Received December 9, 2019, accepted December 30, 2019, date of publication January 6, 2020, date of current version February 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964281

Rectified Softmax Loss With All-Sided Cost Sensitivity for Age Estimation

DAXIANG LI1,2, XUAN MA1, YAQIONG REN1, AND SHYH-WEI TENG3

1School of Telecommunication and Information Engineering, Xi’an University of Posts and Telecommunications, Xi’an 710121, China
2Key Laboratory of Electronic Information Field Inspection and Application Technology, Ministry of Public Security, Xi’an 710121, China
3Faculty of Science and Technology, Federation University Australia, Mount Helen, VIC 3342, Australia

Corresponding author: Xuan Ma (maxuan77777777@gmail.com)

This work was supported in part by the International Cooperation & Exchange and Innovation & Entrepreneurship Dual Tutor Foundation of Shaanxi under Grant 2017KW-013 and Grant 2019JM-604.

I. INTRODUCTION

Age estimation means an accurate age (group) should be determined considering a given face image. Because this technology has important application value in the fields of human-computer interaction, intelligent marketing, intelligent monitoring and criminal reconnaissance, it has become a research hotspot in computer vision in recent years. However, due to the influence of growth environment, lifestyle and ethnic genetic diversity, it is still very challenging to accurately estimate the facial age.

In the machine learning methods, age estimation is usually divided into two parts: facial age representation and age determination. Age representation includes shallow characterization of AAM [1]–[3], LBP [4], BIF [5], [6], and the like. After extracting the appropriate age representation for the face image, some age determination methods are used to predict the corresponding age, such as KNN [6], quadratic function [6], [7], SVM [8], [9], SVR [8], [9], and the like. However, there are some drawbacks: 1. The design of manual age representation is quite cumbersome. 2. The traditional generative model and discriminative model cannot fully obtain the benefits brought by the big data era, namely with the increase in the amount of data, the performance improvement is not obvious. 3. Because not an end-to-end design, it is quite time-consuming to estimate the age for an unacquainted facial image.

Inspired by the excellent performance of deep learning in image classification, many researchers have begun to apply it to facial age estimation. According to different algorithm principles, current methods are mainly divided into regression and multi-classification. In the regression method, the researchers regard the age label as a continuous value. As a result, the output layer of Convolutional Neural Networks (ConvNets) contains only one neuron, so the output is the specific age. In the multi-classification method, the researchers regard the age label as a discrete category. As a result, the number of neurons in the output layer of the ConvNet is consistent with the number of age groups in the experimental dataset, so the output is the specific age category.

INDEX TERMS Age estimation, cost sensitivity, softmax loss.
In this paper, age estimation will be studied based on multi-classification ConvNet.

When using the classification ConvNet for age estimation, the previous method usually uses the plain ConvNet model or the ConvNet model originally used for other image classification tasks (such as: VGG16 [11]), and the improvement in performance is mainly due to fine-tuning on the age image dataset after pre-training in other types of image sets. Because the shallow representation of most images and even all face images in ConvNet can be shared, transfer learning reduces the researchers’ unnecessary pre-training work for ConvNet and often enjoys the performance benefits of increased sample size. However, the obtained model does not satisfy Ockham’s Razor. At this time, it is necessary to internally analyze the age estimation problem and design the ConvNet model in a targeted manner. Since age estimation is a typical cost-sensitive computer vision problem, this method came into being.

Using cost-sensitive theory for age estimation, there are two main problems: 1. Class imbalance problem, as shown in Fig. 1, it is the statistical distribution of the sample size of each age group in the two largest face dataset AFAD [10] and MORPH [12], showing that the sample size is extremely unbalanced. If the classifier is trained directly using class-unbalanced data, the classifier will always tend to misclassify categories that contain fewer samples into categories that contain more samples. In real life, the probability of occurrence of some categories is quite different. For example, the probability of young people appearing in shopping plazas is always higher than that of the elderly and children, but we should still try to make each class have the same effect. In this paper, we name this type of problem as Exterior Cost Sensitivity (ECS), which means that it is interpreted as the cost sensitive problem caused by people collecting samples.

2. Uneven misclassification cost problem. That is, when the age estimation is regarded as a multi-classification problem, the problem of the misclassification between different classes should be considered. For example, the cost of misclassifying a 22-year-old to 23 and misclassifying it as 50 is not the same, as shown in Fig. 3. Similarly, we name it Interior Cost Sensitivity (ICS), which means that it is inherent when age estimation is regarded as a multi-classification problem.

In view of ECS, ICS and the shortcomings of the previous methods in age estimation, this paper embeds cost sensitivity into the deep learning framework during the training phase, and proposes a rectified softmax loss with all-sided cost sensitivity. The core idea of the algorithm is firstly to establish a loss function for each age category in order to reduce the influence of the unbalanced number of samples of different categories on the performance of the whole classification algorithm. In addition, a cost vector is constructed to reflect the cost difference of misclassification among different classes, and a cost sensitive error function is proposed. Finally, the proposed two loss functions are integrated to construct a rectified softmax loss for the CNN model to comprehensively solve the cost sensitivity problem in age estimation. The whole algorithm flow is shown in Fig. 2.
used it as the parameter description of facial images, and then literatures are developed based on AAM. Lanitis et al. [2] reflect the growth process of an individual intuitively, many shape and grey-level appearance. Because the appearance can be considered so comprehensive in the age estimation related studies (i.e., marginal fisher analysis and locality sensitive hashing), to the best of my knowledge, this is the first time that the cost sensitivity problem has been acknowledged by the researchers.

To sum up, the contribution of this paper is summarized as follows:

1) For ECS, we built loss function for each category to eliminate maximally the impact of class-imbalance on the total loss.

2) For ICS, we constructed a novel cost matrix to reflect the difference in misclassification cost based on the Desired Class Maximum Principle (see section III.C). In addition, we also verify its advancement with some examples.

3) For ECS and ICS, we integrated a considerate cost sensitivity loss function for age estimation and prove its practicability with relevant theories. To the best of my knowledge, this is the first time that the cost sensitivity problem has been considered so comprehensive in the age estimation related methods to date.

4) The corresponding Back Propagation (BP) algorithm is ameliorated to analysis the influence of the change of loss function toward the training process of network.

5) Experiments were carried out on the two interracial datasets to reveal the generalization of the proposed method.

