Age Estimation Based on A Single Network with Soft Softmax of Aging Modeling

Zichang Tan[1,2], Shuai Zhou[1,3], Jun Wan[1,2*], Zhen Lei[1,2], and Stan Z. Li[1,2]

[1]Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences
[3]Faculty of Information Technology,
Macau University of Science and Technology, Macau

Abstract. In this paper, we propose a novel approach based on a single convolutional neural network (CNN) for age estimation. In our proposed network architecture, we first model the randomness of aging with the Gaussian distribution which is used to calculate the Gaussian integral of an age interval. Then, we present a soft softmax regression function used in the network. The new function applies the aging modeling to compute the function loss. Compared with the traditional softmax function, the new function considers not only the chronological age but also the interval nearby true age. Moreover, owing to the complex of Gaussian integral in soft softmax function, a look up table is built to accelerate this process. All the integrals of age values are calculated offline in advance. We evaluate our method on two public datasets: MORPH II and Cross-Age Celebrity Dataset (CACD), and experimental results have shown that the proposed method has gained superior performances compared to the state of the art.

1 Introduction

Until today, age estimation is still a very complex pattern classification problem. We judge a person's age mainly through the shin sheen, smooth degree, wrinkles and others, and such appearances are closely linked with various genes, diets, living and working environment and so on, which make age prediction further complicated. Initial work on age estimation goes back to 1990s [1], which simply classifies face images into several age groups based on facial shape features and skin wrinkle analysis. After this work, age estimation attracts more and more scholars' attention. Moreover, in the initial stage of age estimation research, the available age dataset is extremely limited. Fortunately, with the effort of the scholars all over the world, many large datasets are available for age estimation, like FG-NET [2], MORPH II [3], CACD [4], which increases by hundreds of times compared with datasets in the initial stage of age estimation research and significantly promote the development of age estimation.

* corresponding author, jun.wan@ia.ac.cn

Aging is a continuous process and the boundaries between adjacent ages are not obvious. Firstly, each person has different aging speeds and people of the same age may appear to be slightly older or younger comparing with each other. For example, two faces come from different people of the same age may look like in different ages. Secondly, aging is a very slow process and faces at close ages would look similar. We believe that faces labeled with particular age are also close those with neighboring ages and even could be labeled with multiple labels in some way. For instance, Geng et al. [5] treated each face image with a age label distribution rather than a single label and thus during learning, it judged each face by considering its real age and adjacent ages. To be more correct, the label of age is more like to be a set of soft labels rather than a specific evidence, taking various factors related to aging into account.

Therefore, we propose an age estimation framework for exploring the aging information based on the CNN framework. The aging model is embedded into the network to explore more efficient features for age estimation. The main contributions of our work are summarized below:

- We model the randomness of aging with the Gaussian distribution for the chronological age. It is used to calculate the Gaussian integral of an age interval around the true age.
- We propose a new loss function: soft softmax regression function. The new function applies the aging modeling to compute the loss in the training phase. Compared with the classic softmax function, the new function considers the age interval instead of the specific age value.
- Compared with the softmax function, the proposed CNN frameworks with soft softmax function can alleviate the overfitting problems according to our experiments.
- Because of the complexity of Gaussian integral, a look up table is built to accelerate this process.

The rest of the paper is organized as follows. Related works are reviewed in Section 2. The proposed method is presented in Section 3. Then, experiments are provided in Section 4 to evaluate our method and compare with the state-of-the-art methods. Section 5 gives some discussions about the proposed method. Finally, a conclusion is drawn in Section 6.

2 Related works

Early methods for age estimation just classify facial images into several age groups according to some hand-crafted features based on facial geometry features and skin wrinkle analysis [1]. The facial geometry features mainly consist of geometric relationships that computed by the sizes and distances between some primary features (e.g. eyes, mouth, nose, etc.). Facial geometry features are used to distinguish babies, and skin wrinkle feature can distinguish young adults from senior adults. Few years later, on the basis of the former works, Horng et al. [6]

locate eyes, mouths and noses in face images via sobel edge operator and region labeling, then extract geometric and wrinkle features for age estimation.

