Facial-image based Age Estimation Using Imbalanced Datasets

Yasi Liu¹, Tianzi Wang², Limin Zheng¹,* and Lu Yang¹

¹China Agricultural University, College of Information and Electrical Engineering, No. 17, Qinghua East Road, Haidian District, Beijing, China, 100083
²Virginia Tech, ISE department, Blacksburg, USA

*Limin Zheng, E-mail: zhenglimin@cau.edu.cn

Abstract. Facial image based human age estimation is of great application significance. The public-available facial image datasets used for age estimation suffer greatly from the uneven distribution of images of different age groups, which may lead to the low estimation accuracy of the under-sampled age categories and limit the usage of the age estimation in certain applications. We propose a three-stage probability adjustment based CNN algorithm to solve the imbalanced distribution problem of the dataset. In particular, we construct an ENIN neural network structure by applying the Network in Network (NIN) structure to the traditional convolution neural network (CNN) and use the probability vector adjustment to improve the classification accuracy of the under-sampled age categories. Then, we filter out the images with high possibility of being misclassified after the probability vector adjustment and reset their categories by comparing cosine similarity and retraining the ensembled ENIN classifier. We also introduce a population-age-distribution based accuracy metric Accuracy-P to estimate the performance of the age estimation algorithm in real-world applications. Our experimental results confirm that our algorithm can effectively improve the overall estimation accuracy by significantly improving the accuracy of the under-sampled age groups while maintaining satisfactory accuracy for the other age groups.

1. Introduction

In many age-restricted places or applications, coarse-grained facial image based human age estimation plays important roles, such as prohibiting minors from accessing the harmful information on the internet, recognizing face with a long age span and has broad application prospects in the search for lost children.

Computer vision technology leverages the common features on human faces for the facial image based age estimation. In specific, prediction models [1-3] that use the age-related features (e.g., relative distances of the main parts and the wrinkles in the specific areas of the face [4]) for categorizing the facial image into certain age groups are trained. In such model training procedures, most existing approaches use public-available datasets of facial images labeled with age.

One major problem in these approaches is that the public-available datasets they leverage are greatly imbalanced in the amount of images of different age groups, with certain age groups (e.g., the senior, the child) have fewer images than other groups (e.g., adult). Such uneven distribution in the image datasets may cause low estimation accuracy of the under-sampled age groups (i.e., the minority age groups), which will further affect the overall estimation accuracy and limit the usage of the age estimation algorithm in certain application scenarios (e.g., recognizing senior for ticket discount). Such observations are confirmed by our study, as well as the classification results obtained by the
commonly used classification algorithms mentioned above [4-6]. Although there are several resampling algorithms [7-17] to balance the distribution of the sample data and several cost-sensitive learning algorithms [18-21] to adjust the cost of dividing the sample into categories with different sample sizes, directly applying these algorithms won’t increase the accuracy of the under-sampled groups to a satisfactory value.

To solve the problem, this paper introduces a novel accuracy metric Accuracy-\(P\) to evaluate how an age classification algorithm fits real-world applications, as well as a three-stage probability adjustment based CNN algorithm.

The rest of this paper is organized as follows. Section II analyzes the problems caused by the imbalanced age distribution of three popular datasets and points out the irrationality of the overall accuracy while applying it to the imbalanced datasets. Section III introduces the novel population-age-distribution based accuracy measurement Accuracy-\(P\). Section IV introduces the three-stage probability adjustment based CNN algorithm. Section V shows the effectiveness of the algorithm in the image datasets. Section VII concludes this paper.

2. Datasets Analysis

2.1 Facial-Image Dataset

This article covers three datasets. The FG-NET [22] dataset ranges in age from 0 to 69. The MORPH-II [23] dataset ranges in age from 16 to 77. A general way of using the Adience [24] dataset for age classification is to divide the age into 8 categories. Some facial images in these datasets are shown in Figure 1.

The FG-NET dataset and the MORPH-II dataset are often used together as the FG-NET + MORPH dataset [4], since there are no facial images of people age 0 to 16 in the MORPH-II, and the number of facial images in the FG-NET is relatively small. The age is divided into four categories: Children (0-11 years old), Teen Age (12-21 years old), Adult (22-60 years old), and Senior Adult (> 60 years old).