II. RELATED WORKS

Before describing our method systematically, we first briefly introduce age estimation and cost sensitivity learning methods in the previous literature.

A. AGE ESTIMATION

Even though age estimation has not received as much attention as face recognition, it has been studied for more than 20 years. The mainstream methods could be divided into the following aspects.

1) POINT DISTRIBUTION MODEL

Active Appearance Models (AAM) was described by Cootes et al. in 1998 [1], which is a statistical model of the shape and grey-level appearance. Because the appearance can reflect the growth process of an individual intuitively, many literatures are developed based on AAM. Lanitis et al. [2] used it as the parameter description of facial images, and then tested many classifiers to estimate the age (i.e., Quadratic Function, Shortest Distance Classifier, Multilayer Perceptron and Self Organizing Map). Besides, for people who look similar they designed appearance-specific classifier, and similarly, age-specific classifier. AGing pattErn Subspace (AGES) was introduced by Geng et el [3], and it adopts AAM to extract facial feature likewise. In the training stage, Principal Component Analysis (PCA) is used to construct the global aging pattern subspace, which is defined as the sequence of all training images sorted chronologically. For a previously unseen facial image, its aging pattern is selected by the projection in the subspace with the minimum reconstruction error criterion, and the position of image obtained iteratively in the proper aging pattern will indicate its age.

2) MANIFOLD LEARNING METHOD

“Manifold” is a structure which is locally homomorphic with Euclidean space. The main idea of manifold learning is to project the data located in the high-dimensional space into the low-dimensional space so that the low-dimensional data could reflect some essential and structural characteristics of the original high-dimensional data. There is an important premise before it could be applied, which is that the high-dimensional data actually is a low-dimensional manifold embedded in high-dimensional space. Naturally, the use of manifold learning for age estimation also acquiesced to this premise, when the so-called high-dimensional data is the face image (the pixel matrix). By regarding the aging process as the low dimensional distribution in a chronological order, Fu et al. [13] employs the linear manifold learning method, i.e., PCA, Neighborhood Preserving Projections (NPP), Locality Preserving Projections (LPP) and Orthogonal Locality Preserving Projections (OLPP), to find a low-dimensional embedding space, and takes the projection of the image in this space as the aging representation, and finally constructs a quadratic function for regression. Two years later, the same age-manifold method was mentioned again by Fu et al. [7] However, the tool of finding effective low-dimensional embedding space was replaced by Conformal Embedding Analysis (CEA), which followed by the improvement on accuracy.

3) THE SHALLOWER BIO-INSPIRED NETWORK

Riesenhuber et al. [14] described a 4-layers network (usually called BIF) following the organization of visual cortex, which alternates between a simple (S) layer and a complex (C) layer. The S layer was designed based on the multi-orientation and multi-scale Gabor filters, so as to extract the discriminative low-level features. The C layer shows some tolerance to shift and size by max-pooling operation. Guo et al. [6] used the two layers of BIF to estimate facial age. Specifically, they adopted a standard deviation operation (STD) combined with the max-pooling in layer C to better reveal the local variation of textures on faces. Soon afterwards, he [15] combined BIF with a variety of supervised manifold learning methods (i.e., marginal fisher analysis and locality sensitive
discriminant analysis), and then estimated age according to gender and age group.

4) THE DEEPER BIO-INSPIRED NETWORK (i.e., ConvNet)
Levi and Hassner [16] implemented age classification by mean of the classic Alexnet-like network (5 layers). To my knowledge, it is the first successful application of deep learning in age estimation. In addition, the literature [9], [10], [17], [18], [19], [38], [39], [40] (see section IVE) used ConvNet for age estimation. However, they are all based on one or more baseline ConvNet, and do not consider the problem of class-imbalance and the uneven misclassification costs. In contrast, the proposed network structure modification method can finally learn the robust and discriminative parameters about cost sensitivity.

B. COST SENSITIVITY LEARNING
Cost sensitivity learning located on the algorithm-level is similar to the data-resampling method located on the data-level, both often aimed at solving the class-imbalance problem. Specifically, it distributes different misclassification costs to different classes through cost matrix and/or cost vector. Ting [20] introduced cost sensitivity into the decision tree, and the resulting cost sensitive tree is simpler and more efficient. Cai et al. [21] proposed cost sensitive SVM for text classification. In addition, cost sensitivity learning is particularly widely used in the boosting method [22], [23]. Although the theory of cost sensitivity learning is being adopted by many literatures, its research in deep learning is still quite rare. As far as I know, only [24], [25] are available.

Several cost-sensitive age estimation algorithms had been previously considered: Chang et al. [26] proposed a cost-sensitive ranking algorithm, which embedded cost sensitivity in the ranking model. However, the disadvantage was that the cost sensitivity was only considered in the age prediction stage, so they did not learn enough discriminant and robust parameters; Lu et al. [4] proposed a cost-sensitive LBP algorithm, but because LBP only considers local texture information, the method has limited representation ability. In addition, they only consider the misclassification cost problem and ignore the imbalance of the training sample quantity, so the cost-sensitive age estimation model obtained is incomplete. For the above considerations, this paper proposes a modified softmax loss that works in the training phase for age estimation.

The upcoming part of our proposed work is organized as follows. The proposed approach is illustrated in the next section, where we give a detailed description of our scheme. Section IV is our experimental part. In Section V we discuss the transportability of our method and the future research direction. The final conclusion is located in Section VI.

III. THE PROPOSED APPROACH
In this paper, the cost-sensitive problem in age estimation is solved in the training phase, and our main work focuses on the improvement and innovation of the loss function.
B. CONSTRUCT ECS-SOFTMAX LOSS FUNCTION

In the loss function defined in (5), if the loss of each type of sample is considered separately, the formula can be rewritten as:

\[
E(w, b) = \frac{1}{L} \sum_{j=1}^{n} E_j(w, b) = -\frac{1}{L} \sum_{j=1}^{n} \sum_{i=1}^{n} d_{i,j} \log y_i^j(w, b)
\]

where \(E_j(w, b)\) denotes the total loss of samples of category \(j\), and \(L_j\) denotes the number of samples of category \(j\).

It can be intuitively seen from (6) that, if the dataset is class-imbalance (as shown in Fig. 1), the class containing more samples will have a decisive impact on the total loss, while the class containing fewer samples will have a negligible impact on the loss. As a result, classifier will learn skewed weights and biases in favor of the class containing more samples.

