Pattern Recognition 66 (2017) 95–105

Contents lists available at ScienceDirect

Pattern Recognition

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

D2C: Deep cumulatively and comparatively learning for human age estimation

Kai Li[a], Junliang Xing[a,*], Weiming Hu[a,b], Stephen J. Maybank[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] Department of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, United Kingdom

ARTICLE INFO

Keywords:
Age estimation
Deep learning
Convolutional neural network

ABSTRACT

Age estimation from face images is an important yet difficult task in computer vision. Its main difficulty lies in how to design aging features that remain discriminative in spite of large facial appearance variations. Meanwhile, due to the difficulty of collecting and labeling datasets that contain sufficient samples for all possible ages, the age distributions of most benchmark datasets are often imbalanced, which makes this problem more challenge. In this work, we try to solve these difficulties by means of the mainstream deep learning techniques. Specifically, we use a convolutional neural network which can learn discriminative aging features from raw face images without any handcrafting. To combat the sample imbalance problem, we propose a novel *cumulative hidden layer* which is supervised by a point-wise cumulative signal. With this cumulative hidden layer, our model is learnt indirectly using faces with neighbouring ages and thus alleviate the sample imbalance problem. In order to learn more effective aging features, we further propose a *comparative ranking layer* which is supervised by a pair-wise comparative signal. This comparative ranking layer facilitates aging feature learning and improves the performance of the main age estimation task. In addition, since one face can be included in many different training pairs, we can make full use of the limited training data. It is noted that both of these two novel layers are differentiable, so our model is end-to-end trainable. Extensive experiments on the two of the largest benchmark datasets show that our deep age estimation model gains notable advantage on accuracy when compared against existing methods.

1. Introduction

Age estimation, i.e., predicting the age from a face image, has long been an active research topic in computer vision, with many applications such as age-based face retrieval [1], precision advertising [2], intelligent surveillance [3], human-computer interaction (HCI) [4] and internet access control [2].

The typical methodology for age estimation from face images is to extract carefully designed handcrafted features representing the aging information and subsequently solve an age estimator learning problem. Widely used features include local binary pattern (LBP) [5] and Gabor features [6], with some further processing models like the anthropometric model [7], AGing pattErn Subspace (AGES) [8], and the age manifold model [9]. To learn an age estimator, most approaches use either a multi-class classification framework or a regression framework. In multi-class classification the age values are treated as independent labels and a classifier is learnt to predict the age [1,10,8]. However, age estimation is more of a regression problem than a multi-class classification problem due to the continuity of the age space. Based on this observation, many regression based approaches are proposed [9,11–13].

Although these existing methods achieve promising results, the age estimation problem is far from being solved. The main challenges come from the large appearance variations of face images. Fig. 1 shows some face images from the benchmark datasets used in this work. We can see that the face images may be obtained from people of different races, genders, and under conditions of large pose variations, bad illumination, and heavy makeups, which make it difficult to manually design aging features that are robust to all these disturbances. In addition, due to the difficulty of collecting and labeling datasets that contain sufficient samples for all possible ages, the age distributions of most available benchmark datasets in the literature are imbalanced which

* Corresponding author.
E-mail addresses: kai.li@nlpr.ia.ac.cn (K. Li), jlxing@nlpr.ia.ac.cn (J. Xing), wmhu@nlpr.ia.ac.cn (W. Hu), sjmaybank@dcs.bbk.ac.uk (S.J. Maybank).

http://dx.doi.org/10.1016/j.patcog.2017.01.007
Received 14 July 2016; Received in revised form 18 December 2016; Accepted 6 January 2017
Available online 06 January 2017
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

Fig. 1. Examples of faces in the two benchmark datasets used in this work. Top row: the Morph II dataset. Bottom row: the WebFace dataset.

makes accurate age estimation even harder.

In this work, we try to solve the aforementioned challenges in human age estimation. Instead of manually design features, we use a convolutional neural network (CNN) to extract effective and discriminative aging features from raw input face images without any handcrafting. To combat the sample imbalance problem, we propose a novel cumulative hidden layer (Section 3.1). In contrast with the mainstream CNN models which directly map the last hidden layer to the output layer, we insert a cumulative hidden layer before the output layer. This cumulative hidden layer is supervised by a point-wise cumulative signal which encodes the target age labels continuously. Thanks to this cumulative hidden layer, our model can not only learn from one face itself but also from the faces with neighbouring ages and thus alleviate the sample imbalance problem.

In order to learn more effective aging features, we further propose a novel comparative ranking layer (Section 3.2) which is supervised by a pair-wise comparative signal, i.e., who is older. The intuition behind this is that it is difficult to tell accurately the age of one face, but it is relatively easy to tell who is older, given two faces. For example, in Fig. 1, it is hard to guess the exact age of these faces, but it is relatively easy to see that the faces to the right of the figure are older than the faces to the left. This comparative signal helps our model to learn the general concept of “old and young”. This concept is valuable for the exact age estimation task. We argue that this auxiliary pair-wise signal facilitates aging feature learning and improves the performance of the main age estimation task. As one face image can be used in many different pairs, we can make full use of the training data. It's worth noting that both the point-wise and pair-wise supervision signals can be obtained directly from the age labels, so our model does not need any additional manual labeling.