![Figure 1. The facial images in the datasets used in this paper. The number under each image represents the age.](image)

2.2 Imbalanced Image Distribution of Age Categories

The population age distribution refers to the proportion of each age category in the total population. We use the Population Pyramids [25] to observe the population age distributions in different countries or regions in 2018, as shown in Figure 2. The population age distribution of Japan is used to represent a typical aging population distribution. The population age distribution in the world reflects the age distribution of the global population. That is, the proportion of the population in the world decreases with the increase of age. The age distributions in the FG-NET+MORPH dataset and the Adience dataset are shown in Table 1 and Table 2, respectively. It is obvious that the population age distribution under natural conditions and the image distributions of age categories in the datasets are both imbalanced.

In the population age distributions of either Japan or the world, the Senior Adult (over 60) takes a much larger portion of the population (13%-34%), compared with the numbers shown in the datasets (1.6% ~ 4.7%). However, in real-world applications of general purposes, we expect the collected images will basically fit the population age distributions.
Figure 2. Population age distribution in Japan and the world in 2018.

Table 1 Age distribution in the FG-NET+MORPH dataset.

|                | Children | Teen Age | Adult | Senior Adult | Total |
|----------------|----------|----------|-------|--------------|-------|
| FG-NET         | 444      | 302      | 249   | 7            | 1002  |
| MORPH-II       | 0        | 3081     | 11865 | 258          | 15204 |
| FGNET+MORPH    | 444      | 3383     | 12114 | 265          | 16206 |

Table 2 Age distribution in the Adience dataset.

|        | 0-2 | 4-6 | 8-13 | 15-20 | 25-32 | 38-43 | 48-53 | >60 | Total |
|--------|-----|-----|------|-------|-------|-------|-------|-----|-------|
| Adience| 1820| 1586| 1694 | 1121  | 3399  | 1598  | 572   | 583 | 12373 |

2.3 The Problem Caused by the Imbalanced Distribution

The low classification accuracies of the under-sampled age categories are ubiquitous problems of the existing age estimation methods. In the datasets used in this paper, the classification accuracy of the Senior Adult age category is significantly lower than that of other age categories. In the existing image classification applications, the overall accuracy of the classification algorithm represents the probability of correctly classify an arbitrary image in the dataset. However, as the age distribution of the datasets differs greatly from the population age distribution, such accuracy measurement cannot be applied to predict the performance of the algorithm in a real-world application scenario.

3 Population-Age-Distribution based Accuracy Measurement

To evaluate the effectiveness of an age estimation algorithm in real-world applications, we introduce a population-age-distribution based accuracy measurement $\text{Accuracy-P}$, defined as:

$$\text{Accuracy-P} = \sum_{i=1}^{C} \frac{\text{Accu}_i \times \text{Prop}_i}{\sum_{i=1}^{C} \text{Prop}_i}$$

(1)

where $i$ represents different age categories, $\text{Accu}_i$ and $\text{Prop}_i$ represent the classification accuracy and the population proportion of age category $i$ of the natural age distribution, respectively. $\text{Accuracy-P}$ estimates the probability of correctly classifying an image of an arbitrary person, whose age distribution follows the natural population age distribution.

To better understand the novelty of $\text{Accuracy-P}$, we compare it with the accuracy metrics widely used in existing approaches, the accuracy metrics can be denoted as:

$$\text{Accuracy-D} = \sum_{i=1}^{C} \frac{\text{Accu}_i \times \text{Prop}_i'}{\sum_{i=1}^{C} \text{Prop}_i'}$$

(2)

where $i$ represents different age categories, $\text{Accu}_i$ and $\text{Prop}_i'$ represent the classification accuracy and the proportion of images of age category $i$ in the datasets. By comparing equation (1) and equation (2), we can see that the existing accuracy metrics are affected by the imbalanced distribution of the collected image datasets. On the contrary, $\text{Accuracy-P}$ can be used to reflect how accurate an algorithm is while processing facial images collected by real-world applications.
4 Three-Stage Probability Adjustment based CNN Algorithm

4.1 Algorithm Overview
In this section, we introduce our three-stage facial image based age estimation algorithm in details. The entire algorithm focuses on calculating the correct probability vector — the probabilities of an image belonging to each age category. The overall workflow of the algorithm is shown in Figure 3, which mainly contains three steps.