In recent years, fortunately, automatic human age estimation receives increasing attention and more and more new methods have been proposed along with the development of the facial analysis technology. Geng et al. [7, 8] proposed AGing pattErn Subspace (AGES) approach to model the aging pattern, which achieves the mean absolute error (MAE) on FG-NET database to 6.22 years. Moreover, many methods were proposed for age estimation based on manifold learning [9–11]. Those methods firstly learned facial age features in low-dimensional representation with manifold learning, then defined a regression function to fit manifold data for further age prediction. For example, Guo et al. [9] introduced the age manifold learning scheme to extract facial age features, then proposed locally adjusted robust regressor (LARR) method to predict age for face images, which improved performance significantly and reduces MAE on FG-NET to 5.07 years. More recently, local features become very popular for age estimation, such as Gabor [12], Local Binary Patterns (LBP) [13], Biologically-Inspired Features (BIF) [14]. After features extracted by those local image descriptors, classification or regression methods would be used for predicting the age, such as BIF+SVM [14], BIF+SVR [14], BIF+CCA [15].

In the last few years, the CNN has made a lot of progress in age estimation [16–19]. Comparing with traditional methods, CNN learns useful features autonomously instead of hand-crafted ways. Yi et al. [17] designed 46 parallel CNNs with multi-scale facial image patches as input for age estimation, which reduced the MAE to 3.63 years in MORPH II database and achieved the state-of-the-art performance. The parallel CNNs need pre-partion of facial images into different parts according to facial landmarks and predefined scales, and each part would be processed by a separated CNN. Rothe et al. [16] proposed a new method called Deep EXpectation (DEX) model based on VGG-16 network, which won the 1st place at the ChaLearn LAP challenge 2015. However, such deep CNN needs to be pretrained with a large age database and Rothe et al. collected 0.5 million images to do that.

Multi-label learning (MLL) [20] is also a hot topic especially in age estimation research in recent years. MLL assigns training instances with a set of labels rather than a single label to solve the problem of label ambiguity. Geng et al. [5, 21] labeled each facial image with multiple age labels followed a label distribution, showing the advantages dramatically over the methods with single-label.

In this paper, our work is also inspired by the MLL and CNN. We combine the advantages of both MLL and CNN, and propose a novel method with aging information for age estimation.

3 The Proposed Method

In the proposed method, we first model the randomness of aging with Gaussian distribution. Then, we present a soft softmax function using aging modeling. Compared with the softmax function, the new function is more efficient for age

estimation. Later, we introduce a simple CNN framework which is similar to AlexNet [22]. Finally, we present a fast calculation method based on a look up table to accelerate the processing time for Gaussian integral.

3.1 Modeling the Randomness of Aging

For MLL problem, each image would be allowed to be labeled by multiple labels. Following the works [5, 21], each face image is labeled by an age label distribution that is a set of possibilities which represent the description degrees corresponding to each label. One face image x includes some possible discrete age labels $L = \{l_1, ..., l_k\}$ in our aging model. Let $P(l_c, l_i)$ denotes the probability of the possible label l_i corresponding to the chronological age l_c, where $P(l_c, l_i) \in [0\ 1]$, $\sum_{i=1}^{k} P(l_c, l_i) = 1$. $P(l_c, l_i)$ is the maximum value when l_i is equal to l_c.