In this paper, in order to solve the ECS problem in age estimation, the loss of each training sample was calculated separately, and ECS-softmax loss function was proposed as follows:

\[
E(w, b)' = \frac{1}{L_j} \sum_{j=1}^{n} E_j(w, b)
\]

Compared with the original softmax loss (see (5) and (6)), ECS-softmax loss is different in that it adopts a divide and conquer strategy for the loss on each category, that is, it uses a divide-and-conquer strategy to construct a loss function for each category to calculate the loss value separately. In this way, regardless of the sample size of each category, all categories have the same impact on the loss of the entire dataset. Further, the impact of the number toward even the performance of classifier could be ignored.

C. CONSTRUCT ICS-SOFTMAX LOSS FUNCTION

1) CONSTRUCT COST MATRIX

In the standard softmax regression function shown in (3), when calculating the posterior probability of each category, the misclassification cost sensitive problem (that is, ICS) was not considered. At this time, a cost matrix should be designed to reflect the misclassification cost difference between different categories according to the age estimation itself, which is a cost-sensitive computer vision problem.

For ICS, the traditional cost matrix follows the reason-ableness condition [25]: 1. The cost of correct classification should always be kept at the minimum; 2. The cost of two classes with relatively far distance from each other in misclassification should be larger than that of two classes with relatively close distance from each other in misclassification. In addition, another condition needs to be satisfied when cost matrix is applied to softmax regression function: when misclassification cost is equal, the cost-sensitive softmax function should be the same as the original softmax function. To satisfy these properties, we propose a preliminary cost matrix:

\[
c' = \begin{bmatrix} c'_{\rho,k} \end{bmatrix}_{n \times n} = [|\rho - k| + 1]_{n \times n}
\]

where \(c'_{\rho,k}\) represents the cost value of class \(\rho\) mistaken for class \(k\).

Fig. 3 illustrate this cost value when the real age is 22-year-old. That is, for instance, the cost of misclassifying someone belongs to age-22 for age-23 is \(|22 - 23| + 1 = 2\) while the cost of misclassifying someone belongs to age-22 for age-50 is \(|22 - 50| + 1 = 29\). This result does make logical sense.

Except that a situation the cost value is not zero located in diagonal even if the classification result is correct, the property of the preliminary cost matrix is similar to the usual absolute cost matrix. (In fact, the minimum value we can allow is 1, and which is not wrong, because on the one hand, when the misclassification costs of all categories are even, ICS-softmax loss changes back to the original softmax loss, and on the other hand, it makes sense for the proposed final cost matrix).

The reason why we do not directly use the preliminary cost matrix as the weight of the logit value is that it does make the parameter in ConvNet embedded with the factor of misclassification cost in the training process, but ultimately it cannot guarantee that the probability value of the desired class will get better gradually. Here, we propose a general rule that we and even all literatures about cost matrix should follow, called Desired Class Maximum Principle (DCMP), which means that all of the cost matrices have only one purpose, namely, to make the logit value of the correct classification larger and larger (reflected by the increasing \(o_\rho\) value), so as to make the probability value of the correct classification larger and larger (reflected by the increasing \(y_\rho\) value). This is not only the core meaning of the proposed cost matrix, but also the difference from the previous cost matrix. Concretely, we classified logit values of both desired and non-desired (i.e., predicted) classes into positive and negative categories, thus derived a new cost matrix \(c\), and whose specific value can be seen in the following (9).

\[
c_{\rho,k} = \begin{cases} \frac{1}{c_{\rho,k}}, & \text{if } o_\rho \geq 0 \\ c_{\rho,k}', & \text{if } o_\rho < 0 \end{cases}
\]

Different from [25] on automatic learning cost matrix, the cost matrix we proposed is a fixed formula, but it is actually verified as the following (See section III.E.A).

2) ICS-SOFTMAX LOSS

In order to completely solve the ICS problem, we combined (8) and (9), so as to embed cost matrix into the softmax regression function and modify the output value of the last layer of ConvNet (as shown in Fig. 4), and calculate the error value of the ith image that originally belonged to class \(\rho\) and
was mistakenly classified as class k:

$$y_k^i (w, b, c) = \frac{e^{w_k^i c} o^i_k (w, b)}{\sum_{k=1}^{n} e^{w_k^i c} o^i_k (w, b)}$$  \hspace{1cm} (10)$$

The overall loss function is as follows:

$$E(w, b, c)'' = \frac{1}{L} \sum_{i=1}^{L} E^i (w, b, c)'' = - \frac{1}{L} \sum_{i=1}^{L} \sum_{k=1}^{n} d_k^i \log y_k^i (w, b, c)' \hspace{1cm} (11)$$

### D. RECTIFIED SOFTMAX LOSS WITH ALL-SIDED COST SENSITIVITY

1) **CS-SOFTMAX LOSS FUNCTION**

In order to deal with ECS and ICS problems in age estimation at the same time, (7) and (10) are combined and expanded in this paper to construct a rectified softmax loss function for the ConvNet model (see Fig. 5), namely:

$$E(w, b, c) = - \sum_{j=1}^{n} \frac{1}{L_i} \sum_{i=1}^{L_j} \sum_{k=1}^{n} d_k^j \log y_k^j (w, b, c)'$$

$$= \left[ - \frac{1}{L_i} \sum_{i=1}^{L_j} d_k^j \log y_k^j (w, b, c)' \right]$$

$$+ \left[ - \frac{1}{L_2} \sum_{i=1}^{L_2} d_k^2 \log y_k^2 (w, b, c)' \right]$$

$$+ \cdots + \left[ - \frac{1}{L_n} \sum_{i=1}^{L_n} d_k^n \log y_k^n (w, b, c)' \right] \hspace{1cm} (12)$$

2) **CS-BP ALGORITHM**

BP algorithm is the most effective multi-layer neural network learning algorithm and its main characteristic is that the signal is transmitted forward while the error is propagated backward. By constantly adjusting the weight and bias value of the network, the final output of the network is as close as possible to the expected value, and finally the training purpose has been achieved. Therefore, once the loss function is modified, the BP algorithm is bound to be adjusted accordingly.