1. We first calculate the probability vector $P$ using the enhanced NIN (ENIN) network.
2. We generate another probability vector $P'$ considering the imbalanced image distribution in the training set.
3. We further redefine the categorization result by filtering out the images in the under-sampled age category with a high possibility of being miscategorized using $P'$, and reset their categorization result using a new probability vector $P''$ constructed by applying a new cost function to $P'$ and using a new probability vector $P_e$ constructed by retraining an ensemble ENIN classifier.

In the rest of this section, we will introduce in details the procedure of each stage.

![Figure 3. The workflow of the facial image based age estimation.](image)

4.2 Stage 1: Applying the Enhanced NIN Algorithm
First, we follow the facial image preprocessing procedures proposed by Baggio [26] to effectively extract the facial features and convert the original images to the target images of 256×256 pixels. The effect of the facial image preprocessing is shown in Figure 4.

![Figure 4. The effect of the facial image after preprocessing. (a) The original image. (b) The preprocessed image.](image)

We use the term ENIN to refer to the enhanced NIN network model. It enhances the effectiveness of the NIN [27] network by adding a few network structures, as shown in Figure 5. We first apply Mlpconv structure to the network, as the Mlpconv structure has strong abstraction and generalization ability and is more capable of expressing the rich features of facial images than the common convolution layer.

In particular, the structure of the slice layer and the eltwise layer is shown in Figure 6. The slice layer slices the output of the last layer into multiple output layers based on a given dimension and the slice indices. The eltwise layer uses the element-level max operation to take the maximum values of these multiple output layers to reduce the dimension of data. Our network structure is divided into two parts in the 4th Mlpconv structure. The first part contains conv4, conv7, and conv8 layers, and the second part cascades the features of the conv4 and cccp7 layers. Then, cccp9 is used to reduce the noise and select the features. The features of cccp8 and cccp9 are combined and sent into the fully-connected layer.
Each dataset in this paper is divided into three parts: the training set, the verification set, and the test set. The sample size of each part accounts for 80%, 10%, and 10% of the total sample size of the dataset. The test set is used to evaluate the accuracy of the ENIN model and our algorithm.

![ENIN's network structure](image1)

**Figure 5.** ENIN’s network structure.

![The workflow of our algorithm](image2)

**Figure 6.** The slice+eltwise layers.

**Figure 7.** The workflow of our algorithm.

### 4.3 Stage 2: Adjusting the Probability Vector

Due to the influence of imbalanced data distribution, simply applying the ENIN may cause low age classification accuracy for the minority age categories. As a solution, we apply a cost function, which is calculated based on the image distribution of each age category, to adjust the probability vector generated by the ENIN algorithm. Figure 7 shows the detailed workflow of the second stage. The key idea behind the cost function is the threshold moving method [18].

Threshold moving uses a cost function to adjust the probability vector, as shown in equation (3).

\[
P^*_i = \eta \sum_{c=1}^{C} P_{Cost}[i,c]
\]

where \( P_i \) represents the probability output of the trained CNN model, \( C \) represents the total number of categories, \( i \in \{1..C\} \) represents different age categories, \( \sum_{i=1}^{C} P_i = 1, 0 \leq P_i \leq 1 \), the predicted category value determined by the trained CNN model is \( L = \arg \max_i P_i \), \( \eta \) is a normalization item, which
\[ \sum_{c=1}^{C} P'_c = 1. \]

Cost\([i,c] \] represents the cost of misclassifying a sample of the category \( i \) into a category \( c \), which affects the degree of changing the probability.

The first stage of our algorithm outputs the probability vector \( P \) that an image belongs to each age category. Then, we calculate \( P' \) according to equation (3) and equation (4).

\[
\begin{align*}
\text{Cost}_{i,c} &= \begin{cases} 
(\log(N_i / N_i + 1))^2 & \text{if } N_i < N_c & \text{& } c \neq i \\
(\log(N_i / N_i + 1)) & \text{if } N_i > N_c & \text{& } c \neq i \\
0 & \text{if } c = i 
\end{cases}
\end{align*}
\]

where \( N_i \) and \( N_c \) represent the sample size of corresponding age categories in the training set.

From equation (4), we can observe that applying the cost function adjusts the probability vector of an image by increasing the probability of the minority age groups and decreasing the probability of the majority age groups, as the cost of misclassifying an image into a majority category is higher than that of misclassifying it into a minority category.