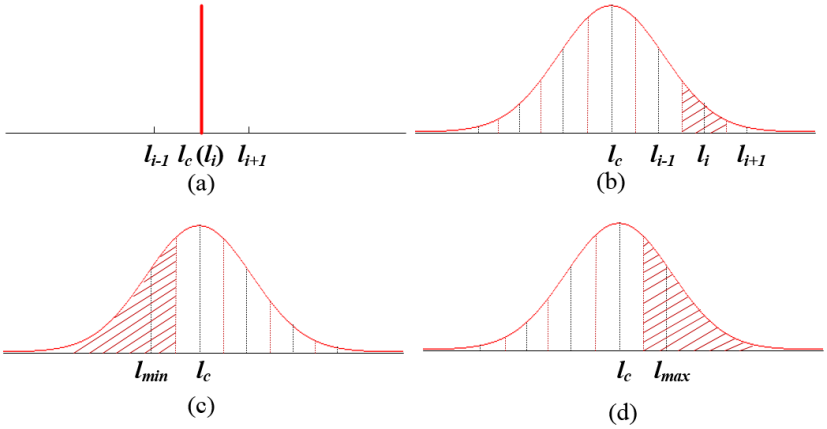

Fig. 1. (a) the Kronecker function $\delta(l_c, l_i)$, $l_c = l_i$; (b) the integral $P(l_c, l_i)$ of gaussian distribution ($l_{min} < l_i < l_{max}$); (c) the integral $P(l_c, l_i)$ of gaussian distribution ($l_i = l_{min}$); (d) the integral $P(l_c, l_i)$ of gaussian distribution ($l_i = l_{max}$).

There are some works [5, 21] that allow each face image to be labeled by a Gaussian distribution for age estimation. Following those works, We also model the aging problem with the Gaussian distribution which is used to calculate the integral of an age interval. The formal formulation of aging model is given in Eq. 1.

$$P(l_c, l_i) = \begin{cases} \int_{-\infty}^{l_i+0.5} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-l_c)^2}{2\sigma^2}} \mathrm{d}x & \text{if } l_i = l_{min} \\ \int_{l_i-0.5}^{l_i+0.5} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-l_c)^2}{2\sigma^2}} \mathrm{d}x & \text{if } l_{min} < l_i < l_{max} \\ \int_{l_i-0.5}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-l_c)^2}{2\sigma^2}} \mathrm{d}x & \text{if } l_i = l_{max} \end{cases} \quad (1)$$

where $l_i \in [l_{min}\ l_{max}]$ is the discrete age label, l_c denotes chronological age, l_{min} and l_{max} is the minimum and maximum age, respectively.

In Fig. 1 (b-d), it shows the calculation of integrals via Eq. 1. Also, we note that if one face image with a single age label only, it also has a special distribution with $P(l_c, l_i) = \delta(l_c, l_i)$, where δ is the Kronecker function shown in Fig. 1 (a).

3.2 Soft Softmax Regression Function

The traditional softmax regression function usually applied remarkably in the field of classification with neural networks [22]. There are also some methods [16, 18] for age estimation with the softmax function, which have achieved promising results. However, age estimation is not a simple classification problem because aging is a continuous process. Also, face images at close age would look very similar. On the other hand, each person has different aging speeds because of many intrinsic and extrinsic factors, in which people with the same age would have various aging features. Hence, different age classes are related rather than independent.

From those observations, we believe that the problem of age estimation can be considered with the aging information as described in Section 3.1. Rothe et al. [16] calculated the expected value among softmax output probabilities and their corresponding age as final predicted age, which achieved better results comparing with the predict age having the maximum probability of the softmax output. This refinement fuses not only the real age's information but also other ages' in the prediction phase. In this section, we design a soft softmax regression function to push the network learn from both truth ages and their age intervals in Eq.1 in the training phase.

For the age estimation, $x_c \in R^d$ denotes the CNN output features for the c^{th} sample, and $l_c \in \{l_{min}, ..., l_{max}\}$ is its corresponding age labels. Here, we set $l_{min} = 0,\ l_{max} = k$) for the convenient description. Given a training set includes m samples $S = \{(x_0, l_0), ..., (x_c, l_c), ..., (x_{m-1}, l_{m-1})\}$, $c \in [0\ m-1]$, the soft softmax loss function with the aging model is defined as:

$$J(\theta) = -\frac{1}{m}\left[\sum_{c=0}^{m-1}\sum_{i=0}^{k} P(l_c, i) log \frac{e^{\theta_i^T x_c}}{\sum_{j=0}^{k} e^{\theta_j^T x_c}}\right] \tag{2}$$

where θ is the parameter matrix of the soft softmax function; $P(l_c, l_i)$ considers the probabilities of the specific label l_c and its adjacent labels. And when $P(l_c, l_i) = \delta(l_c, l_i)$, the soft softmax function becomes the traditional softmax function.

