Pattern Recognition 66 (2017) 106–116

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Diagnosing deep learning models for high accuracy age estimation from a single image

Junliang Xing[a,*], Kai Li[b], Weiming Hu[a,b], Chunfeng Yuan[a], Haibin Ling[c]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China
[b] CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100190, PR China
[c] Dept. of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

ARTICLE INFO

Keywords:
Age estimation
Deep learning
Multi-task learning

ABSTRACT

Given a face image, the problem of age estimation is to predict the actual age from the visual appearance of the face. In this work, we investigate this problem by means of the deep learning techniques. We comprehensively diagnose the training and evaluating procedures of the deep learning models for age estimation on two of the largest datasets. Our diagnosis includes three different kinds of formulations for the age estimation problem using five most representative loss functions, as well as three different architectures to incorporate multi-task learning with race and gender classification. We start our diagnoses process from a simple baseline architecture from previous work. With appropriate problem formulation and loss function, we obtain state-of-the-art performance with the simple baseline architecture. By further incorporating our newly proposed deep multi-task learning architecture, the age estimation performance is further improved with high-accuracy race and gender classification results obtained simultaneously. With all the insights gained from the diagnosing process, we finally build a deep multi-task age estimation model which obtains a MAE of 2.96 on the Morph II dataset and 5.75 on the WebFace dataset, both of which improve previous best results by a large margin.

1. Introduction

Age estimation, i.e., predicting the age from a face image, has long been a challenging problem in computer vision, with many applications like precision advertising, intelligent surveillance, face retrieval and recognition. The main challenges of this problem come from the fact that faces may be shot from people of different races, genders, and under conditions of large pose variations, bad illuminations, and spurious makeups [1]. Human beings ourselves can only give a very rough estimation of the age by only looking at the face.

Classic age estimation methods usually involve two consecutive but relatively independent procedures, feature extraction from the face image and age estimation from the feature. The objective of the feature extraction procedure is to extract invariant features representing the aging information. Many different kinds of features have been used in previous works, such as the local binary pattern (LBP) [2] and the Gabor features [3]. With the extracted features, general machine learning algorithms like the Support Vector Machine [4] can be used to predict the age.

The age estimation accuracy of the above classic methods heavily depends on the manually designed features and the employed learning algorithms. Both selections of the designed feature and learning algorithm need many experiences and efforts. Recently, with the fast development of Convolutional Neural Network (CNN), feature representation and classification model can be effectively learned from end to end. Although deep learning has been successfully applied to many computer vision problems, there are very few studies on how to build a high accuracy deep age estimation model, especially on digging out the underlying oracles for building such a model.

To this end, we intend to perform a comprehensive diagnosis of the deep learning models for the age estimation task, and try to find out the most important and effective factors behind the building of the model. To speedup training and testing of deep age models under different configurations, we design a baseline architecture inspired by [5] as the basic component to start the diagnosis process. Starting from this baseline architecture, we have diagnosed different aspects to build the deep age estimation model, including the types of model formulation, the choices of loss function, as well as the strategies to incorporate information like race and gender via multi-task learning. Our diagnosing studies are helpful to getting better understandings of a deep age estimation model. Moreover, by accumulating the insights from all these investigations, we finally obtain a very deep age estimation model

* Corresponding author.
E-mail addresses: jlxing@nlpr.ia.ac.cn (J. Xing), kai.li@nlpr.ia.ac.cn (K. Li), wmhu@nlpr.ia.ac.cn (W. Hu), cfyuan@nlpr.ia.ac.cn (C. Yuan), hbling@temple.edu (H. Ling).

http://dx.doi.org/10.1016/j.patcog.2017.01.005
Received 14 July 2016; Received in revised form 23 November 2016; Accepted 4 January 2017
Available online 05 January 2017

that outperforms all previous methods by a large margin.

Overall, the main contributions of this work can be summarized in three-fold:

1. We have performed comprehensive diagnoses of the deep learning models for the age estimation problem by investigating three different kinds of formulations with five different loss functions to find that the regression based formulation with MAE loss is the best choice.
2. We have proposed a new architecture for simultaneously performing age estimation, gender and race classification which outperforms other deep multi-task learning architectures.
3. We have obtained a very deep age estimation model which significantly outperforms all previous solutions on two of the largest benchmark datasets.

We hope that these findings along with the whole diagnosing process facilitate the deployment of deep age estimation models for real-world applications.

2. Related work

Early works on age estimation are mainly focused on designing robust aging features and selecting learning algorithms. Some features are specifically designed for the age estimation problem, such as the facial features and wrinkles [6], the learned AGES (AGing pattErn Subspace) [7] features, as well as the biologically inspired features (BIF) [8]. General texture description features like the Local Binary Patterns (LBP) [2,9] and the Gabor feature [3] are also widely employed for age estimation. Given the aging features, classification models like linear SVM [8], Fuzzy LDA [3], Probabilistic Boosting Tree [10], or regression models like Support Vector Regression [8], Kernel Partial Least Squares [11], Neural Network [12] and Semidefinite Programming [13] are explored to estimate the ages.

Early studies also find that, by incorporating other kinds of facial traits like gender and race information, the performance of age estimation can be substantially improved [14,10,15,1,16]. In the experiments conducted by Guo [15], the age estimation error can be reduced by more than 20% if trained separately on male and female. Similar results are also reported from other previous works [10]. Therefore, joint analysis of these facial traits becomes a natural choice for obtaining better age estimation results.

Recently, the deep learning models have been consistently demonstrated as a very powerful framework for solving many computer vision problems, e.g., image classification [17–19], object detection [20–22], face verification [23,24], and facial attribute analyses [25,26]. The core philosophy within the deep learning framework is to let the network directly learn the feature representations and simultaneously train with the prediction tasks from end to end, which helps the deep learning models set new records for many vision tasks. Although with many successes, deep learning models are still mostly thought hard to implement and needs many tips and tricks [27]. To deploy the deep learning models to facial age estimation, although a very few studies have made some attempts [5,28,26], the performance gain obtained from these studies are not as significant as those obtained on other vision problems using deep learning models. Moreover, some of these studies focus on some other objectives, e.g., providing a benchmark dataset [5], or exploit complicated architectures, such as the 23 sub-networks multi-scale architecture in [28], the 36 local sub-networks tree-structured architecture in [26]. Therefore, we believe that the full potentials of deep learning models for the age estimation problem are still not fully explored, which motivates us to perform a comprehensive diagnosis of the deep age estimation model to dig out its most important parts.

Table 1
List of abbreviations used in the diagnosing process.

Abbreviation	Explanation
BF	Black Female
BM	Black Male
LDL	Label Distribution Learning
OH	One Hot
OR	Ordinal Regression
WF	White Female
WM	White Male

3. Diagnosing deep age estimation models

We now introduce our comprehensive diagnosing process of the deep learning based age estimation models. Model formulation and model architecture are two of the most important components for a deep age estimation problem. From model formulation perspective, we have investigated different kinds of formulations for the age estimation problem with the incorporations of commonly used loss functions for each kind of formulation. From the other model architecture perspective, we have studied three different model architectures that incorporate multi-task learning from race and gender classification for the age estimation problem. In the following, we first introduce the basic settings for the diagnosing process. Then we elaborate our diagnosing process on the model formulation and model architecture, respectively. In Table 1, we list all the abbreviations used in our diagnosing process for easy reference.

3.1. Diagnosing settings

To facilitate the diagnosing process, we provide here some basic settings for the diagnosing process, including the architecture of the baseline model, the selections of the benchmark datasets, and the designation of the evaluation metrics.