The experiment results indicate that the probability vector adjustment procedure improves the classification accuracies of the under-sampled age categories while maintaining the satisfactory accuracies of other categories. However, the classification error rate is still high for the under-sampled age categories.

### 4.4 Stage 3: Category Redefinition of the Mis-Categorized Image

To further reduce the misclassification rate of images from the under-sampled age categories, we adopt a category redefinition procedure to redefine the age categories of the images with high possibility of being misclassified after the probability vector adjustment. The procedure contains several steps.

First, we collect the misclassified images from the under-sampled age categories. We apply the ENIN model and the probability vector adjustment to the images in the training set and the validation set, and compare their classification results with their actual age categories and identify the misclassified images from the minority age categories. We name these misclassified images of different minority categories in the training set as \textit{Samtrainmireclassi}, and name these misclassified images of different categories in the validation set as \textit{Samvalreclassi}.

Second, for these misclassified images, we apply a second probability vector adjustment procedure to generate a new probability vector \( P'' \). \( P'' \) is calculated by applying a new cost function to the probability vector \( P' \), as shown by equation (5) and equation (6).

\[ P' = \eta_2 \sum_{i=1}^{C} P \text{Cost}_{i,c} \]

where \( C \) represents the total number of categories, \( i \in \{1..C\} \) represents different age categories. \( P' = \{ P_i | i \in \{1..C\} \} \), \( \sum_{i=1}^{C} P_i = 1.0 \leq P_i \leq 1 \), and \( P'' = \{ P_i'' | i \in \{1..C\} \} \), \( \sum_{i=1}^{C} P_i'' = 1.0 \leq P_i'' \leq 1 \). \( \eta_2 \) is a normalization item, which guarantees \( \sum_{i=1}^{C} P_i'' = 1 \). The predicted category values of a given image determined by the twice probability vector adjustment are \( L'_i = \arg \max_i P'_i \) and \( L''_i = \arg \max_i P''_i \) respectively.

\[
\text{Cost}_{i,c} = \begin{cases} 
(\log(N_i / N_i + 1))^2 & \text{if } c \neq i \\
0 & \text{if } c = i 
\end{cases}
\]

where \( N_i \) and \( N_c \) represent the sample size of corresponding age categories in the training set.

We extract a vector \([P, P', P'']\) for each misclassified image from the minority set as features. The features of all misclassified images are compared with that of the input image, to generate similarity indexes. The cosine similarity of the vectors of a misclassified image and the input image is used as the similarity index, as shown in equation (7).

\[
\cos \theta = \frac{\sum_{i=1}^{C} (A_i \times B_i)}{\sqrt{\sum_{i=1}^{C} (A_i^2) \times \sum_{i=1}^{C} (B_i^2)}}
\]
where $A_i = (AP_i, AFTP_i, ASTP_i)$. $AP_i$ represents the probability $P_i$ that the predicted category value of sample $A$ is $i$, $AFTP_i$ represents the probability $P''_i$ that the predicted category value of sample $A$ is $i$, and $ASTP_i$ represents the probability $P'''_i$ that the predicted category value of sample $A$ is $i$.

We further apply two similarity thresholds $MajoritySimilarityRange$ and $MinoritySimilarityRange$ to determine whether the similarity of the input image and a misclassified image is big enough that the input image should also be marked as misclassified. We calculate the similarity indexes of all images in $Samvalreclassi$ and the $Samtrainmireclassi$ to generate $MajoritySimilarityRange$ and $MinoritySimilarityRange$. The $MajoritySimilarityRange$ represents the similarity range between the correctly classified images in the validation set and the misclassified images in the training set after the probability vector adjustment. The $MinoritySimilarityRange$ represents the similarity range between the misclassified images in the validation set and the training set after the probability vector adjustment.

We also train an ensembled ENIN classifier using the ENIN network structure and the misclassified images in the training set and the validation set after the probability vector adjustment. We create two datasets for each minority age category and we set that the misclassified image number of this minority category is the largest among all the categories. The misclassified images of other categories with more images are randomly selected. We train an ENIN classifier for the two datasets and use the image classification results of the ensembled ENIN classifier to assist the similarity comparison procedure to determine the final classification result of an image.