#### a: Predefined symbols

A multi-layer neural network is assumed to be composed of L layers and there are the following predefined symbols:

1) $W_{\alpha \beta}^{(l)} (1 \leq j \leq s_l; 1 \leq i \leq s_l - 1; 1 \leq l \leq L)$ is the connection weight between the $i$th neuron in the $l$-1th layer and the $j$th neuron in the $l$th layer, and where $s_l$ is the number of neurons in the $l$th layer.

2) $net_k^{(l)} (1 \leq l \leq L; 1 \leq k \leq s_l)$ is the input of the $k$th neuron in the $l$th layer.

3) $h_{\beta}^{(l)} = \{h_{\beta}^{(1)}, h_{\beta}^{(2)}, \ldots, h_{\beta}^{(L)}\} (1 \leq l \leq L)$ is the neuron output of the $l$th hidden layer.

4) $f(\cdot)$ represents the non-linear activation function of the hidden layer, and this paper adopts Rectified Linear Units (ReLU) [33].

#### b: The derivation of BP algorithm

The iterative optimization algorithm that updates the weight and bias matrix adopted here is mini-batch gradient descent, then the key is to calculate the partial derivatives of the error of each training image with respect to the weight and the bias respectively. It is worth noting that we neglect the partial derivatives of the error with respect to the bias, because it always missing the output of a neuron at the previous layer compared with the partial derivatives of the error with respect to the weight.

Suppose a facial image that was originally in class $\rho$ is classified as class $k$, and a conclusion was drawn (see Supplement Material for a detailed derivation):

$$\frac{\partial E'(w, b, c)}{\partial W_{\alpha \beta}^{(l)}} = \begin{cases} 0 & \text{if } (l == L) = \begin{cases} 1 & \text{if } (k == \rho), \\ \frac{1}{L_{\rho}} L_{\rho}^{(L)} h_{\beta}^{(L-1)} & \text{if } (k \neq \rho), \end{cases} \\ \frac{1}{L_{\rho}} \sum_{h=1}^{s_{L-1}} W_{\alpha h}^{(l+1)} s_{h}^{(l+1)} f'(net_{\alpha h}^{(l)}) h_{\beta}^{(l-1)}, & \text{if } (2 \leq l \leq L - 1) = \begin{cases} 1 & \text{if } (k == \rho), \\ \frac{1}{L_{\rho}} \sum_{h=1}^{s_{L-1}} W_{\alpha h}^{(l+1)} s_{h}^{(l+1)} f'(net_{\alpha h}^{(l)}) h_{\beta}^{(l-1)}, & \text{if } (k \neq \rho), \end{cases} \end{cases}$$
where,

$$
\delta_k^{(l)} = \sum_{j=1}^{s_{l+1}} W_{jk}^{(l+1)} \delta_j^{(l+1)} f'(net_k^{(l)}) \\
\text{s.t.} \delta_k^{(L)} = c_{\rho, \rho} (y'_\rho(w, b, c) - 1) = y'_\rho(w, b, c) - 1 \quad (13)
$$

3) CONVNET MODEL AND ALGORITHM STEPS

Before the formal experiment, we carried out some necessary preprocessing steps for the raw facial image, namely face detection, facial landmarks localization and face alignment. Specifically, we adopted the classical cascaded VJ object detector [29] for face detection, and AAM [1] is used for facial landmarks localization. Then, face alignment is conducted based on the location of nasal tip. Finally, the image is resized to the 224×224 for age estimation. In addition, we adopted ResNet-50 [27] as the backbone architecture and integrated the Squeeze-and-Excitation module [28] with it. Finally, in order to accelerate convergence and avoid over-fitting caused by excessive parameters, the fully-connected layer was removed.

After integrating the rectified softmax loss function into ConvNet, we named network as deep cost sensitivity ConvNet. The optimization process of the network is shown as Algorithm 1.

**Algorithm 1 Iterative Optimization About ConvNet**

**Input**: Training set $\{(X_1, y_1), (X_2, y_2), \ldots, (X_L, y_L)\}$. Original weight matrix $W$, Max epochs $M$, Mini batch size $m$, Learning rate $\eta$, Weight decay coefficient $\lambda$, Momentum $\gamma$

**Procedure**:

- feed_forward ConvNet
- for $j$ in range $M$ → Run the entire training set $M$ times
- generate mini_batches using $m$ → Mini_batches is a list of mini_batch
- for mini_batch in mini_batches → Mini_batch gradient descent
- for $X_i, y_i$ in mini_batch → Compute partial derivative $\nabla w$ → See (13)

$$
\partial_t = \gamma \partial_{t-1} + \eta (\nabla w + \lambda w) \\
w' = \partial_t
$$

**Output**: Deep cost sensitivity ConvNet

### E. PRACTICABILITY ANALYSIS

1) EXAMPLES ANALYSIS

In this section, the feasibility of embedding cost matrix proposed in Section III.C.A into softmax regression function is verified by an example. Here, a non-trivial case of $n = 3$ is taken. Therefore, according to the positive and negative of the activation value this case can be divided into $C_3^0 + C_3^1 + C_3^2 + C_3^3 = 8$ specific examples. Below we only verify one of the most representative examples, and the rest of the examples are self-verifying.

**case 1**: $a_1 = 1, a_2 = -2, a_3 = -3$

When the original softmax is used for classification:

$$
y_1 = \frac{e^{a_1}}{e^{a_1} + e^{a_2} + e^{a_3}} = \frac{e^1}{e^1 + e^{-2} + e^{-3}} \approx 0.935.
$$

$$
y_2 = \frac{e^{a_2}}{e^{a_1} + e^{a_2} + e^{a_3}} = \frac{e^{-2}}{e^1 + e^{-2} + e^{-3}} \approx 0.047.
$$

$$
y_3 = \frac{e^{a_3}}{e^{a_1} + e^{a_2} + e^{a_3}} = \frac{e^{-3}}{e^1 + e^{-2} + e^{-3}} \approx 0.018.
$$

At this time, the image is classified into the first category even if it is not. When the proposed ICS-softmax is used for classification, it will adjust the probability of each category according to the actual situation:

1) $\rho = 1$

$$
C'_{1,1} = |1 - 1| + 1 = 1, \quad C'_{1,2} = |1 - 2| + 1 = 2, \\
C'_{1,1} = |1 - 3| + 1 = 3.
$$

Looking up (9), and

$$
c_{1,1} = \frac{1}{c_{1,1}} = 1, \quad c_{1,2} = c'_{1,2} = 2, \quad c_{1,3} = c'_{1,3} = 3.
$$