3.1.1. A baseline deep architecture

The baseline architecture in the following diagnosing is illustrated in Fig. 2, which has three convolutional layers and two fully-connected layers. The input is a 227×227 color image with the *mean image* subtracted. Specifically, the *mean image* is the mean of all the training images. For example, if there are N images in the training set, each image X_i is a $3 \times H \times W$ tensor, where H and W are the height and width of the image X_i. The *mean image* I_{mean} is then calculated as: $I_{mean} = 1/N \sum_{i=1}^{N} X_i$, where the summation and multiplication are element-wise operations. The first convolutional layer (Conv1) has 96 7×7 filters with a stride 4, followed by a 3×3 max pooling layer with a stride 2 and a local response normalization layer [17]. The second convolutional layer (Conv2) has 256 5×5 filters, followed by a 3×3 max pooling layer with a stride 2 and a local response normalization layer. The last convolutional layer (Conv3) has 384 3×3 filters. Again, followed by a 3×3 max pooling layer with a stride 2. The last two layers (FC1 and FC2) are two 512-D fully-connected layers. All the five layers are with the Rectified Liner Units (ReLU) [17]. This baseline architecture is an AlexNet [17] based architecture, which employs large size of convolution kernel and stride in the early layers and reduces their sizes gradually as the layer goes deeper. Our choice of this architecture as the baseline is initially motivated by the previous work [5] which employed this architecture to perform age group classification and demonstrated very good performance. Furthermore, as this baseline architecture is relatively small, it can greatly speedup the diagnosing process and save much time for both model training and testing. Based on these considerations, we therefore employ it as the baseline architecture for all the diagnosing process. It is worth noting that other modern CNN architectures such as VGGNet [19] and GoogLeNet [29] with smaller kernel size can be easily integrated into our final diagnosing

Table 2
The number of images in the three splits of Morph II dataset.

Gender	Race						
	Black			White			Others
Female	S1:1285	S2:1285	S3:3187	S1:1285	S2:1285	S3:31	S3:129
Male	S1:3980	S2:3980	S3:28843	S1:3980	S2:3980	S3:39	S3:1843

model. We have performed this kind of experiments in our final comparisons with the state-of-the-art methods in Section 4.6.

3.1.2. Age estimation datasets

There are many datasets for age estimation in the literature [30–32]. Most of these datasets, however, are relatively small. Since training a good deep neural network generally requires a large amount of training data, we therefore select two of the largest benchmark datasets to perform the diagnoses, the Morph II [33] dataset and the WebFace [34] dataset.

Morph II Dataset: Morph II contains about 55,000 face images of more than 13,000 subjects. Age ranges from 16 to 77 years old. Morph II is a multi-ethnic dataset with additional gender and race labels. It has about 77% Black faces and 19% White faces, while the remaining 4% includes Asian, Hispanic, Indian, and Other. Since Morph II is highly unbalanced in terms of race and gender distributions, in order to get a balanced training set, we follow the previous study [11] to split this dataset into three non-overlapped subsets S1, S2 and S3 (see Table 2). S1 and S2 are balanced in terms of race and gender distributions after this split and being used as training set separately. Specifically, in all experiments, the training and testing are repeated for twice: 1) training on S1, testing on S2+S3 and 2) training on S2, testing on S1+S3.

WebFace Dataset: The WebFace dataset contains 59,930 face images. Age ranges from 1 to 80 years old. The WebFace dataset is also a multi-ethnic dataset with additional gender labels. Unlike Morph II, this dataset is captured in the wild environment, images contain large pose and expression variations, which makes this dataset much more challenging. Following [34], we conduct experiments on this dataset using a four-fold cross validation protocol.

Some example face images in these two datasets are shown in Fig. 1. As we can see, both datasets are very challenging and thus can serve as very good benchmarks for evaluating the performance of age estimation algorithms. We also show the age distributions of the training set of both datasets in Fig. 3. As it can be observed from Fig. 3, the age distributions of these two datasets are imbalanced at some specific ages. For the Morph II dataset some classes are poorly represented, e.g., ages above 60 have few training samples. This is because the Morph II dataset is collected from people in a prison, most of whom are younger than 60 years old. For the WebFace dataset, the situation is much better than that in Morph II, and most of the classes have more than 300 samples. This imbalanced problem has an impact on the results of the deep age estimation models and we consider this problem in all of our model training process. To alleviate the potential bad effects from the imbalanced training samples, we employed an "age-aware sampling" strategy which tried to make the mini-batch during each training iteration as uniform as possible. With this strategy employed for all the diagnosed models in our paper, we believe the diagnosing results of all the models are referable. We will describe the "age-aware sampling" strategy in detail in Section 4.1.

3.1.3. Evaluation metrics

The most widely used evaluation metric for age estimation in the literature is the Mean Absolute Error (MAE), which is defined as follows,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{1}$$

where N is the number of testing samples, y_i is the ground truth age and $\hat{y}_i$ is the predicted age of the i-th sample. This MAE metric has been the standard evaluation metric for the age estimation problem since the very beginning of age estimation research [30]. This evaluation metric has been widely employed in the previous works [8,11,16,34,26,28], and in order to make a direct comparison with previous methods, we therefore employ this metric as one basic metric in this paper.

Although widely adopted, the MAE metric has several shortcomings. When evaluating a large testing set, improvements of some testing samples may not make a significant difference on the MAE value, especially when the MAE value is already very low. For most of the testing samples, the compared deep age estimation models can all produce correct predictions. The differences between different models are their ability to deal with a small number of difficult testing samples. These differences cannot be well reflected by MAE metric. Another shortcoming of MAE is that it cannot reflect the distributions of the estimation errors. To overcome these shortcomings, we design two other metrics to evaluate and compare the performance of different age estimation methods.

The first metric is *Cumulative Correct Score* (CCS) that is used to compare multiple methods. CCS is defined as **the number of** test images such that the absolute age estimation error is not higher than a year threshold t, i.e.,

$$\mathrm{CCS}(t) = \sum_{i=1}^{N} h(|\hat{y}_i - y_i| - t), \tag{2}$$

$$h(x) = \begin{cases} 1, & if\ x \le 0 \\ 0, & otherwise \end{cases}. \tag{3}$$

In practice, people are more concerned about age estimator's performance at different thresholds than a single MAE number. So we can study CCS at different thresholds to clearly locate the difference of performance between multiple methods. For example, if the CCS of one method at different thresholds are consistently larger than other methods, we can conclude that this method performs better than others.

The second metric is the *Relative Cumulative Correct Score* (RCCS) which is specifically used to compare two methods. RCCS is defined as,

$$\mathrm{RCCS}_b^a(t) = \mathrm{CCS}^a(t) - \mathrm{CCS}^b(t), \tag{4}$$

where CCS^a is the CCS of method a, CCS^b is the CCS of method b. From RCCS we can clearly see which method is better at different thresholds. Both of the above two metrics can reflect the error distribution of the

Fig. 1. Some example face images in the two benchmark datasets used in this paper. The images in the top row are from the Morph II dataset, and the images in the bottom row are from the WebFace dataset. As we can see, age estimation is a challenging task in computer vision. Even for human beings ourselves, it is very difficult to tell the accurate age from a face image.

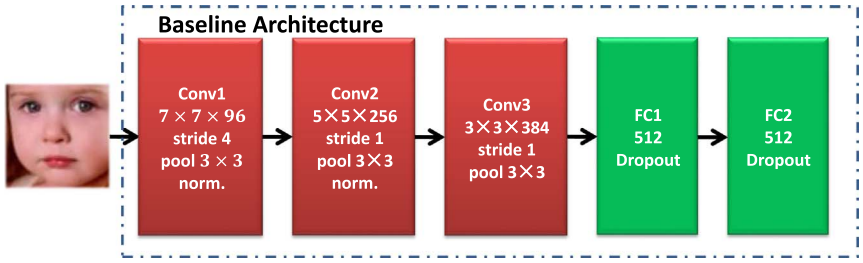


Fig. 2. The baseline architecture employed in our diagnosing process. This architecture is mainly motivated from the AlexNet architecture [17], which is one of the most famous modern deep learning architecture, and the one used in the Adience benchmark [5], which demonstrate good performance for age group classification firstly using a deep architecture. This relatively simple and shallow architecture also greatly speeds up the diagnosing process.

method and thus are more intuitive and expressive than a single MAE number.

3.2. Model formulation

Age estimation can be formulated as a multi-class classification problem, a regression problem, or a ordinal regression problem. In this part, we intend to diagnose the effects of these three different types of formulations in the deep learning framework, which has not been explicitly studied by previous work to the best of our knowledge.