We further apply two similarity thresholds $MajoritySimilarityRange$ and $MinoritySimilarityRange$ to determine whether the similarity of the input image and a misclassified image is big enough that the input image should also be marked as misclassified. We calculate the similarity indexes of all images in $Samvalreclassi$ and the $Samtrainmireclassi$ to generate $MajoritySimilarityRange$ and $MinoritySimilarityRange$. The $MajoritySimilarityRange$ represents the similarity range between the correctly classified images in the validation set and the misclassified images in the training set after the probability vector adjustment. The $MinoritySimilarityRange$ represents the similarity range between the misclassified images in the validation set and the training set after the probability vector adjustment.

We also train an ensembled ENIN classifier using the ENIN network structure and the misclassified images in the training set and the validation set after the probability vector adjustment. We create two datasets for each minority age category and we set that the misclassified image number of this minority category is the largest among all the categories. The misclassified images of other categories with more images are randomly selected. We train an ENIN classifier for the two datasets and use the image classification results of the ensembled ENIN classifier to assist the similarity comparison procedure to determine the final classification result of an image.

For an input image, we first apply stage 1 and stage 2 of the algorithm, to generate $P$ and $P'$. Then, we apply stage 3 and generate $P''$, which improves the classification accuracies of the under-sampled age categories while decreasing those of other categories. If the classification results of $P'$ and $P''$ agree with each other, we output it as the final result. Otherwise, we apply stage 3 of the algorithm and combine the classification results of similarity comparison and the classifier retraining to determine the final classification result. If any similarity index of the input image is closer to $MinoritySimilarityRange$ than $MajoritySimilarityRange$, we set the classification result of the similarity comparison procedure as the classification result of $P''$. We apply the ensembled ENIN classifier to the input image to generate the probability vector $P_e$ and get the predicted classification results. If the classification result of the input image using ensembled ENIN classifier are consistent with the classification results of the similarity comparison and the predicted classification result belongs to the minority category, we set the final category value as the classification result of $P''$. Otherwise, we believe that the image belongs to the majority age category, and output the classification result of $P'$ as the final result.

5 Experiment and Results
In this section, we use the recall values and the Accuracy-P to display the age estimation effects.

5.1 The Classification Accuracy and the Values of the Accuracy Metrics
The age estimation effects of the various age groups at each step in the FGNET + MORPH dataset are shown in Figure 8. The Recall value indicates the correct classification proportion for a category. The Recall values of most age groups using the ENIN model are higher than those using the NIN model. Combined with the age distribution shown in Table 1, we can see that the ENIN network model achieves better estimation accuracy in the majority age groups (e.g., the Adult age category) than the NIN.

Comparing the results using $P'$ with the those of the ENIN, the Recall values of the minority groups (e.g., the Senior Adult age category) have improved significantly, which is consistent with the ability of the probability vector adjustment to improve the classification accuracies of the minority age groups. However, there is still a relatively high proportion of the minority samples that are misclassified.

$P''$ further skews the age estimation results towards the minority age groups. Most of the minority samples misclassified by $P'$ are correctly classified using $P''$. So far the minority age groups have
achieved high accuracies. However, affected by the probability vector adjustment, the Recall values of the majority age groups drop significantly.

We get the final age estimation results after applying Stage 3 of our algorithm. The Recall values of the minority age groups improve significantly than only using the ENIN model. The final age estimation effects in various age groups are relatively balanced and satisfactory.

To evaluate the overall age estimation effect of our algorithm, we calculate three accuracy metrics—Accuracy-D, Accuracy-Japan, and Accuracy-World at each step. The values of the accuracy metrics in the FGNET + MORPH dataset are shown in Figure 9.

Although the three accuracy metrics are dependent on the classification accuracy of each category in the dataset, they reflect the influence of different age distribution characteristics on the age estimation effect.

![Figure 8](image1.png)  
**Figure 8.** The results of each step of our algorithm in the FGNET + MORPH dataset.

![Figure 9](image2.png)  
**Figure 9.** The values of the accuracy metrics in the FGNET + MORPH dataset.