There are three main contributions in this work:

1. We propose a novel cumulative hidden layer which alleviates the sample imbalance problem and thus improves age estimation. To the best of our knowledge, this is the first time that a new layer for the CNN has been designed to combat the sample imbalance problem in human age estimation literatures.
2. We propose a novel comparative ranking layer which facilitates aging feature learning and thus further improve age estimation. We believe that this is the first work that explicitly take account of the pair-wise information between faces during training for human age estimation.
3. By incorporating these two novel layers, we obtain a deep age estimation model which outperforms by a large margin all previous age estimation methods on two of the largest benchmark datasets.

2. Related work

Human age estimation has been studied for decades in the computer vision community. Previous works on age estimation are mainly focused on the manual design of robust ageing features. Typical features designed specifically for age estimation include facial features and wrinkles [7], the learned AGES (AGing pattErn Subspace) [8] features, as well as the biologically inspired features (BIF) [13]. Other more general features devised for texture description are also widely used for age estimation, for example the LBP feature [5,14], the Gabor feature [6], etc.

Based on these carefully designed handcrafted facial aging features, much attention was paid to the age estimator learning step: age estimation by classification or regression. Classification models, e.g.linear SVM [13], Probabilistic Boosting Tree [15], Fuzzy LDA [6], or regression models like Support Vector Regression [13], Kernel Partial Least Squares [16], Neural Network [17] and Semidefinite Programming [18] are all designed to estimate age.

Although a lot of algorithms have achieved promising age estimation results, many challenges still remain in this problem. One of the most prominent challenges is the sample imbalance problem. There are several attempts [19–21] to alleviate this problem which are based on the concept of label distribution learning (LDL) [22]. The label distribution can be seen as an extension of the one-hot encoding in the classic multi-class classification problem. These LDL based age estimation methods represent each target age with a label distribution vector which can capture the correlations between different ages and have been shown to alleviate the sample imbalance problem to a certain extent. Different from these LDL based methods which first design handcrafted aging features and then train the age classifier separately, our model with the proposed cumulative hidden layer learns the aging features and the age regressor in an end-to-end manner, which is more effective to alleviate the sample imbalance problem.

Recently, deep learning models, especially convolutional neural networks (CNNs), have achieved great successes in many computer vision tasks [23–30]. One of the most attractive merits of deep learning is the automatic learning of the features and the classifier at the same time. Although CNNs have been successful in many computer vision problems, there are only a very few studies on using CNNs to perform age estimation [31–33]. Some of these studies are focused on other objectives, e.g., providing a benchmark dataset [31], or exploiting complicated network architectures, such as the multi-scale architecture with 23 sub-networks in [32], and the tree-structured architecture with 36 local sub-networks in [33]. Unlike these existing complicated CNN based models which have many hyper-parameters to tune and which are hard to implement, our model is based on the widely used AlexNet [24] which is easy to reproduce.

In contrast with the existing models, which only use the point-wise age label of one face as supervision signal, our model also exploits the proposed pair-wise comparative supervision signal between two faces and thus outperforms existing models significantly. Pair-wise supervision signal is commonly adopted in hashing. Representative pair-wise supervision based hashing methods include sequential projection learning for hashing [34], minimal loss hashing [35], supervised hashing with kernels [36], two-step hashing [37], fast supervised hashing [38] and deep hashing [39]. The pair-wise supervision signal in hashing methods is used to indicate whether the semantic labels are similar between two items. In contrast, our pair-wise comparative supervision signal is used to indicate the order between the ages of two

faces. The purpose of the pair-wise supervision signal in hashing is to learn compact semantic similarity preserving binary codes. While our pair-wise comparative supervision signal is used to facilitate the aging feature learning.

3. Methodology

In this section we introduce our model, called *Deep Cumulatively and Comparatively* (D2C) supervised age estimation model. Our D2C model simultaneously learns aging features and age estimator in an end-to-end framework. The D2C model exploits our proposed cumulative hidden layer and comparative ranking layer which are supervised by the point-wise cumulative and pair-wise comparative signals, respectively. In the following, we will first introduce the cumulative hidden layer and the comparative ranking layer, and then describe in detail the architecture of the entire D2C age estimation model.

3.1. Cumulative hidden layer

Age estimation can be directly formulated as a multi-class classification problem. This multi-class classification formulation assumes that the images obtained at one particular age are independent of the images obtained at neighbouring ages. In fact, the images obtained at nearby ages are strongly correlated. Based on this observation, it is more natural to formulate age estimation as a regression problem.

Traditional CNN based age regression models directly map the features extracted by the network to the age label (cf. Fig. 2(a)). However, in real-world, usually the age distribution of collected faces is imbalanced. The imbalanced training data causes difficulties in learning the regressor directly since there are only a few samples or even no sample available for certain ages.