3.2.1. Classification formulation for age estimation

Since deep learning models have demonstrated superior performance for many classification problems [17,29], age estimation can be directly formulated as a classification problem by discretizing age in possible year ranges, e.g., between 0 and 80 years in the WebFace dataset. To deploy this formulation into the CNN model, we need to employ a softmax activation function in the output layer and append a classification loss to it.

One-hot encoding based method. The most widely used classification based age estimation method is to use one-hot encoding to represent the age label, and the Softmax loss is adopted as the training objective. Denote the i-th image sample as $\mathbf{X}_i$, the sample's age label as y_i, and the training dataset as $\mathcal{X} = \{\mathbf{X}_i, y_i\}_{i=1}^{N}$, $y_i \in \{1, \ldots C\}$, where C is the number of different age labels. The Softmax loss is defined as follows:

$$\mathcal{L}(\mathcal{X}) = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_{iy_i}\log\mathbf{p}_{iy_i}, \tag{5}$$

where $\mathbf{p}_i$ is the predicted C-dimensional probability vector for $\mathbf{X}_i$ by the CNN, $\mathbf{p}_{iy_i}$ is the y_i-th element of $\mathbf{p}_i$, and $\mathbf{y}_i$ is the one-hot encoding of y_i, i.e., $\mathbf{y}_{iy_i} = 1$ and 0 otherwise.

Label distribution based method. It is noted that in the one-hot encoding based methods, the ages are treated as independent from each other. To consider the correlations between face images with different ages under the classification framework, [35] proposed a label distribution learning method for age estimation. The general idea is extending the one-hot encoding of one face image to a label distribution. In this work, we use the Gaussian label distribution. For the i-th face image $\mathbf{X}_i$ with age label y_i, then the k-th dimension of the corresponding target label distribution is defined as follows:

$$\mathbf{d}_{ik} = \frac{1}{\sigma\sqrt{2\pi}Z}\exp\left(-\frac{(k-y_i)^2}{2\sigma^2}\right),\ k = 1, \ldots, C, \tag{6}$$

where σ is the standard deviation of the Gaussian distribution, and Z is a normalization factor that makes sure $\sum_k \mathbf{d}_{ik} = 1$, i.e.,

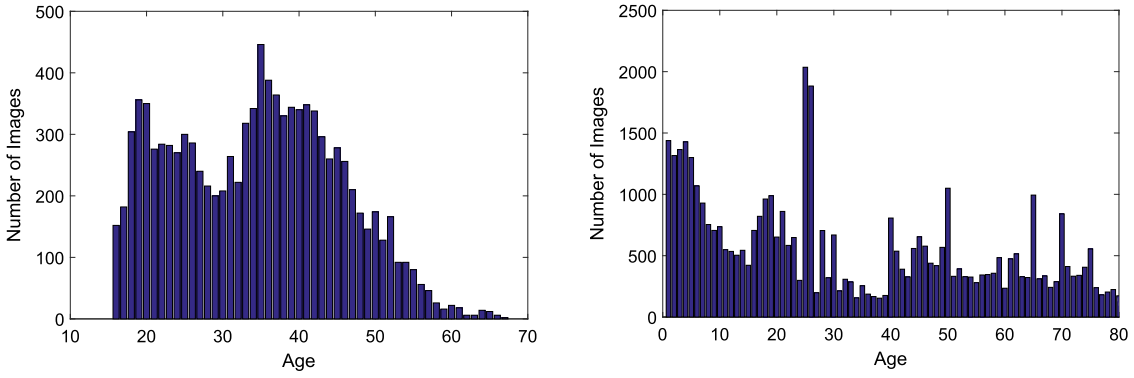
$$Z = \frac{1}{\sigma\sqrt{2\pi}}\sum_{k}\exp\left(-\frac{(k-y_i)^2}{2\sigma^2}\right). \tag{7}$$

The training objective is defined as follows based on the cross-entropy loss:

$$\mathcal{L}(\mathcal{X}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\mathbf{d}_{ik}\log\mathbf{p}_{ik}, \tag{8}$$

3.2.2. Regression formulation for age estimation

The age of an individual is measured by the time passed from the individual's birth, and thus is a continuous value. Therefore, the age estimation problem can also be naturally formulated as a regression problem where the objective becomes to find a regression function that can model the aging process in terms of the feature space. This kind of formulation provides a more natural and accurate formulation than the classification formulation and has fewer parameters in the output layer



(a) Age distribution of the training set of Morph II dataset.

(b) Age distribution of the training set of WebFace dataset.

Fig. 3. Age distributions of the training set of both datasets used in this work.

(i.e., one neuron vs. C neurons). To deploy this formulation into the CNN model, we only need to employ a linear activation function in the output layer and append a regression loss to it.

MSE based method. When formulating age estimation as a regression problem, the Mean Squared Error (MSE) loss can be used as the training objective, which is defined as follows:

$$\mathcal{L}(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{9}$$

where $\hat{y}_i$ is the prediction and y_i is the ground truth age. This MSE loss has very nice mathematical properties like convexity and being continuously differentiable, which makes it widely used in regression based age estimation and many other regression problems.

MAE based method. Besides the MSE loss, we find that the Mean Absolute Error (MAE) can potentially provide a better loss for a deep age estimation model, which is defined as follows:

$$\mathcal{L}(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{10}$$

The main inspiration behind is that the performance of an age estimation model is evaluated using the MAE metric (Eq. (1)). Using this evaluation metric as the loss function for the deep model provides a more straightforward objective for end-to-end model training. To our knowledge, MAE is not directly used as the loss function in previous work for age estimation. The main reason for this may be that the MAE loss is not a smooth function, which makes it hard to optimize for traditional age estimation methods. With recent developments of optimizing non-smoothing functions like ReLU [17] and PReLU [18] in the deep learning framework, the MAE loss function can be optimized effectively using the stochastic gradient descent algorithm.

3.2.3. Ordinal regression formulation for age estimation

The human face matures in different ways depending on the person's age. For example, facial aging effects appear as changes in the shape of the face during childhood and changes in skin texture during adulthood. Based on this observation, previous works [36,37] argue that it is difficult for regression based methods to handle this aging pattern non-stationary problem, and suggest using ordinal regression based methods for age estimation. In this work, we also diagnose this ordinal regression formulation for age estimation using CNN.

Ordinal regression can be considered an intermediate problem in between regression and classification. The most successful and widely used algorithm to solve the ordinal regression problem is to transform it into a series of simpler binary classification subproblems [38,39]. For each age $k \in \{1, 2, \ldots, C-1\}$, a binary classifier is trained according to whether the age of a face is larger than k. Then the age of a test face is predicted based on the classification results of these $C-1$ binary classifiers. Specifically, given the original training face images $\mathcal{X} = \{\mathbf{X}_i, y_i\}_{i=1}^{N}$, for the k-th binary classification subproblem a specific training data is constructed as $\mathcal{X}^k = \{\mathbf{X}_i, y_i^k\}_{i=1}^{N}$, where $y_i^k \in \{0, 1\}$ is the label indicating whether the age of the i-th face image is larger than k. Training on this $\mathcal{X}^k$ one can obtain the k-th binary classifier f_k. After obtaining all these $C-1$ binary classifiers, for a given test face image $\mathbf{X}_j$, the predicted age $\hat{y}_j$ is calculated as follows:

$$\hat{y}_j = 1 + \sum_{k=1}^{C-1} f_k(\mathbf{X}_j), \tag{11}$$

where $f_k(\mathbf{X}_j)$ is the binary classification result for $\mathbf{X}_j$ by f_k.

One may notice that this approach has the problem of unbalanced classes. For example, for the age $C-1$, almost all examples will be of the class less than $C-1$ years. We tackle this problem from two perspectives. First, from the network structure, we adopt one network to collectively implement all these $C-1$ binary classifiers in our experiments. In particular, our network has a multiple-output structure where each output corresponds to a binary classifier. Thus, these $C-1$ classification sub-problems (tasks) are simultaneously trained in an end-to-end manner. So, these tasks can learn and benefit from each other. For example, the tasks with balanced classes may help the learning of tasks faced with unbalanced problems. Second, from the training process, we try to make the mini-batch as uniform as possible (same amount of positive and negative samples) for each task during network training, which can alleviate the unbalanced problem to some extent.