The Accuracy-Japan represents the age estimation effect in real-world applications using the population age distribution of Japan. As can be seen from Figure 9, the Accuracy-Japan of the ENIN is lower than that of the NIN. After that, the Accuracy-Japan improves at each step of our algorithm. This is due to the change of the classification accuracy of each age category and the population age distribution characteristics of Japan's aging population. The population of the Children is small, and the total population of the Adult and the Senior Adult accounts for nearly 80% of the population of Japan. Therefore, the change of the Accuracy-Japan mainly follows the classification accuracy changes of the Adult and the Senior Adult categories. The Accuracy-Japan of the final result reaches a high value because the classification accuracies of various age groups are all satisfactory.

The Accuracy-World reflects the age estimation effect in the application using the age distribution of the global population. The Accuracy-World improves at each step of our algorithm except the step using \( P'' \), which is also caused by the change of the classification accuracy of each age category and the distribution of the global age population. The population of the Adult age category accounts for over 50% of the total population, and the population distributions of other age groups are relatively balanced. Therefore, the change of the Accuracy-World mainly follows the classification accuracy changes of the Adult age category. The Accuracy-World of the final result also reaches a satisfactory level.

Similar to Figure 8 and Figure 9, Figure 10 shows the Recall values of various age groups in the Adience dataset, and Figure 11 shows the values of the accuracy metrics in the Adience dataset.

As shown in Figure 10, the changing trend of the classification accuracies at each step in the Adience dataset is similar to that in the FGNET + MORPH dataset. The ENIN model performs better than the NIN model. The probability vector adjustments skew the age estimation effects towards the minority age groups and decrease the classification accuracies of the majority age groups. The final age estimation results in various age groups after using our algorithm are relatively balanced.
The change of the values of the accuracy metrics in the Adience dataset is also similar to that in the FGNET + MORPH dataset, as shown in Figure 11. However, the age estimation effects in the Adience dataset are not as good as those in the FGNET + MORPH dataset. This is because the images in the Adience dataset are collected in an unconstrained environment, and the age estimation results are affected by factors such as illumination, posture, and expression, etc.

5.2 Comparison with Existing Approaches
For the FGNET+MORPH dataset, the results of our algorithm are compared with the results of the existing approaches [4-5]. As can be seen from Table 3, our algorithm achieves the highest population-age-distribution based accuracies on the population age distribution in the world and achieves Accuracy-Japan that is very close to the highest Accuracy-Japan, which means that our algorithm performs well for any facial image in real-world applications following the natural population age distribution.

![Figure 10. The results of each step of our algorithm in the Adience dataset.]

![Figure 11. The values of the accuracy metrics in the Adience dataset.]

| Method              | Accuracy-D(%) | Accuracy-Japan(%) | Accuracy-World(%) |
|---------------------|---------------|-------------------|-------------------|
| E+A with L+G [6]    | 88.6          | 86.85             | 89.1              |
| E+S with L+G [6]    | 90.2          | 88.6              | 90.7              |
| SVM with SDM [5]    | 91.1          | 89.2              | 91.2              |
| SVM with BB-FCN [5] | 92.9          | 90.7              | 92.5              |
| KNN with SDM [5]    | 92.3          | 91.1              | 92.53             |
| KNN with BB-FCN [5] | 92.9          | 92.4              | 93.21             |
| Ours                | **93.2**      | 92.3              | 93.6              |

We compare the results of our algorithm in the Adience database with the existing approaches [28-29]. As can be seen from Table 4, our algorithm achieves the highest population-age-distribution based accuracies. Although the Accuracy-D of our algorithm is lower than that of the algorithm Levi & Hassncer [28] proposed, our algorithm performs better in real-world applications.

| Method                  | Accuracy-D(%) | Accuracy-Japan(%) | Accuracy-World(%) |
|-------------------------|---------------|-------------------|-------------------|
| Appearance [29]         | 38.3          | 36.9              | 31.3              |
| Appearance+Context [29] | 42.9          | 39.4              | 36.3              |
| Best from Ref. [28]     | **54.5**      | 37.9              | 42.9              |
| Ours                    | 49.24         | **43.87**         | **45.43**         |

6 Conclusions and Future Works
In this paper, we propose a three-stage probability adjustment based CNN algorithm for facial-image based age estimation using imbalanced datasets. We construct an ENIN neural network and use the probability vector adjustment to improve the age estimation effect of the under-sampled age categories.
After that, the category redefinition combined with the cosine similarity comparison and the ensemble ENIN classifier retraining achieves balanced and satisfactory age estimation effects in various age groups. We also introduce a population-age-distribution based accuracy metric $\text{Accuracy-P}$ to estimate the performance of the age estimation algorithm in the real-world application scenarios with the natural population age distribution. The age estimation effects of the algorithm in the two datasets FGNET+MORPH and Adience show that our algorithm achieves satisfactory age estimation accuracy and is suitable for the real-world applications.