To combat the sample imbalance problem, we insert a novel cumulative hidden layer (CHL) before the age output layer (cf. Fig. 2(b)). Our CHL is initially inspired by [40]. In [40], the handcrafted features are designed first and the regressor is learnt separately, while our model learns the aging features and the age regressor in an end-to-end manner. This CHL is supervised by a binary cumulative signal which is obtained directly from the age label. Concretely, suppose given a set of N training face images $\{x_i, l_i\}$, $l_i \in \{1, 2, \ldots, K\}$, $i = 1, 2, \ldots, N$, where x_i denotes the i-th face image, l_i denotes its age label, and K is the number of different ages in the training set. For the i-th face image x_i with age label l_i, we can construct its corresponding K-dimensional binary cumulative signal CuS_i from l_i as follows:

$$\mathrm{CuS}_i^k = \begin{cases} 1, & k \le l_i \\ 0, & k > l_i \end{cases}, \tag{1}$$

where $k = 1, 2, \ldots, K$, and CuS_i^k denotes the k-th element of CuS_i.

This cumulative signal has one appealing property: the cumulative

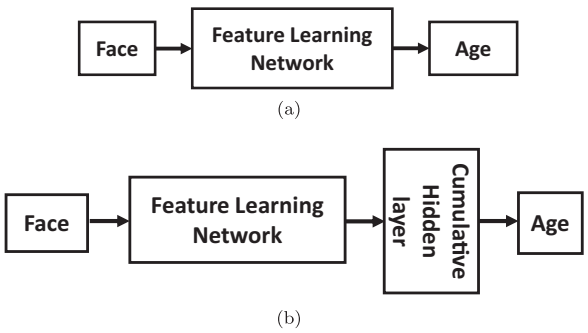

Fig. 2. Schematic diagrams of the traditional CNN based age regression model (top), and our CNN based age regression model with the proposed cumulative hidden layer (bottom). The feature learning network is a series of convolutional layers, pooling layers and fully connected layers.

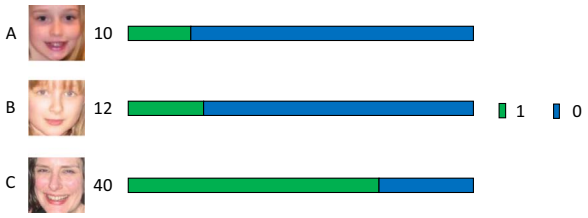

Fig. 3. Three example face images (left), their age labels (middle) and the corresponding cumulative signals (right). It is apparent that A and B are similar, but C is very different from A and B. This is consistent with the differences in their cumulative signals.

signals of neighbouring ages are more similar than those further apart which is consistent with the fact that faces with neighbouring ages are generally more similar in appearance than faces with widely separated ages. For example, in Fig. 3, the 10-year-old face is more similar to the 12-year-old face than to that of the 40-year-old face, and the cumulative signal of the 10-year-old face is also more similar to that of the 12-year-old face (2-bit difference) than that of the 40-year-old face (30-bit difference). This nice property is of help in estimating the ages, especially when the age distribution is imbalanced, because similar ages can be used to partially depict their neighbouring ages that are few or absent in the learning and thus alleviate the sample imbalance problem. Based on the analyses above, we can see that our CHL supervised by this cumulative signal can not only capture the correlations between faces of different ages but also alleviate the sample imbalance problem, both of which are beneficial for accurate age estimation.

For an input image x_i along with its target cumulative signal CuS_i and age label l_i, we use $\boldsymbol{\phi}_i \in \mathbb{R}^D$ to denote the aging feature of x_i learned by the CNN. Then the output of the CHL is:

$$\mathbf{o}_i = \mathbf{W}\boldsymbol{\phi}_i + \mathbf{b}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{K \times D}$, $\mathbf{b} \in \mathbb{R}^K$ are the parameters of the CHL. The input to the final age output layer is the output of the CHL, so the predicted age is calculated as follows:

$$\tilde{l}_i = \mathbf{w}^T \mathbf{o}_i + b, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^K$, $b \in \mathbb{R}$ are the parameters of the output layer. We want to minimize the difference between the output of CHL $\mathbf{o}_i$ and the target cumulative signal CuS_i. At the same time, we want to minimize the difference between the predicted age $\tilde{l}_i$ and the target age l_i. Consequently, the overall loss function of the model in Fig. 2(b) is defined as follows:

$$\mathrm{L}_i = \mathrm{Loss}_i^{\mathrm{age}} + \alpha \mathrm{Loss}_i^{\mathrm{CHL}} = |\tilde{l}_i - l_i| + \alpha \parallel \mathbf{o}_i - \mathrm{CuS}_i \parallel_1, \tag{4}$$

where $\mathrm{Loss}_i^{\mathrm{CHL}}$ is the loss of the CHL with output $\mathbf{o}_i$, $\mathrm{Loss}_i^{\mathrm{age}}$ is the loss of the age output layer with the predicted age $\tilde{l}_i$, and α is the hyper-parameter to tune the importance of each loss. For simplicity, we denote the loss function for a single face image in Eq. (4). The total loss is averaged over all face images in a batch during training. It's worth noting that unlike other regression based age estimation methods which always use L2-norm to calculate the loss, our model uses L1-norm in Eq. (4) which is more robust to outliers.