Then we can calculate the gradient formula of Eq. 2 as:

$$\bigtriangledown_{\theta_v} J(\theta) = -\frac{1}{m}\left[\sum_{c=0}^{m-1}\sum_{i=0}^{k} P(l_c, l_i) x_c \left(1\{v = i\} - \frac{e^{\theta_v^T x_c}}{\sum_{j=0}^{k} e^{\theta_j^T x_c}}\right)\right] \tag{3}$$

where $\bigtriangledown_{\theta_v}$ is the gradient vector of the soft softmax parameters for age v, $1\{\bullet\}$ is indicator function which means $1\{v = j\} = 1$ if and only if $v = j$.

Therefore, the optimized parameters θ can be obtained via the SGD algorithm in the proposed method.

3.3 The Network Architecture

Our convolutional neural network is shown in Fig. 2, which is similar to the AlexNet [22]. The input of the network is RGB facial image with size 224×224. The network includes 5 convolutional layers, 3 max pooling layers and a fully connected layer. The filter size of each layer is also shown in Fig. 2. All the convolutional layers are followed by Rectified Linear Units (ReLU). The network are optimized by Stochastic Gradient Descent (SGD).

We train the network with proposed soft softmax loss function with 101 output neurons corresponding to age numbers from 0 to 100. And it is ok when the dataset lacks of samples corresponding to the output neuron. There are two ways to predict the age value. For the first way, the predicted value can be obtained via the maximum probability of the softmax output. When we used this way to calculate MAE, we call this way as MAE with maximum probability (MP). For the second way, we can conduct a softmax expected value refinement [16] to improve the accuracy, and the final predicting age is $\sum_{i=0}^{k} p_i y_i$, where p_i is the predicting probability of the corresponding age y_i. We call the second way as MAE with expected value (EV).

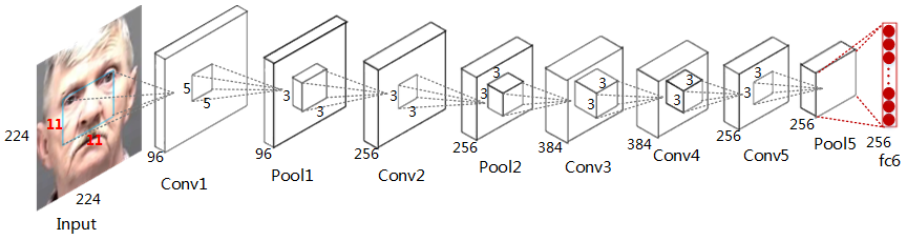

Fig. 2. The network architecture is used in the proposed method.

3.4 Look Up Table for Fast Calculation of Integrals

The integral of Gaussian distribution function among the interval $[a,\ b]$ is:

$$\int_a^b f(x;\mu,\sigma)\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma}\int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{4}$$

where μ is the expected value and σ^2 is the variance of the distribution.

The Gaussian integral in Eq. 4 is usually efficiently calculated by error function $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$ [23] . More specifically, the gaussian intergal can be

calculated with erf as following:

$$\begin{aligned}\int_a^b f(x;\mu,\sigma)\mathrm{d}x &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{b} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x - \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{a} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \\ &= \frac{1}{2}\left[1+erf\left(\frac{b-\mu}{\sqrt{2}\sigma}\right)\right] - \frac{1}{2}\left[1+erf\left(\frac{a-\mu}{\sqrt{2}\sigma}\right)\right] \\ &= \frac{1}{2}\left[erf\left(\frac{b-\mu}{\sqrt{2}\sigma}\right) - erf\left(\frac{a-\mu}{\sqrt{2}\sigma}\right)\right]\end{aligned} \tag{5}$$

From Eq. 5, we can see that each integral calculation need to call erf twice. However, the proposed method needs to calculate integrals both in forward and backward propagation phases for each age label and each face image, which would cost a lot of time accumulatively. Therefore, we prepare an integral table that stores multi-part integrals for each age, thus tedious calculation can be avoided through this table to find corresponding integrals. The consume time of two methods will be discussed in Section 5.