The corresponding probability distribution as the following:

$$
y'_1 = \frac{e^{c'_{1,1} t_{11}}}{e^{c'_{1,1} t_{11}} + e^{c'_{1,2} t_{12}} + e^{c'_{1,3} t_{13}}} \\
= \frac{e^{1 t_{11}}}{e^{1 t_{11}} + e^{-2 t_{12}} + e^{-3 t_{13}}} \approx 0.992.
$$

$$
y'_2 = \frac{e^{c'_{1,2} t_{22}}}{e^{c'_{1,1} t_{11}} + e^{c'_{1,2} t_{12}} + e^{c'_{1,3} t_{13}}} \\
= \frac{e^{-2 t_{22}}}{e^{1 t_{11}} + e^{-2 t_{12}} + e^{-3 t_{13}}} \approx 0.007.
$$

$$
y'_3 = \frac{e^{c'_{1,3} t_{33}}}{e^{c'_{1,1} t_{11}} + e^{c'_{1,2} t_{12}} + e^{c'_{1,3} t_{13}}} \\
= \frac{e^{-3 t_{33}}}{e^{1 t_{11}} + e^{-2 t_{12}} + e^{-3 t_{13}}} \approx 0.001.
$$

2) $\rho = 2$

Similarly, and

$$
c_{2,1} = \frac{1}{c_{2,1}} = \frac{1}{2}, \quad c_{2,2} = c'_{2,2} = 1, \quad c_{2,3} = c'_{2,3} = 2.
$$

The corresponding probability distribution as the following:

$$
y'_1 = \frac{e^{c'_{2,1} t_{11}}}{e^{c'_{2,1} t_{11}} + e^{c'_{2,2} t_{12}} + e^{c'_{2,3} t_{13}}} \\
= \frac{e^{2 t_{21}}}{e^{1 t_{11}} + e^{-2 t_{12}} + e^{-3 t_{13}}} \approx 0.921.
$$

$$
y'_2 = \frac{e^{c'_{2,2} t_{22}}}{e^{c'_{2,1} t_{11}} + e^{c'_{2,2} t_{12}} + e^{c'_{2,3} t_{13}}} \\
= \frac{e^{2 t_{22}}}{e^{1 t_{11}} + e^{2 t_{12}} + e^{3 t_{13}}}.
$$
\[
\begin{align*}
\frac{e^{2x_1}}{e^{2x_1} + e^{-2x_1} + e^{-3x_2}} & \approx 0.077. \\
y_3' & = \frac{e^{2x_1} + e^{2x_2} + e^{3x_3}}{e^{2x_1} + e^{2x_2} + e^{3x_3}} \\
& = \frac{e^{1x_1} + e^{-2x_1} + e^{-3x_2}}{e^{1x_1} + e^{-2x_1} + e^{-3x_2}} \approx 0.002.
\end{align*}
\]

3) \( \rho \Rightarrow 3 \)

Similarly, and
\[
c_{3,1} = \frac{1}{c_{3,1}} = \frac{1}{3}, \quad c_{3,2} = c_{3,2}' = 2, \quad c_{3,3} = c_{3,3}' = 1.
\]

The corresponding probability distribution as the following:
\[
y_1' = \frac{e^{3x_{1,01}}}{e^{3x_{1,01}} + e^{3x_{3,02}} + e^{3x_{3,03}}} \\
& = \frac{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}}{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}} \approx 0.952.
\]
\[
y_2' = \frac{e^{3x_{1,01}} + e^{3x_{3,02}} + e^{3x_{3,03}}}{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}} \\
& = \frac{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}}{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}} \approx 0.013.
\]
\[
y_3' = \frac{e^{3x_{1,01}} + e^{3x_{3,02}} + e^{3x_{3,03}}}{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}} \\
& = \frac{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}}{e^{1x_{1,01}} + e^{-2x_2} + e^{-3x_3}} \approx 0.035.
\]

The remainder cases as the following, and reader can verify for themselves:

- case 2: \( o_1 = 1 \) \( o_2 = 2 \) \( o_3 = 3 \)
- case 3: \( o_1 = -1 \) \( o_2 = -2 \) \( o_3 = -3 \)
- case 4: \( o_1 = 1 \) \( o_2 = 2 \) \( o_3 = -3 \)
- case 5: \( o_1 = 1 \) \( o_2 = -2 \) \( o_3 = 3 \)
- case 6: \( o_1 = 1 \) \( o_2 = -2 \) \( o_3 = 3 \)
- case 7: \( o_1 = -1 \) \( o_2 = -2 \) \( o_3 = 3 \)
- case 8: \( o_1 = -1 \) \( o_2 = -2 \) \( o_3 = -3 \)

From the above, ICS-softmax has been increasing the probability of the desired class, although by a small margin. In this way, it could not only accelerate the convergence of the network, but also improve the accuracy of ConvNet greatly after repeated iterations.

### F. THEORETICAL PRACTICABILITY ANALYSIS

After proposing the method, we need to verify the theoretical practicability of it to analyze whether it satisfies the general nature of the classification loss function. To this end, we need to consider two properties here, namely classification-calibration [25] and guess-aversion [22]. Note that the classification direction of all-sided CS-softmax loss function is determined by ICS-softmax loss, so we only need to consider whether the ICS-softmax loss function for a single sample meets the above properties (for simplicity, we ignored \( i \) in the upper right).

**Classification-calibration.** Classification-calibration is essentially a pointwise form of Fisher consistency for classification. Since the use of surrogate loss function of 0-1 loss introduces additional risk, it can guarantee the consistency of the decision of proposed loss function with the Bayesian decision \( \Phi^* ( \cdot ) \) in terms of statistical property.

**Argument 1:** ICS-softmax error function

\[
E(\mathbf{w}, \mathbf{b}, \mathbf{c})'' = E(\mathbf{d}, \mathbf{o}, \mathbf{c})'' = - \sum_{k=1}^{n} d_k \log y_k (\mathbf{o}, \mathbf{c})
\]

conforms to classification-calibration.