Our model with this novel CHL is similar to the very successful attribute based models used in many computer vision problems [41–43]. Structurally, these attribute based models are *two-stage mapping*, i.e., they first map the visual features to the attribute space and then map this attribute space to the label space. The attribute space is design to capture the correlations between different classes, so the model can be learned indirectly even if there is little or no samples of a class. Similarly, our deep age estimation model first maps the aging features to the cumulative space by using the CHL, and then maps this cumulative space to the output age label space. The cumulative space captures the correlations between different ages and thus alleviates the sample imbalance problem effectively.

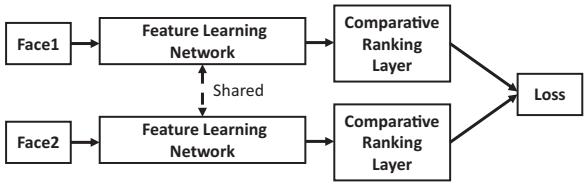

Fig. 4. Schematic diagram of our proposed comparative ranking layer.

3.2. Comparative ranking layer

It is worth noting that learning a function from face images to ages is a relatively difficult task. Even human beings find it difficult to estimate age accurately from a face image, but it is relatively easy to tell who is older between two face images. As shown in Fig. 1, it is difficult to tell the exact age of each face, but we can relatively easy to see that faces on the right are older than the faces on the left even though we do not know the exact ages of those faces. Based on this observation, we propose a novel comparative ranking layer (CRL) which is supervised by a pair-wise comparative signal, i.e., who is older. This auxiliary comparative signal helps the model to learn the general concept of "old and young". This concept is valuable for the exact age estimation task.

The schematic diagram of the network with our proposed CRL is shown in Fig. 4. Given a pair of face images (x_i, x_j) along with their ground-truth age labels (l_i, l_j), the comparative signal CoS_{ij} is defined as follows:

$$\mathrm{CoS}_{ij} = \begin{cases} 1, & if \quad l_i > l_j \\ 0.5, & if \quad l_i = l_j \\ 0, & if \quad l_i < l_j \end{cases}. \tag{5}$$

We can think of CoS_{ij} as the target probability of x_i is older than x_j, i.e., $\mathrm{CoS}_{ij} = 1$ represents that x_i is older than x_j, $\mathrm{CoS}_{ij} = 0$ represents that x_j is older than x_i, and $\mathrm{CoS}_{ij} = 0.5$ represents that x_i is the same age as x_j. This pair of images (x_i, x_j) go though two feature extraction networks with shared weights, this procedure maps the face images onto D-dimensional feature vectors $(\boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j)$. The aim of the CRL is to learn a *ranking* function $f: \mathbb{R}^D \mapsto \mathbb{R}$ that shows who is older, e.g., $f(\boldsymbol{\varphi}_i) > f(\boldsymbol{\varphi}_j)$ indicates that x_i is older than x_j. Based on this consideration, we choose the CRL to be a fully connected layer with a single output neuron, i.e.,

$$f(\boldsymbol{\varphi}_i) = \mathbf{w}^T \boldsymbol{\varphi}_i + b, \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^D$, $b \in \mathbb{R}$ are the parameters of the CRL. After we obtain the scores of two face images, i.e., $f(\boldsymbol{\varphi}_i)$ and $f(\boldsymbol{\varphi}_j)$. In a similar way to [44], we map from these scores to the posterior probability $p_{ij} = P(x_i \succ x_j)$ using a logistic function, i.e.,

$$p_{ij} = P(x_i \succ x_j) = \frac{1}{1 + e^{-(f(\boldsymbol{\varphi}_i) - f(\boldsymbol{\varphi}_j))}}, \tag{7}$$

where $x_i \succ x_j$ denotes that x_i is older than x_j. The definition of p_{ij} in Eq. (7) has a nice *consistency* property, i.e., given $p_{ij} > 0.5$ and $p_{jk} > 0.5$, based on the definition of Eq. (7), we can derive $p_{ik} > 0.5$. In other words, when $x_i \succ x_j$ and $x_j \succ x_k$ then $x_i \succ x_k$.

We use the binary cross entropy loss function to calculate the loss for a face image pair (x_i, x_j) along with the target CoS_{ij}:

$$\mathrm{Loss}_{ij}^{\mathrm{rank}} = -\mathrm{CoS}_{ij} \log p_{ij} - (1 - \mathrm{CoS}_{ij}) \log(1 - p_{ij}). \tag{8}$$

Fig. 5 shows the value of $\mathrm{Loss}_{ij}^{\mathrm{rank}}$ as a function of $f(\boldsymbol{\varphi}_i) - f(\boldsymbol{\varphi}_j)$ for the three values of the target CoS_{ij}. We can see that when the target $\mathrm{CoS}_{ij} = 1$ (0), i.e., x_i is older (younger) than x_j, minimizing the loss in Eq. (8) pushes $f(\boldsymbol{\varphi}_i)$ to be larger (smaller) than $f(\boldsymbol{\varphi}_j)$ which meets our requirements that the score output by f can reflect who is older. Note that when the target $\mathrm{CoS}_{ij} = 0.5$, i.e., x_i is the same age as x_j. The loss in Eq. (8) becomes symmetric (the green line in Fig. 5) and with its minimum at the origin, i.e., $f(\boldsymbol{\varphi}_i) = f(\boldsymbol{\varphi}_j)$. This gives us a principled way of training on face pairs that are known to have the same age.