Acknowledgments
This work was supported by the [National Key R&D Program of China] under Grant [No. 2017YFC1601800].

References
[1] Cootes, T. F, Edwards, G. J and Taylor, C. J 2001 Pattern Analysis & Machine Intelligence IEEE Transactions on 23 681-85
[2] Geng, X, Zhou, Z. H and Smithmiles, K 2007 IEEE Transactions on Pattern Analysis & Machine Intelligence 29 2234-40
[3] Chi, N, Quach, K. G, Luu, K, Le, H. B and Ricanek, K 2011 Proceedings of the 36th Intl. Conf. on Acoustics, Speech and Signal Processing
[4] Kwon, Y. H and Lobo, N. D V 1999 Computer Vision & Image Understanding 74 1-21
[5] Huang, J, Li, B, Zhu, J and Chen, J 2017 Multimedia Tools & Applications 76 1-17
[6] Wang, J. G, Yau, W. Y and Wang, H. L 2009 Proceedings of the Applications of Computer Vision
[7] Buda, M, Maki, A and Mazurowski, M. A 2017 A systematic study of the class imbalance problem in convolutional neural networks
[8] Chawla, N. V, Bowyer, K. W, Hall, L. O and Kegelmeyer, W. P 2002 Journal of Artificial Intelligence Research 16 321-57
[9] Mease, D, Wyner, A. J and Buja, A 2007 Journal of Machine Learning Research 8 409-39
[10] Drummond, C and Holte, R. C 2003 Proc of the Icml Workshop on Learning from Imbalanced Datasets II 1-8
[11] He, H and Garcia, E. A 2009 IEEE Transactions on Knowledge & Data Engineering 21 1263-84
[12] Han, H, Wang, W. Y and Mao, B. H 2005 Proceedings of the International Conference on Advances in Intelligent Computing 878-87
[13] Bunkhumpornpat, C, Sinapiromsaran, K and Lursinsap, C 2009 Proceedings of the 13th Pacific-Asia Conference on Knowledge and Data Mining 475-82
[14] He, H, Bai, Y, Garcia, E. A and Li, S 2008 Proceedings of the IEEE International Joint Conference on Neural Networks 1322-28
[15] Jo, T and Japkowicz, N 2004 Acm Sigkdd Explorations Newsletter 6 40-9
[16] Liu, X. Y, Wu, J and Zhou, Z. H 2009 IEEE Transactions on Systems Man & Cybernetics Part B 39 539-50
[17] Kubat, M and Matwin, S 1997 Proceedings of the 14th International Conference on Machine Learning 179–86
[18] Elkan, C 2001 Proceedings of the 17th International Joint Conference on Artificial Intelligence 973-78
[19] Zhou, Z. H and Liu, X. Y 2006 IEEE Transactions on Knowledge & Data Engineering 18 63-77
[20] Khan, S. H, Hayat, M, Bennamoun, M, Sohel, F. A and Togneri, R 2018 IEEE Transactions on Neural Networks & Learning Systems 29 3573-78
[21] McCann, B and Weiss, G 2005 Proceedings of the 1stInternational Workshop on Utility-Based Data Mining 69-77
[22] FG-NET (Face and Gesture Recognition Network). http://www.fgnet.rsunit.com/.
[23] Jr, K. R and Tesafaye, T 2006 *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition* 341-45

[24] Eidinger, E, Enbar, R and Hassner, T 2014 *IEEE Transactions on Information Forensics & Security* 9 2170-79

[25] Population Pyramids (https://www.populationpyramid.net).

[26] Baggio, D. L 2012 Mastering opencv with practical computer vision projects

[27] Lin, M, Chen, Q and Yan, S 2013 Network in Network

[28] Levi, G and Hassner, T 2015 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 34-42

[29] Gallagher, A. C and Chen, T 2009 *Proceedings of the IEEE-Computer-Society Conference on Computer Vision and Pattern Recognition (CVPR)* 256-63