4 Experiments

4.1 Datasets

We evaluated the proposed method on MORPH II [3] and CACD [4] datasets, which are available standard datasets for facial age estimation. Some samples from both datasets are shown in Fig. 3.

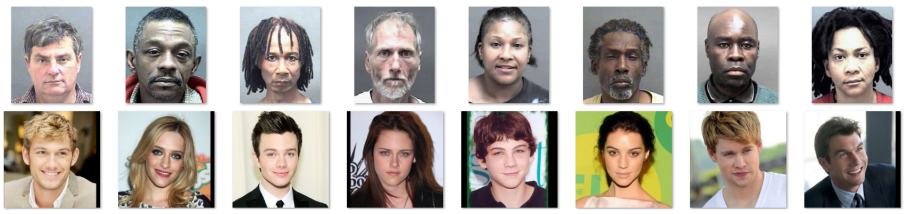

Fig. 3. Face samples from the MORTH II (see the first row) and CACD (see the second row) datasets.

MORPH II includes about 55,000 face images and age ranges from 16 to 77 years. It provides the personal information, such as age, gender, and ethnicity. This dataset is more abundant in the age information, but the faces are recorded under uneven illumination.

CACD is collected from Internet Movie DataBase (IMDB), and it is the largest public cross-age database. This database includes more than 160 thousands images of 2000 celebrities taken from 2004 to 2013 (10 years in total). The age ranges from 16 to 62. Compared with MORPH II, CACD has the biggest total quantity and average number of each subject.

Fig. 4. It shows face samples from the MORPH II dataset in the first row, and their corresponding results of face alignment in the second row.

4.2 Experimental Setting

In our experiments, all images are resized into 224×224. We use SGD and mini-batch size of 64. The learning rate starts from 0.001, and the models are trained for up to 300000 iterations. We use a weight decay of 0.0005 and a momentum of 0.9.

We follow the work [17, 24] to split MORPH II into three non-overlapped subsets S_1, S_2, S_3 randomly. These three subsets are constructed by two rules: 1) Male-Female ratio is equal to three; 2) White-Black ratio is equal to one. In our experiments, we totally use the same test protocols[1] provided by Yi et al. [17]. That is all experiments are repeated two times: 1) Training set: S1, testing sets: S2+ S3; 2) Training set: S2, testing sets: S1+ S3.

Only a few works [25] conducted evaluation on the CACD database owing to its noise. Note that only 200 celebrities of the database are checked and their noisy images are removed, and images of other celebrities contain much noises. Thus, these 200 celebrities are used for testing and the others for training in our experiments.

In our experiments, all images would be processed by a face detector [26] and non-face images would be removed. After processing, there are 55244 images in MORPH II and 162941 images in CACD. Then, we use active shape models (ASM) [27] to detect the facial landmarks and all facial images would be aligned and cropped via the locations of the eyes center and the upper lip (see Fig. 4). When evaluating on MORPH II, images for training is about a quarter of testing images, which is extremely insufficient. Therefore, we augment training images with flipping, rotating by $\pm 5^0$ and $\pm 10^0$, and adding Gaussian white noises with variance of 0.001, 0.005, 0.01, 0.015 and 0.02.

To further improve the performance on MORPH II, our network (with the softmax function) are pretrained on IMDB-WIKI [16]. Note that we don't con-

[1] http://www.cbsr.ia.ac.cn/users/dyi/agr.html

duct such operation for CACD evaluation because some images from IMDB-WIKI and CACD are duplicated.

4.3 Parameters Discussion

For aging modeling, the parameter σ controls the shape of Gaussian distribution at each age. σ is smaller, the Gaussian distribution is sharper and neighboring ages contribute less to the learning of chronological age in the training stage. Likewise, the contribution of neighboring ages would increase as σ rises.

To find an appropriate value for σ, we conducted experiments with a variety of σ on MORPH II and the results are shown in the Table 1. As we can see, the results is not sensitive to the parameter $\sigma \in [0.5, 1.2]$ and the well-done performance can be achieved when σ is about to 1. Thus, we set σ to 1 in our following experiments.