3.3. Model architecture

Besides the model formulation, the model architecture is also very important for the deep age estimation problem. For the model architecture, we mainly study the different ways to design the architecture to incorporate multi-task learning for more effective age estimation. Since age, gender, and race are three closely related facial traits of a human, early studies on age estimation suggest that incorporating these three different kinds of traits together can improve the results of age estimation [10,16]. Moreover, it will be very beneficial to use a single network in a practical system which can save much computation and memory. These facts naturally motivate us to investigate multi-task learning of gender, race, and age to obtain additional performance gain. To perform multi-task learning of age, gender, and race upon the baseline architecture, we will introduce three different multi-task learning architectures in the following.

3.3.1. Parallel multi-task learning architecture

The parallel multi-task learning architecture fuses the different tasks concurrently and has been widely adopted in previous deep learning models [24,40,25,22]. This type of architecture has show very good performances in problems like face recognition [24,40] and object detection [21,22].

To deploy this multi-task learning strategy to the baseline architecture, we attach three fully-connected layers to the last layer (i.e., FC2) of the baseline architecture (Fig. 4(a)). Each fully-connected layer is then followed by a loss function designed for the task. The weights $W_1, \ldots, W_5$ associated with the first five layers are shared by all the three tasks. W_a, W_g, and W_r are the task specific parameters. Denote the training dataset as $\mathcal{X} = \{\mathbf{X}_i, y_i^a, y_i^r, y_i^g\}_{i=1}^{N}$, where $y^a{}_i$, $y^r{}_i$, and $y^g{}_i$ are the labels for age, race and gender, respectively. The loss for the parallel multi-task learning model is defined as follows:

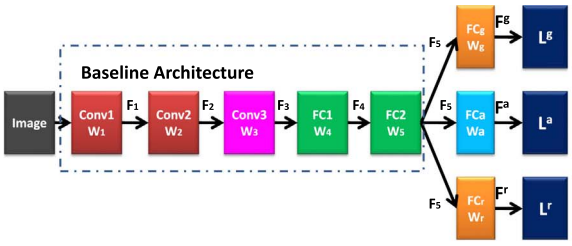
$$\mathcal{L}(\mathcal{X}) = L^a + \alpha L^g + \beta L^r = \frac{1}{N}\sum_{i=1}^{N}|\mathrm{F}_i^a - y_i^a| - \alpha\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{2}\mathbf{y}_{ik}^g \log \mathbf{F}_{ik}^g - \beta\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C_r}\mathbf{y}_{ik}^r \log \mathbf{F}_{ik}^r, \tag{12}$$

where α and β are hyper-parameters to tune the importance of each task, F_i^a is the predicted age for the i-th sample, $\mathbf{F}_i^g$ is the predicted gender probability vector of the i-th sample, $\mathbf{F}_{ik}^g$ is the k-th element of $\mathbf{F}_i^g$, and $\mathbf{y}_{ik}^g = 1$ if the i-th sample has gender label k and 0 otherwise. The meanings of $\mathbf{F}_i^r$, $\mathbf{F}_{ik}^r$, $\mathbf{y}_{ik}^r$ are similar for race, and C_r is number of race labels. Note that here we use the MAE loss for the regression based age estimation task (F_i^a is a scalar), and Softmax loss for the race and gender classification tasks.

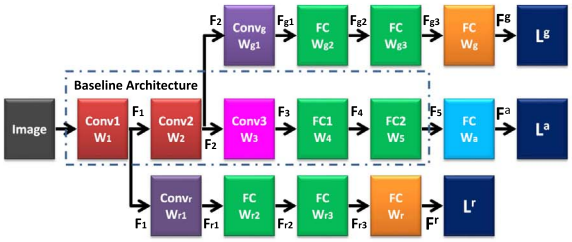
3.3.2. Deeply supervised multi-task learning architecture

The deeply supervised multi-task learning architecture fuses different tasks progressively in different layers, which is inspired by previous work like deeply supervised nets [41] and many other variants like GoogLeNet [29] and DeepID2 [24]. This kind of architecture has also demonstrated performance improvement on tasks like image classification.

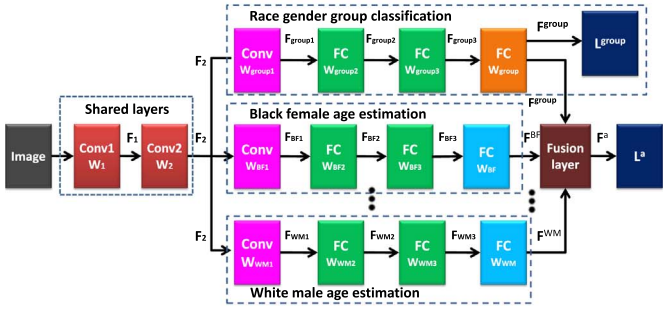
To deploy this multi-task learning strategy to the baseline archi-



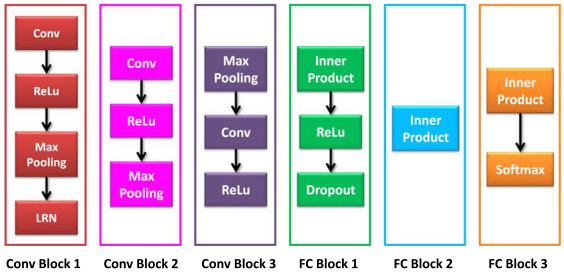
(a) Parallel multi-task learning architecture.



(b) Deeply supervised multi-task learning architecture.



(c) Hybrid multi-task learning architecture.



(d) Internal structure of each computational block.

Fig. 4. Illustration of our three multi-task learning architectures. F denote the intermediate layer outputs and W are the weights for each computational block. Computational blocks of the same type are shown in the same color. Fig. 4(d) shows the internal structure of each kind of block. For example, the Conv1 and FC1 in Fig. 4(a) are instantiations of "Conv Block 1" and "FC Block 1" in Fig. 4(d), respectively. We use a, g, and r to denote age, gender and race. Best viewed in color.

tecture, we add supervision branches after certain intermediate layers for gender and race classification tasks. Age estimation is done at last layer, which is similar to the parallel architecture. Please see Fig. 4(b) for more details. The intuition behind the deeply supervised multi-task learning architecture is that the three tasks (age, gender, race) are not of the same difficulty. Age estimation is more difficult than the other two tasks, and it requires more layers of abstractions with large capacity. Therefore, the loss function for the age estimation is connected to the highest-level features. The race classification task on Morph II only involves black and white people, which is relatively easy to distinguish based on only low-level features (e.g., color and texture), it is thus connected to the first convolutional layer. Gender classification is slightly more difficult than race classification, so it requires slightly higher-level features and is connected to the second convolutional layer. Since feature maps at lower convolutional layers may be noisy and not discriminative enough, following [41], we add a dimensionality reduction layer (e.g., Conv_g and Conv_r layers with kernel size 1×1 in Fig. 4(b)) and two discriminative non-linear mapping layers before the final classification. The overall loss is the same as the parallel architecture discussed above, and the network is trained using stochastic gradient descent algorithm.

3.3.3. Hybrid multi-task learning architecture

Although the above two architectures use gender and race information, they do not consider the relationship between tasks. Previous studies on age estimation [15] suggest that age estimation can be influenced by the gender and race differences dramatically, i.e., age estimation errors can be increased when across gender and/or race.

Inspired by these findings, we further propose a hybrid multi-task learning architecture for age estimation by explicitly incorporating this prior knowledge into the network architecture. See Fig. 4(c) for more details. This architecture mainly comprises of four parts: 1) shared part (i.e., Conv1 and Conv2), 2) race gender group (i.e., BF, BM, WF, WM) classification part, 3) group specific age estimation part, and 4) a fusion layer that fuses the prediction made by each group specific age estimator.

$\mathbf{F}^{\text{group}}$ is the output of group classification part indicating the probability of an input belonging to each race gender group. Each of the group specific age estimator part excels in estimating the age of images belongs to one specific race gender group. After we obtain the group probabilities $\mathbf{F}^{\text{group}}$ and all high accurate group specific age predictions (i.e., F^{BF}, …, F^{WM}) for an input, we employ the average fusion strategy to get the final prediction F^a as follows:

$$\text{F}^a = \sum_{k\in\{\text{BF,BM,WF,WM}\}} \mathbf{F}_k^{\text{group}} \text{F}^k, \tag{13}$$

where $\mathbf{F}_k^{\text{group}}$ is the element of $\mathbf{F}^{\text{group}}$ corresponding to race gender group k.