Proof: If there is a sample \( X \), assuming it belongs to the \( \rho \)th class and the classification result is correct, and then its ICS-softmax error is

\[
E(\rho, \mathbf{o}, \mathbf{c})'' = - \log \left( \frac{e^{\mathbf{w}_\rho, \mathbf{o}_\rho}}{\sum_{k=1}^{n} e^{\mathbf{w}_k, \mathbf{o}_k}} \right).
\]

At this point, the expected risk is

\[
R_{\text{exp}} (\mathbf{o}) = E \left[ E(\mathbf{d}, \mathbf{o}, \mathbf{c})'' \right]
\]

\[
= \sum_{\rho=1}^{n} P(\rho | X) E(\mathbf{d}, \mathbf{o}, \mathbf{c})''
\]

\[
= \sum_{\rho=1}^{n} P(\rho | X) \log \left( \frac{e^{\mathbf{w}_\rho, \mathbf{o}_\rho}}{\sum_{k=1}^{n} e^{\mathbf{w}_k, \mathbf{o}_k}} \right).
\]

We let \( \frac{\partial R_{\text{exp}} (\mathbf{o})}{\partial \mathbf{o}_t} |_{\rho=t} = 0 (\forall \mathbf{t} \in [1,n]) \) to find the ideal set of ConvNet predicted output \( \mathbf{o} \), and then:

\[
\frac{\partial R_{\text{exp}} (\mathbf{o})}{\partial \mathbf{o}_t} |_{\rho=t} + \frac{\partial R_{\text{exp}} (\mathbf{o})}{\partial \mathbf{o}_t} |_{\rho \neq t} = 0. \quad (14)
\]

\[
\frac{\partial R_{\text{exp}} (\mathbf{o})}{\partial \mathbf{o}_t} |_{\rho=t} = \frac{\partial}{\partial \mathbf{o}_t} \left( - \sum_{\rho=1}^{n} P(\rho | X) \log \left( e^{\mathbf{w}_\rho, \mathbf{o}_\rho} \right) \right)
\]

\[
+ \sum_{\rho=1}^{n} P(\rho | X) \log \left( \sum_{k=1}^{n} e^{\mathbf{w}_k, \mathbf{o}_k} \right)
\]

\[
= -c_{t,t} P(t | X) + c_{t,t} P(t | X) \frac{e^{c_{t,t} \mathbf{w}_t}}{\sum_{k=1}^{n} e^{c_{t,k} \mathbf{w}_t}}. \quad (15)
\]

\[
\frac{\partial R_{\text{exp}} (\mathbf{o})}{\partial \mathbf{o}_t} |_{\rho \neq t} = \frac{\partial}{\partial \mathbf{o}_t} \left( - \sum_{\rho=1}^{n} P(\rho | X) \log \left( e^{\mathbf{w}_\rho, \mathbf{o}_\rho} \right) \right)
\]

\[
+ \sum_{\rho=1}^{n} P(\rho | X) \log \left( \sum_{k=1}^{n} e^{\mathbf{w}_k, \mathbf{o}_k} \right)
\]

\[
= -c_{t,t} P(t | X) + c_{t,t} P(t | X) \frac{e^{c_{t,t} \mathbf{w}_t}}{\sum_{k=1}^{n} e^{c_{t,k} \mathbf{w}_t}}.
\]
Since the proposed cost matrix is based on DCMP (see also the example in the section III.C.A), there is:

\[ y'_\rho = \frac{e^{\rho,\rho o}}{\sum_{k=1}^{n} e^{\rho,k o_k}} > y_{\rho} = \frac{e^{\rho}}{\sum_{k=1}^{n} e^{o_k}} > \frac{1}{n}. \]

\[ \Rightarrow - \log \left( \frac{\sum_{k=1}^{n} e^{\rho,k o_k}}{\sum_{k=1}^{n} e^{o_k}} \right) < - \log \frac{1}{n}. \]

\[ \Rightarrow E(\rho, o, c)'' < E(d, \Lambda, c)''. \]

Argument 2 is satisfied.

**IV. EXPERIMENT**

In this section, we first introduced dataset, evaluation criteria and the details of experiment, and finally carried out two groups of experiments: ablation experiment and contrast experiment.

**A. DATASETS**

We first chose MORPH Album2 [12], which is basically composed of the white race and the black race, as an option. In addition, for the sake of universality of our algorithm, we also adopt Asian Face Age Dataset (AFAD) [10], a dataset composed of the yellow race.

1) MORPH

The full MORPH Album2 contains over 94,000 facial images, but since it is partially public, we only use and discuss a subset of it, i.e., MORPH Public release (for short, MORPH). This dataset contains 55,608 unique images of more than 13,000 individuals and the ages range from 16 to 77. For a detailed distribution and some examples of MORPH at each age, see Fig. 1(a) and Fig. 6, respectively. Since MORPH does not provide an official split standard for the training set and the testing set, in order to compare with other age estimation methods, the split standard in [10], [11], [16]–[19] was followed, namely, randomly divide MORPH into 80\% for the training set and remaining 20\% for the testing set.

![FIGURE 6. The example in MORPH.](image-url)
only the largest public dataset for the age estimation so far, but also a powerful dataset for researching facial age in the wild. For a detailed distribution some examples of AFAD at each age, see Fig. 1(b) and Fig. 7, respectively. The split standard of it is analogous to MORPH.

**FIGURE 7.** The example in AFAD.

### B. EVALUATION CRITERIA

In age estimation, the commonly used evaluation criteria are Mean Absolute Error (MAE) and the Cumulative Scores (CS).

**MAE.**

\[
MAE = \frac{1}{L_t} \sum_{i=1}^{L_t} \left| \rho^i - k^i \right|.
\]  

(17)

where \( L_t \) is the amount of the testing set, and \( \rho \) is the desired class while \( k \) is the predicted class. MAE is used to calculate the mean age error of the all facial images in the testing set. For an example, the MAE of an age estimation algorithm is 5, that is, given a facial image, the estimated age will be 5 years younger or 5 years older on average than the actual age.

**CS.**

\[
CS = \frac{1}{L_t} \sum_{i=1}^{L_t} bool(|\rho^i - k^i| \leq m).
\]  

(18)

where bool is the Boolean function and \( m \) represents the tolerable age error. CS could be understood as the accuracy rate under the allowable error.