Moreover, we can see that the performance of MAE with EV is better than MAE with MP used the same value σ in Table 1. The same conclusion will also be shown in the next section.

Table 1. Results with different σ on the MORPH II dataset (the lower the better). The top 2 performances are shown in boldface, which are from the average MAE with MP under training set $S1$ and $S2$ respectively. From the best performances, we set $\sigma = 1$ in our experiments.

σ	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
MAE with MP (Train:S_1, Test:S_2+S_3)	3.37	3.37	3.33	3.31	3.30	3.27	3.30	3.30
MAE with MP (Train:S_2, Test:S_1+S_3)	3.06	3.05	3.09	3.04	3.06	3.05	3.07	3.04
Average MAE with MP	3.215	3.21	3.21	3.175	3.18	**3.16**	3.185	**3.17**
MAE with EV (Train:S_1, Test:S_2+S_3)	3.28	3.29	3.26	3.25	3.25	3.24	3.26	3.25
MAE with EV (Train:S_2, Test:S_1+S_3)	3.01	3.01	3.05	3.01	3.04	3.03	3.05	3.04
Average MAE with EV	3.145	3.15	3.155	**3.130**	3.145	**3.135**	3.155	3.145

4.4 Comparisons

Results on the MORPH II dataset We conduct our experiments with the softmax and soft softmax regression function respectively and the results are shown in Table 2. We can see that the results of the soft softmax function are superior than the softmax function under MAE with MP or EV. Without pretrained model, the average MAE is 3.16 with MP and 3.14 with EV.

Table 2. Results based on the CNN mentioned in Section 3.3 with different objective functions on the MORPH II dataset (the lower the better). It shows that the best performances are from the soft softmax function with pretrained model whenever MAE with MP or EV is used.

Methods	Train Set	Test Set	MAE with MP	Avg. MAE with MP	MAE with EV	Avg. MAE with EV
softmax	$S1$	$S2+S3$	3.45	3.28	3.28	3.16
	$S2$	$S1+S3$	3.10		3.03	
soft softmax	$S1$	$S2+S3$	3.27	3.16	3.24	3.14
	$S2$	$S1+S3$	3.05		3.03	
pretrained model, softmax	$S1$	$S2+S3$	3.34	3.20	3.19	3.08
	$S2$	$S1+S3$	3.06		2.97	
pretrained model, soft softmax	$S1$	$S2+S3$	3.19	3.06	3.14	**3.03**
	$S2$	$S1+S3$	2.93		2.92	

Table 3. Comparisons with the state-of-the-art methods on MORPH II under the same testing protocol (the lower the better).

Methods	Train Set	Test Set	MAE	Avg. MAE
Our method	$S1$	$S2+S3$	3.14	3.03
	$S2$	$S1+S3$	2.92	
Multi-scale CNN [17]	$S1$	$S2+S3$	3.72	3.63
	$S2$	$S1+S3$	3.54	
BIF+KCCA [15]	$S1$	$S2+S3$	4.00	3.98
	$S2$	$S1+S3$	3.95	
BIF+KPLS [28]	$S1$	$S2+S3$	4.07	4.04
	$S2$	$S1+S3$	4.01	
BIF+rCCA [15]	$S1$	$S2+S3$	4.43	4.42
	$S2$	$S1+S3$	4.40	
BIF+PLS [28]	$S1$	$S2+S3$	4.58	4.56
	$S2$	$S1+S3$	4.54	
CNN [29]	$S1$	$S2+S3$	4.64	4.60
	$S2$	$S1+S3$	4.55	
BIF+KSVM [15]	$S1$	$S2+S3$	4.89	4.91
	$S2$	$S1+S3$	4.92	
BIF+LSVM[15]	$S1$	$S2+S3$	5.06	5.09
	$S2$	$S1+S3$	5.12	
BIF+CCA [15]	$S1$	$S2+S3$	5.39	5.37
	$S2$	$S1+S3$	5.35	

To further improve the performance, we also pretrain the model on the IMDB-WIKI database, and it achieves the best performance with the soft softmax function and EV, which the average MAE is 3.03. Compared with the softmax method, the pretrained model with soft softmax reduce the average MAE dramatically from 3.20 to 3.06 and from 3.08 to 3.03 with MP and EV respectively.

