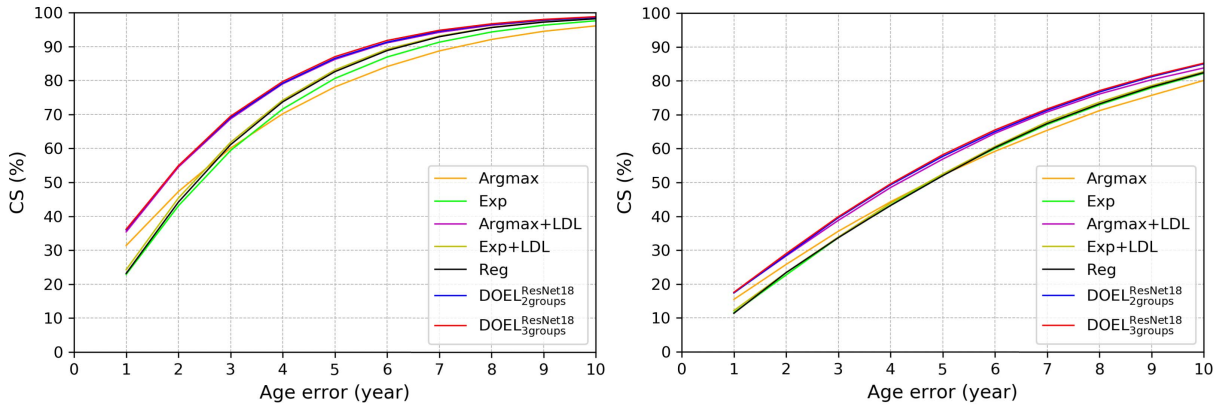

Fig. 4. CS curves of different model formulations. (a) Comparison results on MORPH II database. (b) Comparison results on AgeDB database.

TABLE II
COMPARISONS AMONG DIFFERENT MODEL FORMULATIONS

Methods	MAE	
	on MORPH II	on AgeDB
Argmax	3.54	6.66
Exp [53]	3.08	6.06
Argmax+LDL [45]	2.88	6.01
Exp+LDL [37]	2.88	5.90
Reg	2.93	5.94
$\text{DOEL}^{\text{ResNet18}}_{\text{2groups}}$	2.85	5.83
$\text{DOEL}^{\text{ResNet18}}_{\text{3groups}}$	**2.81**	**5.80**

referred to as the "Exp" [53]. In addition, the Label Distribution Learning (LDL) [44] technique was also applied to the previous two models and integrated new versions were named the "Argmax + LDL" [45] and the "Exp + LDL" [37], respectively. For the kind of regression, a conventional regression model was constructed and we used the "Reg" to represent it. Our two proposed methods both belong to the ordinal regression type. Here we also used the ResNet18 for facial feature extraction in each model. Experimental results were given in Table II and Fig. 4.

At the first glance of these results, all age estimation methods employing different model formulations uniformly performed worse on AgeDB than on MORPH II. This performance degeneration should be attributed to the wild capturing environment in the former database, since faces in it are with large variations of poses, expressions and lighting, whereas the ones in the latter are not. Then more careful comparisons of respective performance under different metrics were made. Under the MAE metric, it could be found that: (1) Although the conventional classification model (the "Argmax") performed worse than the regression model (the "Reg"), performance of the former could be improved by using the expectation of all possible age classes as the estimation age or incorporating the LDL technique in the training process. Note that above two operations could be superimposed to further enlarge the improvement. For example, the deeply modified version "Exp + LDL" even surpassed the "Reg"; (2) The proposed two methods (the $\text{DOEL}^{\text{ResNet18}}_{\text{2groups}}$ and the $\text{DOEL}^{\text{ResNet18}}_{\text{3groups}}$) both based on ordinal regression modelling achieved better results than other methods, of which the one with three-group classification performed the best. In the experiments on MORPH II, although the improvement of the proposed approaches compared with [45] and [37] was relatively small, the significant superiority of our proposed methods was showed on the AgeDB database. Note that the second database is wild while the first one is controlled, thus age prediction on the AgeDB database is more challenging and the experimental results on it reveal the effectiveness of a specific estimation approach more veritably. On the other hand, if just taking the CS metric into consideration, we would have almost the same observations like those when only using the MAE metric. Besides, a new difference also exists. Specifically, two modified versions, the "Exp" and the "Exp + LDL", both deteriorated the performance of the original one (the "Argmax") when acceptable error level l was relatively small. However, this situation reversed when it came to higher levels. On the other hand, another optimized version, the "Argmax + LDL", consistently outperformed the original version and the regression model under all considered error levels, but was still a little inferior to the two ordinal regression models. Note that these two findings had not been reported in the past related work [56] which also studied on different model formulations. We hope the findings may provide some guidance for model selection when a specific error tolerance is given.