### C. THE DETAILS OF EXPERIMENTS

Our experiments were based on the caffe open source framework [30] (solver mode: GPU) in the Ubuntu18.04 system. Deep cost sensitivity learning adopted the strategy of transfer learning, and caffe model derived from [31]. On this weight basis, we made fine-tuning on MORPH and AFAD. The hyper-parameters in the solver prototxt were set as follows: basic learning rate is 0.0015, and decaying exponentially; weight decay coefficient is 0.0005; Stochastic gradient descent is on, and mini-batch size is 64; momentum is acquiescent 0.9.

### D. ABLATION EXPERIMENT

In this section, ablation experiment with control variables method was conducted to compare the effects of different loss functions on age estimation performance, so as to further highlight the superior performance of our idea. We use one of the four loss function include our proposed ECS-softmax loss, ICS-softmax loss, CS-softmax loss and original softmax loss, as a loss function during the training phase while the rest of the network stays the same. As a result, four ConvNets with parameters attached to different meanings to complete age estimation during the testing phase. The final results are shown in Table 1, and the following Q& A are unfolded:

**TABLE 1.** Comparison of different cross entropy loss functions.

| Loss Function         | MORPH  | AFAD  |
|-----------------------|--------|-------|
| Original softmax loss | 3.377  | 3.806 |
| ECS-softmax loss      | 3.284  | 3.667 |
| ICS-softmax loss      | 3.141  | 3.549 |
| CS-softmax loss       | 2.885  | 3.177 |

Q1: Why is it that age estimation performance of AFAD worse than MORPH (3.806 > 3.377), even though the same original softmax loss is used?

A1: Because MORPH is collected in a restricted environment, the image resolution and lighting conditions are easy to control. On the contrary, the most facial images in AFAD are collected in a natural way, so many images are blurred (e.g. bad front-facing camera), noisy (e.g. weird hairstyle that cover even the whole face) and so on. To sum up, the gap is quite reasonable.

Q2: Why is it that AFAD better than MORPH ((3.806-3.667) > (3.377-3.284)), in terms of the revenue from ECS-softmax loss?

A2: This indicates that compared with MORPH, the ECS problem of AFAD is more serious, that is, its number distribution at each age is slanting, which can be seen from the Fig. 1.

Q3: Why is it that ICS-softmax loss improve performance more than ECS-softmax loss in both datasets ([(3.377-3.141) > (3.377-3.284)] && [(3.806-3.549) > (3.806-3.667)])?

A3: This shows that the ICS is a crucial factor in the age estimation, and the proposed cost matrix is effective.

Q4: Why is CS-softmax loss the best improvement in performance?

A4: First of all, it combines ICS and ECS formally, and secondly, the significance of CS for age estimation is evident.

### E. CONTRAST EXPERIMENT

To illustrate the advantage of deep cost sensitivity learning, we compared the result with the state-of-the-art age estimation methods. Both MAE and CS are shown in the TABLE 2, Fig. 8 and Fig. 9, respectively. The following is a brief introduction to the methods involved in the table and figure. Deeply learn feature [9] uses the feature map of the convolution layer and the pooling layer in a 4-layers ConvNet as the representations, and finally feeds it to SVR for age regression. MR-CNN [10] and C-CNN [11] regard age estimation as regression and classification problems, respectively.
OH-ranker [26], Ranking-CNN [19] and OR-CNN [10] consider the relative order of age labels [37] and then perform a series of two classifications. The difference between the three is the selection of the classifier, the first one uses SVM and the last two use ConvNets. The difference between OR-CNN and Ranking-CNN is that the former model is more concise because it shares the middle characterization of ConvNet. For the improvement of ConvNet’s fitting ability, the technique of the multi-scale analysis and the local aligned patches have been added to Multi-scale [18]. DEX [17] proposes a softmax expected value refinement after the softmax loss layer during the testing phase to estimate apparent age. CNN-ELM [38] uses CNN to extract features from the input face image, and then use Extreme Learning Machine (ELM) to classify the intermediate results. D2C [39] proposes a novel cumulative hidden layer which is supervised by a point-wise cumulative signal. In this cumulative hidden layer, their model is learned indirectly through faces of neighboring ages. GA-DFL [40] splits ordinal ages into a set of discrete groups and learn deep feature transformations across age groups to project each face pair into the new feature space, where the intra-group variances of positive face pairs from the training set are minimized and the inter-group variances of negative face pairs are maximized, simultaneously. In addition, in order to illustrate the portability of proposed rectified softmax loss function, we have embedded it to Ranking-CNN and OR-CNN respectively, forming Ranking-CSCNN and OR-CSCNN.

As can be seen from Table 2, our method has achieved the first place in the white race and black race (MORPH) and the yellow race (AFAD), which shows that our method is simple, but perfectly in line with the idea of Occam’s razor: Entities should not be multiplied unnecessarily. In addition, some conclusions can be drawn:

1. There are three methods based on rank theory [37], and they all achieve good results, which indicates that the relative order among age labels is a very important factor in age estimation.
2. Except for OH-ranker, the rest of the most advanced methods are based on deep learning, but unfortunately, they mainly rely on migration learning to improve performance.
3. The only method that can be compared to our approach is Ranking-CNN, but its downside is that it uses multiple ConvNets, which is a waste of computing resources and space.
4. Two methods with the rectified softmax loss function performed very well, and even the result of Ranking-CSCNN is better than our method, which shows that our method is well adapted to the method of facial age estimation based on CNN.

As can be seen from Fig. 8 and Fig.9, when the tolerable age error is 0-15 years, our method is ahead of other methods most of the time. This adds further arguments to our thinking: ConvNet structure modification method that considers specific factor for specific task is more reliable than blind migration learning. It is worth mentioning that the specific task of this paper is age estimation, and the specific factor is cost-sensitive learning.

In Fig. 10 and Fig. 11, we further compare the accuracy of the original loss function and the rectified loss function.
Although deep cost sensitivity learning we discuss is aimed at age estimation, it is undeniable that when there is a multi-classification problem or a binary classification problem, the classification accuracy can be improved by only modifying the preliminary cost matrix $c'$ towards the specific problems. (e.g., if the misclassification costs are all equal, let $c'$ be a uniform matrix, i.e., $c' = (1_{n \times n})$). After all, even if the misclassification costs in pattern recognition tasks are likely to differ, the class-imbalance problem is almost everywhere in all benchmark datasets.

In addition, in the future work, some other factors inherent in age estimation can be considered to further improve its performance.