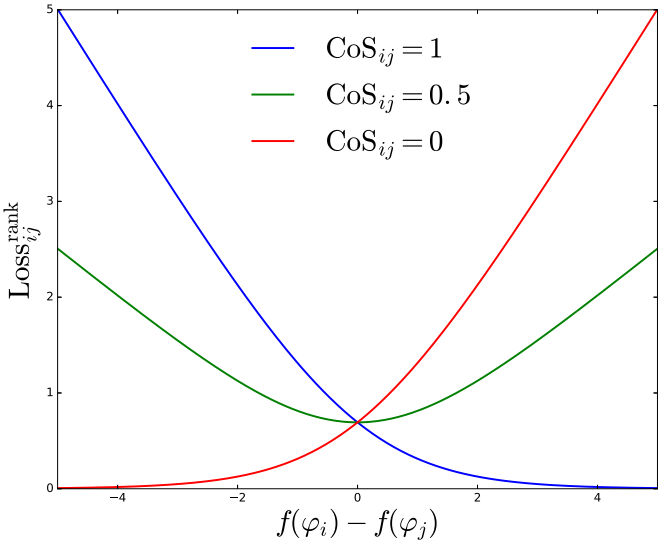

Fig. 5. The value of $\mathrm{Loss}_{ij}^{\mathrm{rank}}$ for three values of the target CoS_{ij}.

It is noteworthy that not all the training face pairs have the same degree of difficulty. For example, suppose given two face pairs (x_a, x_b) and (x_c, x_d), where l_a=50, l_b=10, l_c=30, and l_d=25. It is easier to judge $x_a \succ x_b$ than to judge $x_c \succ x_d$. We use $|l_i - l_j|$ to measure the difficulty of a face pair (x_i, x_j). Inspired by the concept of "curriculum learning" proposed in [45], we use the easy face pairs at the beginning and gradually increase the difficulty of the face pairs. By using this strategy our model can gradually learn more complex and discriminative aging features from the subtle facial difference between face pairs which are critical to accurate age estimation. In addition, we can make full use of the small amount of face images with specific age since one face image can be used in a lot of different training pairs, and thus alleviate the sample imbalance problem to some extent.

It is noted that our comparative ranking layer does not take account of the exact age of each face. Instead, it only uses the relative order between faces. This information is more stable than exact age values. Compare to the exact age label supervision signal which only contains the information of one face, this comparative signal considers the pair-wise information between two faces which provides complemental information. By training with face pairs, the model learns more discriminative aging features by directly learning from the difference between faces. As is mentioned before, it is easier to distinguish who is older between two faces than to tell the exact age of one face. We argue that this related and relatively easy task is beneficial to the aging feature learning and thus improve the main exact age estimation task. This is also been verified in other works such as [30,46] that some related and easy tasks can boost the performance the main difficult task.

3.3. D2C network architecture

Fig. 6 shows the entire end-to-end architecture of our deep cumulatively and comparatively (D2C) supervised age estimation model which incorporates the proposed cumulative hidden layer (CHL) and comparative ranking layer (CRL) discussed above. Note that there are two CNNs in Fig. 6, however, these two CNNs are identical in that they have the same structure and parameters. We use two CNNs to get a better illustration for the comparative ranking layer which is based on a pair of face images. We exploit the widely used AlexNet [24] as the "root" net (the gray part in Fig. 6). Other modern CNN architectures [26,47] can also be used as the root net, but a comparison of different network architectures is not the focus of this work. Next, we describe in detail our D2C age estimation model.

The root net is the gray network in Fig. 6. The network has five

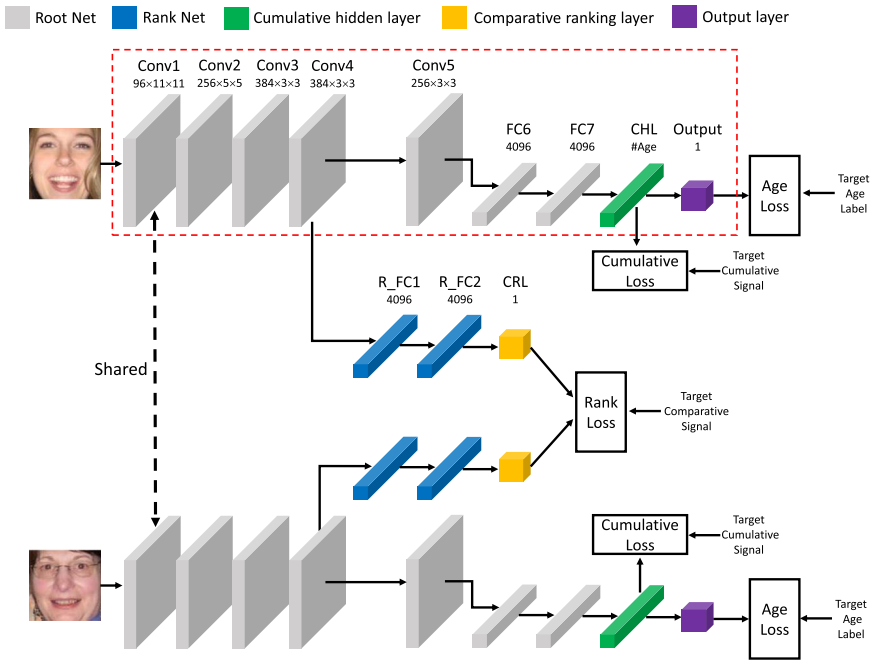

Fig. 6. The end-to-end deep architecture of our D2C age estimation model.