Above comparison experiments demonstrated the ordinal regression modelling is more suitable than others for the age estimation task.

F. $OR_{2groups}$ vs. $DOEL_{2groups}$ vs. $DOEL_{3groups}$

Although theoretical comparison analyses among the three ordinal regression based methods have been given in Section III-D, we are also concerned with their respective performance in the real world. Thus, several comparison experiments on AgeDB dataset were carried out.

TABLE III
COMPARISONS AMONG DIFFERENT ORDINAL REGRESSION MODELLINGS

Methods	on AgeDB	
	MAE	IR
$OR^{ResNet18}_{2groups}$	5.83422	30/16488
$DOEL^{ResNet18}_{2groups}$	5.83408	-
$DOEL^{ResNet18}_{3groups}$	**5.79756**	-
$DOEL^{ResNet101}_{2groups}$	5.73700	-
$DOEL^{ResNet101}_{3groups}$	**5.69204**	-

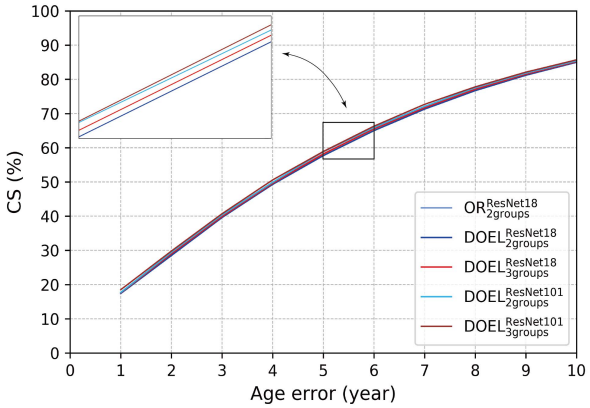

Fig. 5. CS curves of different ordinal regression modellings.

For convenience, the ResNet18 network was employed again to extract aging related facial features in each method. Analyses in Section III-D have explained that, main difference between the $OR^{ResNet18}_{2groups}$ and the $DOEL^{ResNet18}_{2groups}$ is their different aggregating manners on results produced by those individual classifiers, whereas their model structures and loss functions are both the same. Therefore, in our experiments, only one network model for them two was trained actually but different aggregation strategies were leveraged to predict human age in the testing phrase. This operation made the comparison between them more fair. Concerning the $DOEL^{ResNet18}_{3groups}$, another independent model was constructed and learned. Experimental results were presented in Table III and Fig. 5.

For comparisons between the $OR^{ResNet18}_{2groups}$ and the $DOEL^{ResNet18}_{2groups}$, when we only considered the MAE and the CS metrics, it could be found that although these two age estimation methods shared the same network model in practice, result of the latter was still slightly better than that of the former. However, it was the three grouping version of the proposed ensemble methods, the $DOEL^{ResNet18}_{3groups}$, that provided the best performance. These findings were consistent with related theoretical analyses in previous Section III-D, and both of them proved that a three-group classification is superior than a two-group version. In addition, enough attention should be paid to their performance under the IR metric. As discussed in Section III-D, the aggregation strategy of the $OR^{ResNet18}_{2groups}$ has some theoretical defects. To be more specific, the logicality of that strategy calls for consistent predictions from individual classifiers. However, in practical experiments, this consistency requirement was not satisfied for dozens of times. On the contrary, the proposed ensemble methods enable the fusion of individual estimations at any circumstances.

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON MORPH II DATABASE

Methods	MAE
Xing et al. [56]	2.96
D2C [43]	3.06
Tan et al. [46]	2.86
Yoo et al. [31]	2.89
Wan et al. [30]	2.93
Sheng et al. [47]	2.98
Xie et al. [32]	2.81
$DOEL^{ResNet101}_{2groups}$	2.79
$DOEL^{ResNet101}_{3groups}$	**2.75**

G. Performance Under Different Network Depths

In previous experiments, to save computation and time costs, the ResNet18 was employed to extract facial features. Note that an important advantage of the ResNet framework is that it enables the stacking of more layers while the modified one is still easy to be optimized. Moreover, comprehensive experiments had showed that a deeper structure leads to accuracy improvements in image classification tasks [55]. We were interested in whether a deeper network would do age estimation a favor or not. Therefore, additional experiments were conducted through replacing the original ResNet18 with the ResNet101 in the proposed two methods and corresponding results on AgeDB database were given in Table III and Fig. 5. Results clearly showed that a deeper network could further improve the performance of the two proposed methods in age estimation. We argued that this should be owed to the more accurate representations of aging related facial features after increasing the network's depth.