As shown in Table 2, the results show that the soft softmax regression function is superior than the softmax function for age estimation whenever MP, EV or the pretrained model is used.

Then, we compared our method with the state-of-the-art methods. The comparisons are shown in Table 3 under the same testing protocol. We can see that our method can achieve the best performance with the average MAE of 3.03, which is reduced 0.6 comparing with the previous best method [17] with the average MAE of 3.63. Specially, whenever training with S1 and testing with S2+S3 or training with S2 and testing with S1+S3, our method is the best with MAE of 3.14 and 2.92, respectively.

Results on the CACD dataset The images in this database are taken in the unconstrained environment, which are more close to the real life. We conduct the experiments on this database with the softmax and soft softmax regression function. The experimental results are shown in table 4. The soft softmax method can achieve better result than the softmax method whenever using MP or EV. Our method reduces the MAE to 5.22 with MP and 5.19 with EV. And the proposed method is also better than the DFDNet method [25] which has achieved 5.57 in MAE.

Table 4. Results based on the CNN mentioned in Section 3.3 with different objective functions on the CACD dataset (the lower the better).

Methods	Train Set	Test Set	MAE with MP	MAE with EV
softmax	1800 celebrities	200 celebrities	5.43	5.28
soft softmax (ours)	1800 celebrities	200 celebrities	5.22	**5.19**

5 Discussion

5.1 Anti-overfitting Analysis

Fig. 5 (a) shows the loss trend when training on S1 and testing on S2+S3 on the MORPH II dataset. Compared with the soft softmax regression function, the training loss decreases sharper used the softmax function after 15000 iterations. It indicates that the overfitting problem may be more serious when the softmax function is used.

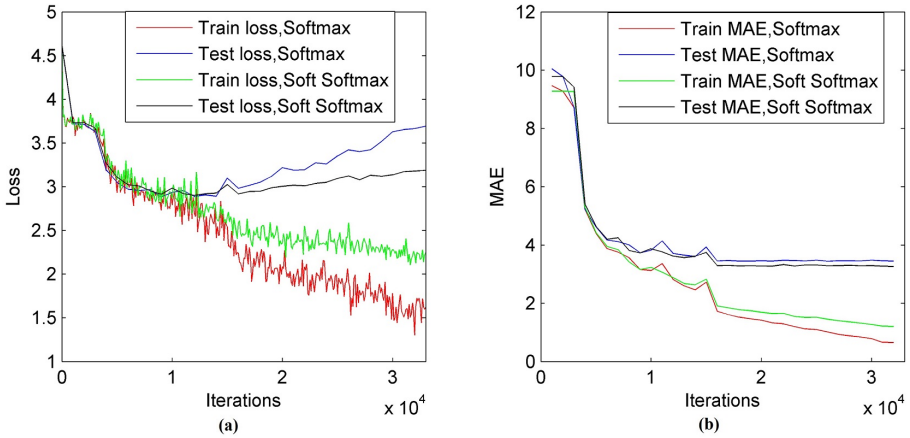

Fig. 5. (a) It shows the training and testing loss by iterations; (b) It shows the MAE trend by iterations.

Moreover, both two loss functions are designed to reducing testing error for age estimation. Thus, we draw the MAE trends as well, which are shown in Fig. 5 (b). With the softmax function, training MAE are lower while testing MAE are higher than the soft softmax function. Obviously, the model with the softmax function is more likely to be overfitted. That is because it has poor performances compared with the soft softmax function.

Form the above discussions, the soft softmax regression function has the anti-overfitting characteristic in some way.

5.2 Computational Time Analysis

In this section, we mainly compare the consuming time among two methods for integrals calculation in the training process, looking up table and Gaussian integral calculation with the error function [23]. Our comparative experiments conducted under the same conditions with GTX TITAN X GPU and the results are shown in Table 5. Compared with online Gaussian integral calculation, the way of look up table reduce the consuming time by about 18.84%. In the training process with Gaussian integral calculation with the error function, for each instance, we need to calculate the interval for each age in the forward and backward pass phase, and each interval calculation need to call error function twice, which is very time consuming.