convolutional layers and two fully connected layers. We use Rectified Liner Units (ReLu) as the activation function. The first convolutional layer (Conv1) consists of 96 kernels with size of 11×11, followed by a local response normalization (LRN) layer and a 3×3 max pooling (MP) layer. The second convolutional layer (Conv2) has 256 5×5 kernels, followed by a LRN layer and a 3×3 MP layer. The third convolutional layer (Conv3) has 384 3×3 kernels. It is followed by the fourth convolutional layer (Conv4) with 384 3×3 kernels. The fifth convolutional layer (Conv5), with 256 3×3 kernels, is followed by a 3×3 MP layer. The convolutional layers are followed by two 4096-dimensional fully connected layers (FC6 and FC7). The FC7 layer is followed by the cumulative hidden layer discussed in Section 3.1. The dimension of the cumulative hidden layer is equal to the number of different ages (#Age) in the training data. The last layer outputs the predicted age.

Similar to the auxiliary intermediate supervision branch in [47], the input to the rank net (the blue part in Fig. 6) is obtained from the Conv4 output of the root net. This choice is also based on the consideration that the main age estimation task and the auxiliary ranking task are not of the same difficulty. The main age estimation task is a difficult task and thus requires the highest-level features. Compared to the main age estimation task, the ranking task introduced by the comparative layer is a relatively easy task (i.e., binary classification) which requires slightly lower-level features. This network passes the input through a 3×3 MP layer followed by two 4096-dimensional fully connected layers (R_FC1 and R_FC2). The resulting data is passed to the comparative ranking layer discussed in Section 3.2.

The overall loss of our D2C age estimation model for a pair of input face images (x_i, x_j) with the target age labels (l_i, l_j), the target cumulative signals $(\mathrm{CuS}_i, \mathrm{CuS}_j)$, and the target comparative signal CoS_{ij} is defined as the weighted sum of Eqs. (4) and (8), i.e.,

$$\mathrm{Loss}_{ij}^{\mathrm{overall}} = \sum_{m=i,j} \mathrm{Loss}_m^{\mathrm{age}} + \alpha \sum_{m=i,j} \mathrm{Loss}_m^{\mathrm{CHL}} + \beta \mathrm{Loss}_{ij}^{\mathrm{rank}}, \tag{9}$$

where α, β are hyper-parameters to tune the importance of each loss. $\mathrm{Loss}^{\mathrm{age}}$ and $\mathrm{Loss}^{\mathrm{CHL}}$ are equally important since they are the loss functions of the main age estimation task. Therefore, we fix $\alpha = 1$ throughout the experiments. $\mathrm{Loss}^{\mathrm{rank}}$ is the loss function of the auxiliary task which facilitates aging feature learning during training and β is used to balance this auxiliary task and the main age estimation task. Therefore, we only adjust the value of β in our experiments. We choose $\beta = 0.5$ based on a held-out validation set. Unlike the mainstream CNN architectures, our D2C model is not a chain-like net. However, it is based on a directed-acyclic graph which can be trained end-to-end from scratch using back-propagation and stochastic gradient descent. Since our main purpose is age estimation, the rank net is only used to facilitate aging feature learning which is easier than and converges faster than the main age regression task. Based on this observation, we early stop the rank net which is similar to the procedure proposed in [30] to avoid overfitting. Specifically, we remove $\mathrm{Loss}_{ij}^{\mathrm{rank}}$ in Eq. (9) when its value no longer decreases. At testing time, we only use the network inside the red dashed line in Fig. 6 to predict the age of an input face. This procedure is very efficient because it only requires one forward pass through the network.

4. Experiments

In this section, we first describe the age estimation benchmark datasets used in this work, the age estimation performance evaluation metric, and the experimental settings. Then, we will conduct detailed experiments to validate the effectiveness of our proposed cumulative hidden layer and comparative ranking layer. Finally, we will compare our D2C age estimation model with the state-of-the-art age estimation methods.

4.1. Datasets and experimental settings

4.1.1. Datasets

There are many datasets for age estimation in the literature [48,9,49]. Most of these datasets, however, are relatively small. Since training a good deep neural network generally requires a large amount of training data, we select two of the largest benchmark datasets, i.e., the Morph II [50] dataset and the WebFace [51] dataset as our testbeds.