VI. CONCLUSION

We came up with a rectified softmax loss with all-sided cost sensitivity for age estimation called CS-softmax loss. Specifically, it covers the most common problems in age classification, which consider the uneven distribution problem of facial datasets and the cost of age misclassification. For the latter, we further propose a novel cost matrix that maximizes the desired class probability while considering the misclassification cost. Finally, the effectiveness of the method is demonstrated by theoretical analysis and different experiments based on the interracial benchmark datasets.

REFERENCES

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” in Proc. Eur. Conf. Comput. Vis., 1998, pp. 484–498.
[2] 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.
[3] X. Geng, Z. H. Zhou, Y. Zhang, L. Gang, and H. D. Hong, “Learning from facial aging patterns for automatic age estimation,” in Proc. Multimedia ACM Int. Conf. Multimedia, 2006, pp. 307–316.
[4] J. Lu, V. E. Liong, and J. Zhou, “Cost-sensitive local binary feature learning for facial age estimation,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5356–5368, Dec. 2015.
[5] C. Ng, Y. T. Cheng, G. S. Hsu, and H. Y. Moi, “Multi-layer age regression for face age estimation,” in Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA), 2017, pp. 294–297.
[6] G. D. Guo, G. W. Mu, Y. Fu, and T. S. Huang, “Human age estimation using bio-inspired features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Aug. 2009, pp. 112–119.
[7] 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.
[8] G. Guo, Y. Fu, T. S. Huang, and R. D. Charles, “Locally adjusted robust regression for human age estimation,” in Proc. IEEE Workshop Appl. Comput. Vis., Jan. 2008, pp. 1–6.
[9] X. L. Wang, R. Guo, and C. Kambhamettu. “Deeply-learned feature for age estimation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., Jan. 2015, pp. 534–541.
[10] Z. X. Niu, M. Zhou, L. Wang, X. B. Gao, and H. Gang, “Ordinal regression with multiple output CNN for age estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4920–4928.
[11] 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.
[12] W. Rawls and K. Riechek, “MORPH: Development and optimization of a longitudinal age progression database,” in Biometric ID Management and Multimodal Communication. Berlin, Germany: Springer, 2009, pp. 17–24.
[13] Y. Fu, X. Ye, 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., Jul. 2007, pp. 1383–1386.
[14] M. Riesenhuber and T. Poggio, “Hierarchical models of object recognition in cortex,” *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.

[15] G. D. Guo, G. W. Mu, and Y. Fu, “A study on automatic age estimation using a large database,” in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1986–1991.

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

[17] R. Rothe, R. Timofte, and L. V. Gool, “DEX: Deep expectation of apparent age from a single image,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 252–257.

[18] Z. Lei and S. Z. Li, “Age estimation by multi-scale convolutional network,” in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 144–158.

[19] 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), Jul. 2017, pp. 742–751.

[20] K. M. Ting, “Inducing cost-sensitive trees via instance weighting,” in *Proc. Eur. Symp. Principles Data Mining Knowl. Discovery*, 1998, pp. 139–147.

[21] L. Cai and T. Hofmann, “Hierarchical document categorization with support vector machines,” in *Proc. 13th ACM Int. Conf. Inf. Knowl. Manage.*, 2004, pp. 78–87.

[22] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos, “Guess-averse loss functions for cost-sensitive multiclass boosting,” in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 586–594.

[23] H. Masnadi-Shirazi and N. Vasconcelos, “Cost-sensitive boosting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 294–309, Feb. 2011.

[24] M. Kur and I. Kononenko, “Cost-sensitive learning with neural networks,” in *Proc. 13th Eur. Conf. Artif. Intell.*, 1998, pp. 445–449.

[25] S. H. Khan, M. Hayat, M. Bennamoun, A. S. Ferdous, and T. Roberto, “Cost sensitive learning of deep feature representations from imbalanced data,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2015.

[26] 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.*, Jun. 2011, pp. 585–592.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–15.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 7132–7141.

[29] 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.*, Aug. 2001, pp. 511–518.

[30] Y. Jia, E. Shelhamer, J. Donahue, and S. Karayev, “Caffe: Convolutional architecture for fast feature embedding,” in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[31] Q. Cao, L. Shen, W. Xie, and O. Parkhi, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. IEEE Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 49–57.

[32] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2012, pp. 60–78.

[33] Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. 25th Int. Conf. neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[34] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[35] S. Xie and Z. Tu, “Holistically-nested edge detection,” *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 3–18, Dec. 2017.

[36] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2015, pp. 46–54.

[37] L. Li and H. T. Lin, “Ordinal regression by extended binary classification,” in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2006, pp. 865–872.

[38] M. X. Duan, K. L. Li, and C. Q. Yang, “A hybrid deep learning CNN-ELM for age and gender classification,” *Neurocomputing*, vol. 275, pp. 448–461, Jan. 2018.

[39] 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.

**DAXIANG LI** received the M.S. degree from the Department of Electronics Science, Northwest University, China, in 2005, and the Ph.D. degree from Northwest University, in 2015. He is currently an Associate Professor with the School of Telecommunication and Information Engineering, Xi’an University of Posts and Telecommunications, China. His research interests include image retrieval, image classification, and machine learning.

**XUAN MA** received the B.S. degree from the Shaanxi University of Technology, China, in 2017. He is currently pursuing the master’s degree with the School of Telecommunication and Information Engineering, Xi’an University of Posts and Telecommunications, China. His research interests include age estimation, machine learning, and pattern recognition.

**YAQIONG REN** received the B.S. degree from the Shaanxi University of Technology, China, in 2017. She is currently pursuing the master’s degree with the School of Telecommunication and Information Engineering, Xi’an University of Posts and Telecommunications, China. Her research interests include age estimation and pattern recognition.

**SHYH-WEI TENG** received the B.S. degree in computing and the Ph.D. degree in information technology from the Faculty of Information Technology, Monash University, Australia, in 1998 and 2004, respectively. He is currently an Associate Professor and the Deputy Head of the School of Engineering and Information Technology, Faculty of Science and Technology, Federation University Australia. His research interests include image processing, machine learning, and bioinformatics. He regularly serves on the reviewer boards of a number of top international journals, including *Pattern Recognition* and *Neurocomputing*, and as a Program Member for several international conferences, such as the IEEE CEC and ICME. He was also nominated for the 2007 Staff Excellence Award in Teaching and the 2008 Excellence Award in Gippsland Community Engagement Team.