H. Comparisons With State-of-the-Art Methods

Finally, we decided to compare the two proposed ensemble methods with other state-of-the-art age estimation approaches. Experiments were first carried on two widely evaluated chronological age databases, MORPH II and FG-NET. Related results were given in Tables IV and V, respectively. Our approaches showed competitive performance on both datasets. Specifically, on MORPH II database, to the best of our knowledge, this is the first time that an age estimation model succeeds in reducing MAE below 2.80 under the "S1-S2-S3" protocol without pretraining on extra face data. Note that DHAA reported a lower MAE value than our methods recently. However, with the same backbone Resnet18 for feature extraction and the same baseline for comparison, the proposed approach acheived a larger improvement than DHAA (0.27 vs. 0.206). On the other hand, our methods

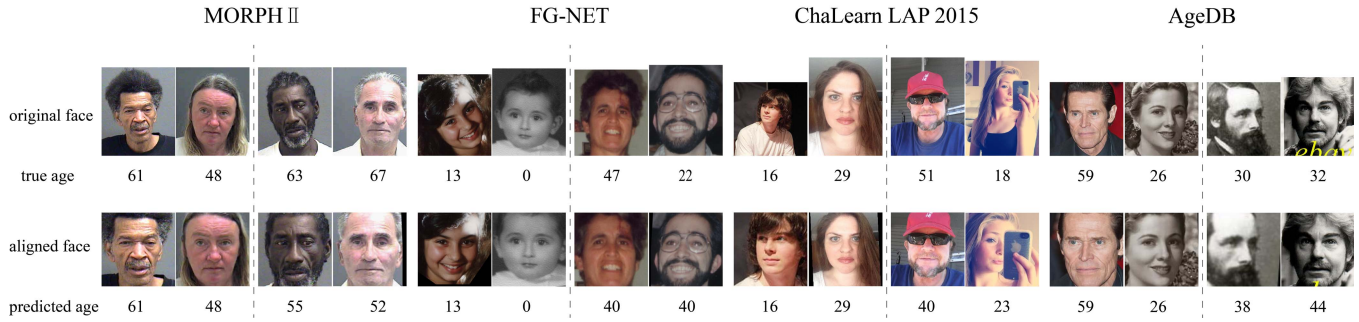

Fig. 6. Some age estimation results by the $\text{DOEL}_{3\text{groups}}^{\text{ResNet101}}$. For results under a specific face aging database, on the left side of the dashed line are good ones and on the other side are bad ones.

TABLE V
COMPARISONS WITH STATE-OF-THE-ART METHODS ON FG-NET DATABASE

Methods	MAE	CS($l = 5$)	CS($l = 10$)
DEX [53]	4.63	-	-
Liu et al. [57]	3.93	76.0%	91.0%
LSDML [58]	3.92	75.0%	89.0%
ODFL [42]	3.89	80.0%	91.0%
Tan et al. [46]	4.34	76.0%	86.0%
Shen et al. [47]	3.85	80.6%	-
Pan et al. [40]	4.10	78.0%	88.0%
Xie et al. [32]	3.58	78.3%	92.3%
ODL [59]	3.71	81.0%	**96.0%**
SADAL [60]	3.68	-	-
DHAA [61]	3.72	-	-
$\text{DOEL}_{2\text{groups}}^{\text{ResNet101}}$	3.52	81.1%	92.6%
$\text{DOEL}_{3\text{groups}}^{\text{ResNet101}}$	**3.44**	**82.7%**	93.0%

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS ON CHALEARN LAP 2015 DATABASE

Methods	Validation Set		Test Set		Num.
	MAE	ϵ-error	MAE	ϵ-error	of Networks
DEX [53]	3.25	0.28	-	0.265	20
Tan et al. [46]	3.21	0.28	2.94	0.264	8
DHAA [61]	3.052	0.265	-	0.252	1
BridgeNet [62]	2.98	0.26	2.87	0.255	1
$\text{DOEL}_{2\text{groups}}^{\text{ResNet101}}$	**2.933**	**0.258**	**2.713**	**0.247**	1
$\text{DOEL}_{3\text{groups}}^{\text{ResNet101}}$	2.965	0.259	2.726	0.249	1

also made new performance records during evaluations on FG-NET dataset. Note that the algorithms of Wan *et al.* [30] and Xie and Pun [32] mainly benefited from considering the influence of gender and race factors. However, the proposed methods still did better than them without taking those factors into consideration, although it would be operable if given related attribute labels.