Morph II dataset: The Morph II dataset contains about 55,000 face images of more than 13,000 subjects with ages ranging from 16 to 77 years old. Morph II is a multi-ethnic dataset. It has about 77% Black faces and 19% White faces, while the remaining 4% includes Asian, Hispanic, Indian, and Other. We follow the previous study [16], and

Table 1
The number of images of the three splits of the 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

split this dataset into three non-overlapping subsets S1, S2 and S3 (cf. Table 1). In all the experiments the training and testing are repeated twice: 1) training on S1, testing on S2+S3 and 2) training on S2, testing on S1+S3. This training and testing set split protocol has become the standard for the Morph II age estimation dataset.[1]

WebFace dataset: The WebFace dataset contains 59,930 face images. The ages range from 1 to 80 years old. The WebFace dataset is also a multi-ethnic dataset. In contrast with the Morph II dataset, this dataset is captured in the wild. The images contain large pose and expression variations, which make this dataset much more challenging. Following [51], we conduct experiments on this dataset using a four-fold cross validation protocol.

Fig. 1 shows some example face images in these two datasets. As we can see, both datasets are very challenging and thus can serve as very good benchmarks for evaluating the performance of different age estimation methods.

4.1.2. Evaluation metric

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

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{10}$$

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. Smaller MAE values mean better age estimation performance.

4.1.3. Experimental settings

The face images in the datasets are preprocessed in a standard way, i.e., the faces in the images are detected and aligned, then cropped and normalized to 256×256. Fig. 7 shows some examples of the original images and their corresponding preprocessed versions. In all the following experiments, we use the Caffe [52] 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.[2] We train all the networks using mini-batch (set to 256) 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 [24], i.e., randomly cropping of 227×227 pixels from the 256×256 input face image, then randomly flipping it before feeding it to the network. The initial learning rate is 10^{-3} which is divided by 10 when the training curve reaches a plateau. These hyper-parameters are chosen based on the validation set. We found that all networks converge well under these settings, so we use the same hyper-parameters for different models to make fair comparisons.

4.2. Analyses of our novel cumulative hidden layer

To demonstrate the effectiveness of our cumulative hidden layer, we train two networks, the first without and the second with this layer. The networks are denoted by Net_{base} and Net_{CHL} respectively. The age estimation results of these two models on the Morph II and WebFace datasets are shown in Tables 2, 3. We can clearly see that Net_{CHL} outperforms Net_{base} on both datasets. These experimental results validate the effectiveness of our cumulative hidden layer for age estimation.

4.2.1. Missing data experiments

In real-world, usually the age distribution of face images collected is imbalanced or say incomplete with some ages lost. To more explicitly demonstrate that our cumulative hidden layer can alleviate this problem, we evaluate Net_{base} and Net_{CHL} while making the training data more and more imbalanced. To simulate such a scenario, we remove all the face images every T years, where $T \in \{6, 5, 4\}$, so the training data become more and more imbalanced as T decreases. We retrain Net_{base} and Net_{CHL} on both datasets at different values of T. Tables 4, 5 show the age estimation results. It is evident from these two tables that when more training data are removed and the training data become more and more imbalanced, the performance of both Net_{base} and Net_{CHL} degrades. However, Net_{CHL} performances consistently better than Net_{base} on both datasets under different values of T. These results show that our proposed cumulative hidden layer dose alleviate the sample imbalance problem and therefore improve the age estimation performance.

4.2.2. More parameters lead to better performance?

The Net_{CHL} has a total of 9 learnable layers. On the other hand, the Net_{base} has 8 learnable layers. As increasing the number of learnable parameters can enlarge the model capacity and in some cases lead to better performance, one could argue that the performance improvement in our Net_{CHL} comes merely from the additional parameters introduced by the cumulative hidden layer. To disprove this, we train another model $\text{Net}_{\text{base}}^{\text{Aug}}$ by augmenting Net_{base} with an additional layer such that the number of parameters of $\text{Net}_{\text{base}}^{\text{Aug}}$ is the same as Net_{CHL}. We found that the additional layer leads to a degradation rather than to an improvement in performance for Net_{base}: the MAE increases from 3.31 to 3.32 on the Morph II dataset. Similarly, the MAE increases from 6.34 to 6.36 on the WebFace dataset. This suggests that the gain in performance of Net_{CHL} over Net_{base} derives from our proposed cumulative hidden layer and the cumulative supervision signal rather than from an increased number of parameters.

4.2.3. L2-norm vs. L1-norm

The L2-norm is widely used in regression based age estimation problem since it has very nice mathematical properties such as convexity and continuously differentiable. However, the L2-norm is sensitive to errors in the labels. Since label errors are inevitable in real world datasets, we use the more robust L1-norm to calculate the loss in Eq. (4). To demonstrate the superiority of the L1-norm for age estimation, we train another model $\text{Net}_{\text{CHL}}^{\text{L2}}$ using the L2-norm in the loss function. Compared with Net_{CHL} which uses L1-norm in the loss function, the MAE of $\text{Net}_{\text{CHL}}^{\text{L2}}$ increases from 3.16 to 3.18 on the Morph II dataset, and from 6.12 to 6.53 on the WebFace dataset. Since the WebFace dataset is automatically compiled from images on the Web and contains many more label errors than the Morph II dataset, the performance gap between Net_{CHL} and $\text{Net}_{\text{CHL}}^{\text{L2}}$ is much larger on the WebFace dataset than on the Morph II dataset. This clearly demonstrates the effectiveness of L1-norm for age estimation when faced with a noisy data set. We can also see that even though the Morph II dataset was compiled in a controlled environment and has few label errors, Net_{CHL} still performs slightly better than $\text{Net}_{\text{CHL}}^{\text{L2}}$ on this dataset. This is because MAE is the evaluation metric for age estimation (Eq. (10))

[1] http://csee.wvu.edu/~gdguo/Data/AgingDataPartition.htm
[2] https://github.com/BVLC/caffe/wiki/Model-Zoo

Fig. 7. Examples of the original face images and their corresponding preprocessed versions after face detection and alignment. Left two: the Morph II dataset. Right two: the WebFace dataset.