To train the hybrid multi-task age estimation model, we design a three-step procedure: 1) Pre-training race gender group classification part using all the training data; 2) Pre-training group specific age estimation part using group specific training data; and 3) Fine-tuning the whole network from end to end using all the training data. The testing procedure is simple, we can use Eq. (13) to obtain the final age estimations. We can also obtain the race and gender predictions from the group classification part easily.

4. Results and analyses

In this section, we first describe some details about our experimental settings. Then, we give the experimental results on different model formulations and different multi-task architectures. We will also analyze the model depth for age estimation. Finally, we compare our best model with the state-of-the-art age estimation methods.

4.1. Experimental settings

The face images in the dataset were preprocessed in a standard way, i.e., the faces in the images are detected and aligned, then cropped and normalized to 256×256. For all the following experiments, we use the Caffe [42] toolbox, which provides a flexible framework to develop new deep learning models, and makes our work easy to reproduce. All the model protocol files and training results in our experiments will be released in the Caffe model zoo. We train all the network using mini-batch (set to 128) stochastic gradient descent with momentum (0.9) and weight decay (5×10^{-4}). For all fully-connected layers we use a

dropout ratio of 0.5. We use data augmentation similar to [17], i.e., randomly cropping of 227×227 pixels from the 256×256 input face images, then randomly mirroring it before feeding it to the network. The initial learning rate is 10^{-3}. We divide the learning rate by 10 every 10,000 iterations, and training stops at 50,000 iterations. These hyper-parameters are chosen based on a hold-out validation set. We found that all the networks converge well under these settings, so we use the same parameters for different models to make fair comparisons between different methods.

As shown in Fig. 3, the datasets used in this paper are imbalanced at some specific ages. To alleviate its impact on the results of deep age estimation models, we apply an "age-aware sampling" strategy during training. In practice, we use two types of lists, one is age list, and the other is per-age face image list. At each iteration during training, we first sample an age M in the age list, then sample a face image in the per-age face image list of age M, and repeat this process multiple times to create a mini-bath. When reaching the end of the per-age face image list of age M, a random shuffle operation is performed to reorder the face images of age M. When reaching the end of age list, we also perform a random shuffle operation to reorder the ages. With this sampling strategy for the mini-batch based training process, we can get a mini-batch as uniform as possible with respect to ages, and thus alleviate the imbalanced problem to some extent.

4.2. Analyses of classification based deep age estimation methods

To study the classification formulation for age estimation using CNN, we train two models Net_{OH} and Net_{LDL} from the baseline architecture (Fig. 2). Net_{OH} is based on the One-Hot encoding and Net_{LDL} is based on the Label Distribution Learning introduced in Section 3.2.1. The age estimation results of these two models on both datasets are shown in Tables 3, 4.

We can clearly see that Net_{LDL} outperforms Net_{OH} on both datasets. This is because in Net_{OH}, the age labels are assumed to be independent to one another. However, Net_{LDL} considers the correlations between different ages. These results show that it is better to use label distribution learning for classification based deep age estimation.

4.3. Analyses of regression based deep age estimation methods

To study the regression formulation for age estimation using CNN, we train two models Net_{MSE} and Net_{MAE} from the baseline architecture. Net_{MSE} is based on the Mean Squared Error and Net_{MAE} is based on the Mean Absolute Error introduced in Section 3.2.2. The age estimation results of these two models on Morph II and WebFace datasets are shown in Tables 3, 4.

We can see that Net_{MAE} is better than Net_{MSE} on both datasets. There are two reasons to explain these results. First, the MAE loss is more robust to outliers than the MSE loss, which is very important in practice, because label noises are inevitable in real-world datasets. For example, the WebFace dataset contains many more label errors than the Morph II dataset, so the performance gap between Net_{MAE} and Net_{MSE} is larger on WebFace dataset than on the Morph II dataset.

Table 3
The age estimation results of the three different formulations on Morph II dataset using the training and testing set split protocol in Table 2.

Method		s2+s3 MAE	s1+s3 MAE	Average MAE
Classification based methods	Net_{OH}	3.85	3.89	3.87
	Net_{LDL}	3.48	3.49	3.49
Regression based methods	Net_{MSE}	3.44	3.43	3.44
	Net_{MAE}	**3.40**	**3.39**	**3.40**
Ordinal regression based method	Net_{OR}	3.46	3.48	3.47

Table 4
The age estimation results of the three different formulations on WebFace dataset using the four-fold cross validation protocol.

Method		Fold1 MAE	Fold2 MAE	Fold3 MAE	Fold4 MAE	Average MAE
Classification based methods	Net_{OH}	6.65	6.76	6.63	6.64	6.67
	Net_{LDL}	6.20	6.24	6.10	6.26	6.20
Regression based methods	Net_{MSE}	6.40	6.46	6.30	6.44	6.40
	Net_{MAE}	**6.12**	**6.13**	**5.99**	**6.22**	**6.12**
Ordinal regression based method	Net_{OR}	6.19	6.28	6.20	6.30	6.24

Second, MAE is the evaluation metric for age estimation algorithm (see Eq. (1)), so directly optimize this MAE loss can improve the age estimation performance. For example, even though the Morph II dataset was complied in a controlled environment and has few label errors, Net_{MAE} which directly optimizes the final evaluation metric still performs better than Net_{MSE} on this dataset.

Since the widely adopted Morph II benchmark is collected in controlled environment, the performance on it is already very high. For most of the testing samples, the compared deep age estimation models can all produce correct predictions, and improvements of some hard testing samples may not make a significant difference on the MAE value. In order to better show the gain in performance of Net_{MAE} over Net_{MSE}, we use our RCCS metric (Eq. (4)) to compare Net_{MAE} and Net_{MSE}. The results are show in Fig. 5. We can see that Net_{MAE} can make more correct predictions than Net_{MSE} at all the thresholds (from 0 to 6 years) since all the numbers in Fig. 5 are positive numbers.

Thanks to this MAE loss function which is not only robust to outliers but also can be used directly to optimize the evaluation metric, this regression based methods Net_{MAE} outperforms both classification based methods Net_{OH} and Net_{LDL} on both datasets. This also validates the philosophy of deep learning, i.e., direct optimization of what you want can always improve the performance.

4.4. Analyses of ordinal regression based deep age estimation methods

We use Net_{OR} to denote the Ordinal Regression based deep age estimation model introduced in Section 3.2.3. From the age estimation results in Tables 3, 4, we can see that Net_{OR} performs better than Net_{OH} and is comparable to Net_{LDL}. Surprisingly, the regression base methods Net_{MAE} still outperforms Net_{OR} on both datasets. We also plot the age estimation performances at different testing ages on both datasets for Net_{LDL}, Net_{MAE} and Net_{OR} which are representative methods of classi-

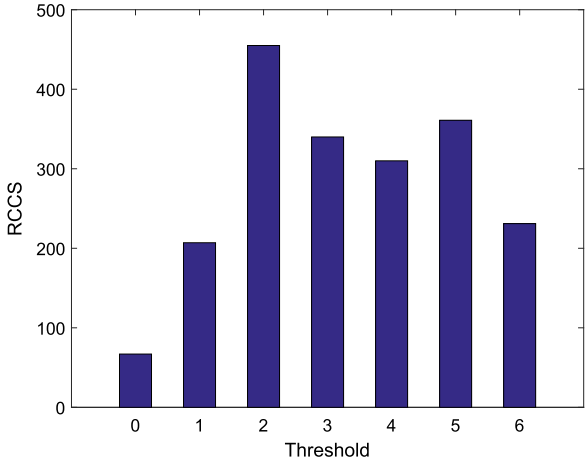


Fig. 5. The RCCS between Net_{MAE} and Net_{MSE} on Morph II dataset.