Then we conducted apparent age estimation on Chalearn LAP 2015 dataset. For fair comparison with other methods, few tricks adopted in previous works [53], [61], [62] were also used in our training and testing processes. To be more clear, we pretrained our models on the filtered IMDB-WIKI data before respective finetuning on apparent age database. Note that all involved competitors use the same additional dataset for pretraining. Besides, in the testing phase, when given a target face no matter from the validation subset or the test subset, we fed its four corner and the central crop plus their horizontally flipped versions to the trained model. Their output were averaged and then taken as the final prediction. Comparison results were given in Table VI. Our methods again outperformed other approaches. However, a notable thing is that, for the two proposed methods, the two-group version did a little better than the three-group one, which is different from their consistent performance on three chronological age datasets. The reason lies in the pretraining phase. It was found that the $\text{DOEL}_{2\text{groups}}^{\text{ResNet101}}$ already got better performance than the $\text{DOEL}_{3\text{groups}}^{\text{ResNet101}}$ after their respective pretraining on IMDB-WIKI. As mentioned in the description part of this database, ages of faces were obtained by web crawling techniques, which inevitably brought many mistakes. In other words, this dataset is noisy in age label. We argue that this noisy condition badly influence the capability of $\text{DOEL}_{3\text{groups}}^{\text{ResNet101}}$ more than that of $\text{DOEL}_{2\text{groups}}^{\text{ResNet101}}$. Thus it is suggested to choose the two-group version of the proposals when the pretraining data include noisy labels.

No matter on controlled or wild databases, no matter making chronological or apparent age estimation, our methods all made remarkable performance. As mentioned before, age estimation is an ordinal regression problem essentially. Following this fact, we set up two special ensembles for age prediction, both of which implicitly encoded the ordinal relationship. The effectiveness of the two proposed approaches partially proves the reasonableness and significance of treating a specific problem as it should be in essence. In addition, some visual estimation results were shown in Fig. 6 for qualitative evaluation, including good and bad ones. Among them, those results of typical bad predictions deserve more attention. We found that some faces were not well aligned when they were fed into the estimation system. In addition, some facial images have different expressions, poses and bad illuminations, and few of them even contain occlusion. All these negative factors may account for the bad estimations.

To further improve the prediction performance, more effective image preprocessing methods should be explored and integrated in our age estimation system in the future work.

V. Conclusion

In this paper, we propose two ensemble learning methods both based on ordinal regression modelling for age estimation. They show good interpretability in theory as well as excellent performance in practice. Concerning the future work, since it is hard to collect facial images with exact age labels, we are interested in learning an effective prediction model on a small amount of training data. In addition, several aspects related to the current topic also deserve to be further studied. For instance, cross-population age prediction and face aging synthesis, both of which are open problems.

Acknowledgment

The authors appreciate W. Li for his selfless sharing of pretraining details.

References

[1] I. Dotu, M. A. Patricio, A. Berlanga, J. García, and J. M. Molina, "Boosting video tracking performance by means of tabu search in intelligent visual surveillance systems," *J. Heuristics*, vol. 17, no. 4, pp. 415–440, Aug. 2011.

[2] C. Shan, F. Porikli, T. Xiang, and S. Gong, Eds., *Video Analytics for Business Intelligence* (Studies in Computational Intelligence), vol. 409. Berlin, Germany: Springer, 2012.

[3] F. Fernandez-Navarro, P. Campoy-Munoz, M.-D. L. Paz-Marin, C. Hervas-Martinez, and X. Yao, "Addressing the EU sovereign ratings using an ordinal regression approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2228–2240, Dec. 2013.

[4] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Berlin, Germany: Springer-Verlag, 2011.

[5] 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. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 585–592.

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

[7] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 742–751.

[8] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[10] M. Perez-Ortiz, P. A. Gutierrez, and C. Hervas-Martinez, "Projection-based ensemble learning for ordinal regression," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 681–694, May 2014.