6 Conclusion

In this paper, we proposed a novel approach based on a single CNN for age estimation. First, the randomness of aging is modeled by the Gaussian integral that

Table 5. Comparison of the execution time between look up table method and direct Guassian integral.

Methods	Look Up Table (Our)	Online Gaussian Integral Calculation
Time/10,000 iterations	1702 seconds	2097 seconds

not only considers the chronological age but also includes the age intervals nearby the truth age. Second, we present a soft softmax regression function instead of classic softmax function, which is combined with aging model. Moreover, to further speed up the computation of gaussian integral, we build a look up table to store pre-compute gaussian integrals. So this way only requires one memory access from the look up table. Evaluations on two age datasets show that the proposed method achieves state-of-the-art performances.

Acknowledgement. This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects ♯61473291, ♯61572501, ♯61502491, ♯61572536, Science and Technology Development Fund of Macau (No. 019/2014/A1), NVIDIA GPU donation program and AuthenMetric *R&D* Funds.

References

1. Young HoKwon, N.D.V.: Age classification from facial images. Computer Vision and Image Understanding **74** (1999) 1–21
2. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **34** (2004) 621–628
3. Rawls, A.W., Ricanek Jr, K.: Morph: Development and optimization of a longitudinal age progression database. In: Biometric ID Management and Multimodal Communication. Springer (2009) 17–24
4. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: Computer Vision–ECCV 2014. Springer (2014) 768–783
5. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35** (2013) 2401–2412
6. Horng, W.B., Lee, C.P., Chen, C.W.: Classification of age groups based on facial features. Tamkang Journal of Science and Engineering **4** (2001) 183–192
7. Xin Geng, Zhihua Zhou, K.S.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2234–2240
8. Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM (2006) 307–316

9. Guodong Guo, Yun Fu, C.R.T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing **17** (2008) 1178–1188
10. Fu, Y., Xu, Y., Huang, T.S.: Estimating human age by manifold analysis of face pictures and regression on aging features. In: Multimedia and Expo, 2007 IEEE International Conference on, IEEE (2007) 1383–1386
11. Yun Fu, T.S.: Human age estimation with regression on discriminative aging manifold. IEEE Transactions on Multimedia **10** (2008) 578–584
12. Gao, F., Ai, H.: Face age classification on consumer images with gabor feature and fuzzy lda method. In: Advances in biometrics. Springer (2009) 132–141
13. Günay, A., Nabiyev, V.V.: Automatic age classification with lbp. In: Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on, IEEE (2008) 1–4
14. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 112–119
15. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: Cca vs. pls. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE (2013) 1–6
16. Rothe, R., Timofte, R., Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 10–15
17. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Computer Vision–ACCV 2014. Springer (2014) 144–158
18. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2015) 34–42
19. Kuang, Z., Huang, C., Zhang, W.: Deeply learned rich coding for cross-dataset facial age estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 96–101
20. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006)
21. Geng, X., Wang, Q., Xia, Y.: Facial age estimation by adaptive label distribution learning. In: 2014 22nd International Conference on Pattern Recognition (ICPR), IEEE (2014) 4465–4470
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
23. Andrews, L.: (Special functions of mathematics for engineers (1998))
24. Guo, G., Mu, G.: Human age estimation: What is the in fluence across race and gender? In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2010) 71–78
25. Liu, T., Lei, Z., Wan, J., Li, S.Z.: Dfdnet: Discriminant face descriptor network for facial age estimation. In: Biometric Recognition. Springer (2015) 649–658
26. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001 IEEE Conference on. Volume 1., IEEE (2001) I–511
27. Cootes, T.F., Taylor, C.J., Cooper, D.M.L., Graham, J.: Active shape modelstheir training and application. Computer Vision and Image Understanding **61** (1995) 38–59

28. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 657–664
29. Yang, M., Zhu, S., Lv, F., Yu, K.: Correspondence driven adaptation for human profile recognition. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 505–512