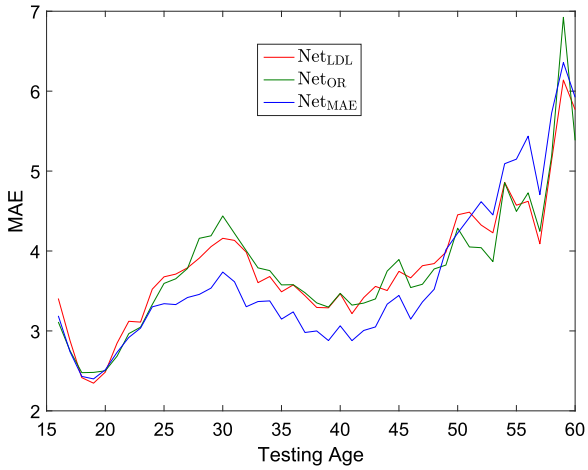


Fig. 6. Age estimation performances of Net_{LDL}, Net_{OR} and Net_{MAE} at different testing ages on Morph II dataset.

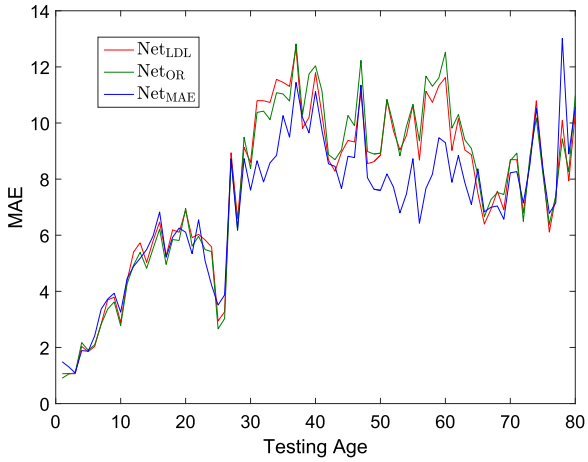
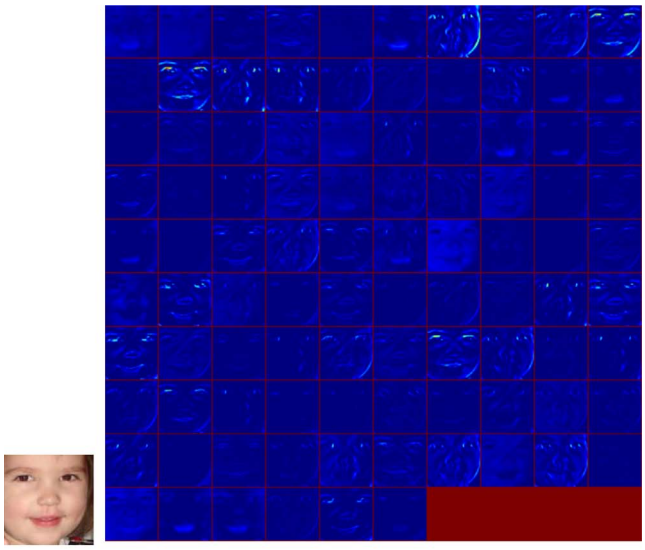


Fig. 7. Age estimation performances of Net_{LDL}, Net_{OR} and Net_{MAE} at different testing ages on WebFace dataset.

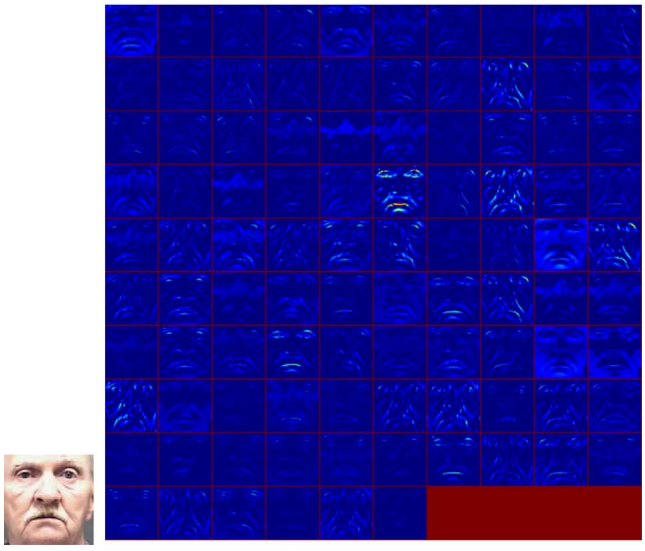
fication, regression, and ordinal regression based methods, respectively. The results are show in Figs. 6 and 7. We can see that the difficulties of age estimation from different ages are not the same. The MAE of the young people are smaller than that of the adult people. The reason for this can be explained as follows. From birth to adulthood, the greatest change of the face is the craniofacial growth, and during adult aging, the most perceptible change becomes texture change which is subtler than the craniofacial growth. Therefore, it is more difficult to estimate the age for the adult people than for the young people, and this fact is also true for our human beings. We can also see that Net_{MAE} performs better than Net_{LDL} and Net_{OR} at most of the different testing ages.

As mentioned in Section 3.2.3, previous works argue that it is difficult for regression based age estimation methods to handle the aging pattern non-stationary problem and it is better to use ordinal regression based methods. In this work, we show that this claim may not be true under the deep learning framework. Fig. 8 shows the output responses of the first convolutional layer of the regression based deep age estimation model Net_{MAE} on two face images. We can see that for the young face (Fig. 8(a)), most of the features extracted by Net_{MAE} contain shape information (see the facial contours in Fig. 8(b)). While for the old face (Fig. 8(c)), the Net_{MAE} model can extract rich texture features (see the wrinkles in Fig. 8(d)). These intuitive visualizations demonstrate that the regression based deep model Net_{MAE} can learn this non-stationary aging process automatically and effectively.

Based on the above analysis of the three formulations for age



(a) (b) Responses for Face(a).



(c) (d) Responses for Face(c)

Fig. 8. Visualizing output of the first convolutional layer of Net_{MAE}. For the input faces (a) and (c), (b) and (d) show the output responses of the first convolutional layer of Net_{MAE} for these two faces respectively. The first convolutional layer of Net_{MAE} contains 96 feature maps, whose outputs are shown here on a 10×10 grid.

estimation using CNN, we can see that the regression based deep model Net_{MAE} is the most promising one despite of its simplicity. Compared with Net_{OH} which assumes that each age is independent from other ages, Net_{MAE} provides a natural formulation which takes account of the continuous nature of age. Compared with Net_{MSE}, Net_{MAE} is more robust to outliers which is very important in read-world datasets. Compared with Net_{LDL} and Net_{OR}, Net_{MAE} can not only directly optimize the final evaluation metric but also can automatically capture the non-stationary aging process and thus obtains competitive results.

4.5. Analyses of multi-task learning with race and gender classification

Since the regression formulation with MAE loss function works best in our experiments, we use it by default in this set of experiments. Note that our Net_{MAE} has already beaten other state-of-the-art methods on both datasets. We are curious about whether using multi-task learning

Table 5
The results of our three multi-task learning architectures on Morph II dataset.

Method	Training Set	Testing Set	CCS(0)	CCS(1)	CCS(2)	CCS(3)	CCS(4)	MAE (yrs.)	Average MAE (yrs.)
Net_{MAE}	S1	S2+S3	4642	13676	21316	27400	32194	3.40	3.40
	S2	S1+S3	4734	13741	21564	27621	32341	3.39	
Parallel architecture $Net_{Parallel}$	S1	S2+S3	4627	13575	21547	27499	32154	3.42	3.43
	S2	S1+S3	4682	13557	21478	27397	32185	3.43	
Deeply supervised architecture Net_{Deeply}	S1	S2+S3	4684	13649	21363	27435	32036	3.40	3.42
	S2	S1+S3	4751	13707	21196	27264	32088	3.43	
Hybrid architecture Net_{Hybrid}	S1	S2+S3	**4720**	**14013**	**21758**	**27924**	**32738**	**3.32**	**3.31**
	S2	S1+S3	**4814**	**13973**	**21825**	**28080**	**32871**	**3.30**	

Table 6
The results of our three multi-task learning architectures on WebFace dataset.