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

[12] A. Lanitis, C. Taylor, and T. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[13] 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)*, Santiago, Chile, Dec. 2015, pp. 252–257.

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

[15] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 267–273.

[16] 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 (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 801–809.

[17] Z. Huo *et al.*, "Deep age distribution learning for apparent age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 722–729.

[18] M. Uricar, R. Timofte, R. Rothe, J. Matas, and L. V. Gool, "Structured output SVM prediction of apparent age, gender and smile from deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 730–738.

[19] 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.

[20] A. Gunay and V. V. Nabiyev, "Automatic age classification with LBP," in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, Istanbul, Turkey, Oct. 2008, pp. 1–4.

[21] F. Gao and H. Ai, "Face age classification on consumer images with Gabor feature and fuzzy LDA method," in *Proc. Int. Conf. Biometrics (ICB)*, Alghero, Italy, Jun. 2009, pp. 132–141.

[22] 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 (MM)*, Santa Barbara, CA, USA, Oct. 2006, pp. 307–316.

[23] 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.

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

[25] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer, "Locally adjusted robust regression for human age estimation," in *Proc. IEEE Appl. Comput. Vis. Workshops*, Copper Mountain, CO, USA, Jan. 2008, pp. 1–6.

[26] S. Yan, H. Wang, X. Tang, and T. S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[27] 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. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 657–664.

[28] 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, Sep. 2010.

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

[30] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2531–2541, Sep. 2018.

[31] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 808–812, Jun. 2018.

[32] J.-C. Xie and C.-M. Pun, "Chronological age estimation under the guidance of age-related facial attributes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2500–2511, Sep. 2019.

[33] 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.

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

[35] Y. Chen, Z. Tan, A. P. Leung, J. Wan, and J. Zhang, "Multi-region ensemble convolutional neural networks for high accuracy age estimation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., Sep. 2017.

[36] S. Taheri and Ö. Toygar, "On the use of DAG-CNN architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, Feb. 2019.

[37] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, Jul. 2018, pp. 712–718.

[38] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, Jun. 2018, pp. 1078–1084.

[39] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12587–12596.
[40] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5285–5294.
[41] Y. Liu, A. W. K. Kong, and C. K. Goh, "A constrained deep neural network for ordinal regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 831–839.
[42] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep feature learning for facial age estimation," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 157–164.
[43] K. Li, J. Xing, W. Hu, and S. J. Maybank, "D2C: Deep cumulatively and comparatively learning for human age estimation," *Pattern Recognit.*, vol. 66, pp. 95–105, Jun. 2017.
[44] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[45] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[46] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2610–2623, Nov. 2018.
[47] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2304–2313.
[48] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Southhampton, U.K., Apr. 2006, pp. 341–345.
[49] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Singapore, Nov. 2014, pp. 144–158.
[50] *The FG-NET Aging Database*. Accessed: Nov. 5, 2019. [Online]. Available: https://fipa.cs.kit.edu/433_451.php
[51] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1997–2005.
[52] 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 (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 243–251.
[53] 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.*, vol. 126, nos. 2–4, pp. 144–157, Apr. 2018.
[54] 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.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[56] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Pattern Recognit.*, vol. 66, pp. 106–116, Jun. 2017.
[57] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Pattern Recognit.*, vol. 66, pp. 82–94, Jun. 2017.
[58] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 292–305, Feb. 2018.
[59] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep learning for facial age estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 486–501, Feb. 2019.
[60] P. Sun, H. Liu, X. Wang, Z. Yu, and S. Wu, "Similarity-aware deep adversarial learning for facial age estimation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shanghai, China, Jul. 2019, pp. 260–265.
[61] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Deeply-learned hybrid representations for facial age estimation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Macao, China, Aug. 2019, pp. 3548–3554.
[62] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1145–1154.

Jiu-Cheng Xie received the M.Sc. degree in pattern recognition and intelligent system from the Nanjing University of Posts and Telecommunications, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer and Information Science, University of Macau, China. His current research interests include biometrics, computer vision, and machine learning.

Chi-Man Pun (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau in 1995 and 1998, respectively, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2002. He is currently an Associate Professor and the Head of the Department of Computer and Information Science, University of Macau. He has investigated several funded research projects and authored or coauthored more than 150 refereed scientific articles in international journals, books, and conference proceedings. His research interests include digital image processing, multimedia security, pattern recognition, and computer vision. He is also a professional member of the ACM. He has also served as the editorial member/referee for many international journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.