Method	Testing Fold	CCS(0)	CCS(1)	CCS(2)	CCS(3)	CCS(4)	MAE (yrs.)	Average MAE (yrs.)
Net_{MAE}	Fold1	1372	4017	5942	7296	8403	6.12	6.12
	Fold2	1427	4089	5990	7345	8466	6.13	
	Fold3	1442	4137	6117	7534	8576	5.99	
	Fold4	1345	4067	5995	7323	8412	6.22	
Parallel architecture $Net_{Parallel}$	Fold1	1322	3932	5904	7314	8419	6.16	6.15
	Fold2	1347	4046	5941	7342	8437	6.17	
	Fold3	1405	4072	6015	7482	8602	6.05	
	Fold4	1355	3954	5892	7320	8441	6.23	
Deeply supervised architecture Net_{Deeply}	Fold1	1350	3953	5965	7343	8485	6.13	6.15
	Fold2	1346	3975	5870	7227	8342	6.20	
	Fold3	1400	4042	6098	7512	8588	6.05	
	Fold4	1351	3955	5859	7291	8446	6.22	
Hybrid architecture Net_{Hybrid}	Fold1	**1813**	**4504**	**6399**	**7695**	**8667**	**6.03**	**6.03**
	Fold2	**1790**	**4438**	**6312**	**7674**	**8701**	**6.08**	
	Fold3	**1777**	**4453**	**6359**	**7785**	**8871**	**5.91**	
	Fold4	**1751**	**4290**	**6212**	**7589**	**8660**	**6.08**	

with race and/or gender information can further improve the performance of age estimation. Tables 5, 6 show the results of our three multi-task learning architectures introduced in Section 3.3 on both datasets.

From the MAE evaluation metric, we can see that the parallel architecture $Net_{Parallel}$ is comparable to or slightly worse than the single task model Net_{MAE}, which is also reported in [28]. Deeply supervised architecture Net_{Deeply} is slightly better than $Net_{Parallel}$, but again shows no improvement over the single task model Net_{MAE}. These results suggest that simply inserting multiple loss functions to the network and forcing it to learn from each other dose not work well on age estimation. Our proposed hybrid architecture Net_{Hybrid}, on the contrary, works better than the other two architectures and shows improvements over the single task model Net_{MAE}.

From our new CCS metric (Eq. (2)), we can also see that the Net_{Hybrid} obtains the largest CCS at different thresholds (from 0 to 4 years) on both datasets. These results demonstrate that our hybrid architecture can make more accurate predictions which is very important for practical use. The main reason behind this is that our hybrid architecture considers the relationship between tasks, and encodes this information directly into the network design.

4.6. Analyses of the model depth for age estimation

Very deep CNNs, such as VGGNet [19] and GoogLeNet [29] have achieved great success for many computer vision tasks. Since our baseline architecture in Fig. 2 is relatively shallow compared to these very deep architectures, we are curious about whether using these very deep architectures can further improve the performance of age estimation. Based on this consideration, we train another mode Net^{VGG}_{Hybrid} which is a hybrid multi-task architecture based on the very deep VGGNet. Compare to our baseline architecture which is shallow (3 convolutional layers) with large kernel size (7×7), the VGGNet is deeper (16 layers) with smaller kernel size (3×3). Tables 7, 8 show the age estimation results of Net^{VGG}_{Hybrid} on both dataset. We can see that Net^{VGG}_{Hybrid} based on the very deep VGGNet performs much better than Net_{Hybrid} which is based on the shallow baseline architecture. These results demonstrate that using deeper model with smaller convolutional kernel size can further improve the age estimation performance.

Table 7
The age estimation results of Net_{Hybrid} and Net^{VGG}_{Hybrid} on Morph II dataset using the training and testing set split protocol in Table 2.

Method	S2+S3 MAE	S1+S3 MAE	Average MAE
Net_{Hybrid}	3.32	3.30	3.31
Net^{VGG}_{Hybrid}	**2.96**	**2.95**	**2.96**

Table 8
The age estimation results of Net_{Hybrid} and Net^{VGG}_{Hybrid} on WebFace dataset using the four-fold cross validation protocol.

Method	Fold1 MAE	Fold2 MAE	Fold3 MAE	Fold4 MAE	Average MAE
Net_{Hybrid}	6.03	6.08	5.91	6.08	6.03
Net^{VGG}_{Hybrid}	**5.72**	**5.78**	**5.70**	**5.80**	**5.75**

4.7. Analyses of age estimation accuracy under different races and genders

We use Morph II dataset to analysis the age estimation accuracy for people of different races and genders, since this dataset has both race and gender labels. The results are show in Fig. 9. We can see that the MAE of white people is smaller than that of black people. The main reason behind this is that it may be easier to detect the facial appearance changes of white people than those of black people. We can also see that the MAE of male is smaller than that of female, this is because males and females may have different face aging patterns.

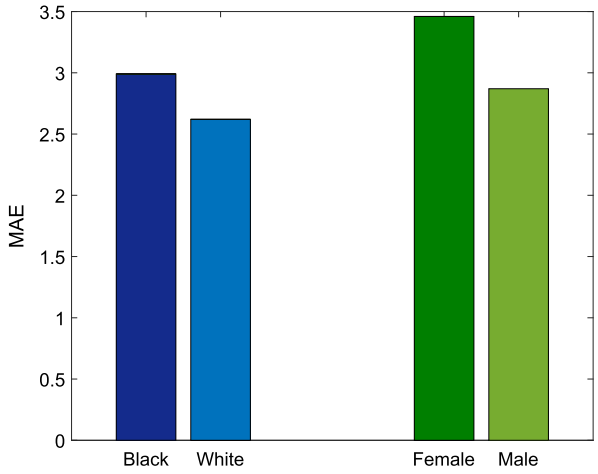


Fig. 9. The age estimation results of $\text{Net}^{\text{VGG}}_{\text{Hybrid}}$ for different races (Black vs. White) and genders (Female vs. Male) on Morph II dataset.

Table 9
Comparison with the state-of-the-art methods on Morph II dataset.

Methods	Gender Accuracy	Race Accuracy	Age MAE
BIF [8]	96.6	–	5.09
KPLS [11]	98.4	98.9	4.18
KCCA [16]	98.5	99.0	3.98
Ridge [34]	97.7	–	4.80
Tree-a-CNN [26]	98.4	–	3.61
Multi-scale-CNN [28]	98.0	98.6	3.63
$\text{Net}^{\text{VGG}}_{\text{Hybrid}}$	**98.7**	**99.2**	**2.96**

Table 10
Comparison with the state-of-the-art methods on WebFace dataset.

Methods	Gender Accuracy	Age MAE
BIF [8]	79.3	10.65
RF [43]	–	9.38
Ridge [34]	87.0	9.75
Tree-a-CNN [26]	89.7	7.72
$\text{Net}^{\text{VGG}}_{\text{Hybrid}}$	**92.3**	**5.75**

Many female faces may potentially show younger appearances than male face due to the different extent in using makeups and accessories [1], and this fact makes it more difficult to estimate the age of females.

4.8. Comparison with the state-of-the-art methods

Tables 9, 10 compare our best model (i.e., $\text{Net}^{\text{VGG}}_{\text{Hybrid}}$) with several recently published methods on Morph II and WebFace datasets. All of the methods evaluated in this section are using the same training and testing set partition protocol which is discussed in Section 3.1.2 for fair comparisons. Our results outperform all the other state-of-the-art methods on both datasets by a large margin. On the Morph II dataset, our best model reduces MAE by 0.65 years which is a significant improvement. To the best of our knowledge, this is for the first time an MAE below 3 years has been obtained on the Morph II age estimation dataset.

On the WebFace dataset, our model improves the best results by about 2 years. Since the WebFace dataset is built from faces in the wild, few methods conducted experiments on this challenging dataset. We have compared our model to all other published results we can find on this dataset, including the latest one in [26]. Our 1.97 years reduction of MAE is a significant improvement over the state-of-the-art methods considering the difficulty of this dataset. This competing performance of our model indicates the effectiveness of our diagnoses in model formulation, loss function, multi-task learning architecture, and model depth for age estimation.

5. Conclusion and future work

In this paper, we investigate the deep learning based age estimation problem. We have performed in-depth diagnoses of the deep learning models for the age estimation problem. Started from a simple baseline architecture, we have progressively improved its performance by investigating three different kinds of formulations of the model using five different loss functions, as well as the model architecture design for multi-task learning. By accumulating all these findings, we finally obtain a very deep age estimation model with high prediction accuracy. We hope these findings and results to be useful for the research and application of the deep age estimation techniques.

In our future work, we plan to study the age estimation problem regarding to a specific age group or a specific person, using the deep learning models. For the age-specific age estimation problem, we need to find a principled way to learn age-dependent optimization objectives for the deep age estimation model. For the person-specific age estimation problem, we plan to design some transfer learning based mechanism to adapt the knowledge learned from a general deep age estimation model to a dedicated deep age estimation model for a specific person.

Conflict of interest

None declared.

Acknowledgments

This work is partly supported by the 973 Basic Research Program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421, U1636218, 61672519, and 61303178), the Strategic Priority Research Program of the CAS (Grant No. XDB02070003), and the CAS External Cooperation Key Project. We thank NVIDIA Corporation for donating a GeForce GTX Titan X GPU used in this project.

References

[1] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey, IEEE Trans. Pattern Anal. Mach. Intel. 32 (11) (2010) 1955–1976.
[2] Z. Yang, H. Ai, Demographic classification with local binary patterns, in: Proceedings of the IEEE International Conference Comput. Bio., 2007, pp. 464–473.
[3] F. Gao, H. Ai, Face age classification on consumer images with gabor feature and fuzzy LDA method, in: Proceedings of the IEEE International Conference Comput. Bio., 2009, pp. 132–141.
[4] H. C. Lian, B. L. Lu, Age estimation using a min-max modular support vector machine, in: Proceedings of the International Conference Neural Info. Process., 2005, pp. 83–88.
[5] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit. Workshops, 2015, pp. 34–42.
[6] Y.H. Kwon, N. da Vitoria Lobo,, Age classification from facial images, Comput. Vis. Image Underst. 74 (1) (1999) 1–21.
[7] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intel. 29 (12) (2007) 2234–2240.
[8] G. Guo, G. Mu, Y. Fu, T. Huang, Human age estimation using bio-inspired features, in: Proceedings IEEE International Conference Comput. Vis. Pattern Recognit., 2009, pp. 112–119.
[9] A. Gunay, V. Nabiyev, Automatic age classification with LBP, in: Proceedings of the International Symp. Comput. Info. Sci., 2014, pp. 1–4.
[10] W. Gao, H. Ai, A probabilistic boosting tree for face gender classification on consumer images, in: Proceedings of the IEEE International Conference Comput. Bio., 2009, pp. 169–178.
[11] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2011, pp. 657–664.
[12] S. Yan, H. Wang, T. Huang, X. Tang, Ranking with uncertain labels, in: Proceedings of the IEEE Conference Multimed. Expo., 2007, pp. 96–99.

[13] S. Yan, H. Wang, X. Tang, T. Huang, Learning autostructured regressor from uncertain nonnegative labels, in: Proceedings of the IEEE International Conference Comput. Vis., 2007, pp. 1–8.
[14] T. Fujiwara, H. Koshimizu, Age and gender estimations by modeling statistical relationship among faces, in: Proceedings of the International Conference Knowledge-Based Intel. Info. Eng. Sys., 2003, pp. 559–566.
[15] G. Guo, G. Mu, Y. Fu, C. Dyer, T. Huang, A study on automatic age estimation using a large database, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2009, pp. 1986–1991.
[16] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: CCA vs. PLS, in: Proceedings of the IEEE International Conference Face Gesture, 2013, pp. 1–6.
[17] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Adv. Neural Info. Process. Systems, 2012, pp. 1097–1105.
[18] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference Comput. Vis., 2015, pp. 1026–1034.
[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, abs/1409.1556.
[20] W. Ouyang, X. Wang, Joint deep learning for pedestrian detection, in: Proceedings of the IEEE International Conference Comput. Vis., 2013, pp. 2056–2063.
[21] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2014, pp. 580–587.
[22] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proceedings of the Adv. Neural Info. Process. Systems, 2015, pp. 91–99.
[23] G. Huang, H. Lee, E. Learned, Learning hierarchical representations for face verification with convolutional deep belief networks, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2012, pp. 2518–2525.
[24] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the Adv. Neural Info. Process. Systems., 2014, pp. 1988–1996.
[25] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: Proceedings of the Eur. Conference Comput. Vis., 2014, pp. 94–108.
[26] S. Li, J. Xing, Z. Niu, S. Shan, S. Yan, Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2015, pp. 222–230.
[27] G.B. Orr, K.-R. Müller, Neural Networks: Tricks of the Trade, 2nd edition, Springer, Berlin, Heidelberg, 2012.
[28] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Proceedings of the Asian Conference Comput. Vis., 2014, pp. 144–158.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2015, pp. 1–9.
[30] A. Lanitis, C. Taylor, T. Cootes, Toward automatic simulation of aging effects on face images, IEEE Trans. Pattern Anal. Mach. Intel. 24 (4) (2002) 442–455.
[31] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2007) 578–584.
[32] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensic Secur. 9 (12) (2014) 2170–2179.
[33] K. Ricanek, T. Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in: Proceedings of the IEEE International Conference Face Gesture, 2006, pp. 341–345.
[34] Z. Song, Visual image recognition system with object-level image representation, (Ph.D. thesis), National University of Singapore, 2012.
[35] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, IEEE Trans. Pattern Anal. Mach. Intel. 35 (10) (2013) 2401–2412.
[36] K. Chang, C. Chen, Y. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2011, pp. 585–592.
[37] K.-Y. Chang, C.-S. Chen, A learning framework for age rank estimation based on face images with scattering transform, IEEE Trans. Imag. Process. 24 (3) (2015) 785–798.
[38] E. Frank, M. Hall, A simple approach to ordinal classification, in: Proceedings of the Eur. Conference Mach. Learn., 2001, pp. 145–156.
[39] L. Li, H.-T. Lin, Ordinal regression by extended binary classification, in: Proceedings of the Adv. Neural Info. Process. Systems., 2006, pp. 865–872.
[40] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the IEEE International Conference Comput. Vis. Pattern Recognit., 2015, pp. 2892–2900.
[41] C.-Y. Lee, S. Xie, P.W. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Proceedings of the Arti. Intel. Stat. Conference, 2015, pp. 562–570.
[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference multimed., 2014, pp. 675–678.
[43] S. Li, S. Shan, X. Chen, Relative forest for attribute prediction, in: Proceedings of the Asian Conference Comput. Vis., 2013, pp. 316–327.

Junliang Xing received the B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Xi'an, China, in 2007, and the Ph.D. degree in computer sci ence from Tsinghua University, Beijing, China, in 2012. He is currently an associate p rofessor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Dr. Xing was the recipient of Google Ph.D. Fellowship 2011, the Excellent Student Scholarships at Xi'an Jiaotong Universit y from 2004 to 2007 and at Tsinghua University from 2009 to 2011. He has published more than 40 papers on international journals and conferences. His current research in terests mainly focus on computer vision problems related to faces and humans.

Kai Li received the BE degree from Dalian University of Technology, China, in 2013. Currently, he is a Ph.D. student training in the Institute of Automation, Chinese Academy of Sciences. His research interests include visual attributes analyses and image classification.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University in 1998. From 1998 to 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Currently, he is a full professor in the Institute of Automation, Chinese Academy of Sciences. He has published more than 200 papers on international journals and conferences. His research interests include visual motion analysis and recognition of web objectionable information.

Chunfeng Yuan received the B.S. and M.S. degrees in information science and technology from the Qingdao University of Science and Technology, China, in 2004 and 2007, respectively, and the Ph.D. degree in 2010 from the National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. She is currently working as an assistant professor at Institute of Automation, Chinese Academy of Sciences. Her main research interests include activity analysis and pattern recognition.

Haibin Ling received the BS and MS degrees from Peking University, China, in 1997 and 2000, respectively, and the PhD degree from the University of Maryland, College Park, in 2006. From 2006 to 2007, he worked as a postdoctoral scientist at UCLA. In 2008, he joined Temple University where he is now an associate professor. His research interests include computer vision and medical image analysis.