Table 2
The age estimation results of Net_{base} and Net_{CHL} on the Morph II dataset using the training and testing set split protocol in Table 1.

Method	S2+S3 MAE	S1+S3 MAE	Average MAE
Net_{base}	3.31	3.30	3.31
Net_{CHL}	**3.15**	**3.16**	**3.16**

Table 3
The age estimation results of Net_{base} and Net_{CHL} on the WebFace dataset using the four-fold cross validation protocol.

Method	Fold1 MAE	Fold2 MAE	Fold3 MAE	Fold4 MAE	Average MAE
Net_{base}	6.39	6.33	6.32	6.31	6.34
Net_{CHL}	**6.13**	**6.14**	**6.07**	**6.14**	**6.12**

Table 4
The age estimation results of Net_{base} and Net_{CHL} on the Morph II dataset at different T values.

Method	T=6 MAE	T=5 MAE	T=4 MAE
Net_{base}	3.54	3.60	3.87
Net_{CHL}	**3.33**	**3.37**	**3.50**

Table 5
The age estimation results of Net_{base} and Net_{CHL} on the WebFace dataset at different T values.

Method	T=6 MAE	T=5 MAE	T=4 MAE
Net_{base}	6.64	6.86	7.02
Net_{CHL}	**6.39**	**6.50**	**6.70**

which is defined using the L1-norm, so we can directly optimize this metric by using the L1-norm as a loss function. This is also the philosophy of deep learning, i.e., direct optimization of what you want can always improve the performance. Some people may concern that the loss function in Eq. (4) has many indifferentiable points which may not be easy to optimize. In fact, with recent developments in optimizing non-smoothing functions like ReLu [24] and PReLu [25] in the deep learning framework, the loss function in Eq. (4) can be optimized effectively using the stochastic gradient descent algorithm. In order to make this clear, we plot the validation MAE of Net_{CHL} and $\text{Net}^{\text{L2}}_{\text{CHL}}$ during training on the WebFace dataset in Fig. 8 (we don't plot the training loss because the training loss based on L1-norm and L2-norm can't be directly compared). We can see that Net_{CHL} converges without any difficulties and obtains consistently better validation performance than $\text{Net}^{\text{L2}}_{\text{CHL}}$ during training. These experimental results and analyses validate the effectiveness of our choice of using L1-norm as the loss function for age estimation.

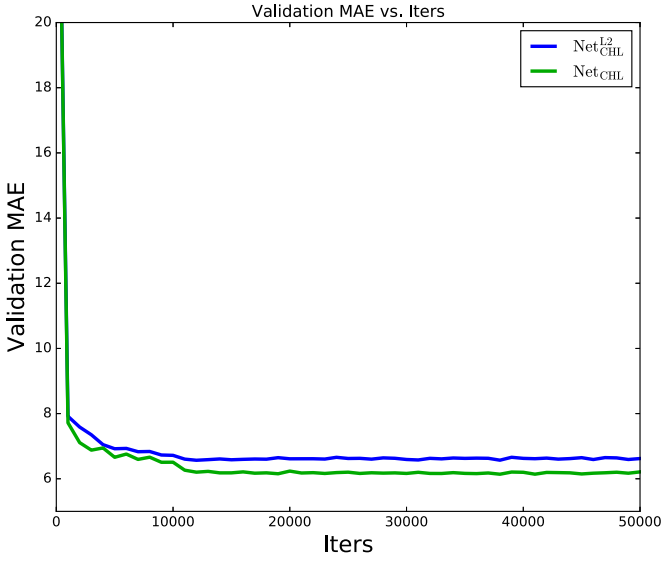

Fig. 8. Validation MAE of Net_{CHL} and $\text{Net}^{\text{L2}}_{\text{CHL}}$ observed during training on the WebFace dataset.

4.2.4. Comparisons with label distribution learning based methods

Label distribution learning (LDL) based methods are very effective to deal with the sample imbalance problem in age estimation. Different from the classic one-hot encoding based multi-class classification for age estimation, the LDL based methods represent each age label with a label distribution vector which captures the correlations between different ages and thus can alleviate the sample imbalance problem to some extent. In order to compare our Net_{CHL} with these LDL based methods, we train two other networks Net_{CLC} and Net_{LDL}. Net_{CLC} is the classic one-hot encoding multi-class classification based age estimation network, and Net_{LDL} is an age estimation network based on the LDL proposed by Geng et al. [19]. The age estimation results of these three networks on both datasets are shown in Tables 6, 7. We can see that Net_{LDL} outperforms Net_{CLC} on both datasets. This is because compared with Net_{CLC} which treats each age label independently, Net_{LDL} captures the correlations between different ages and improves the age estimation performance. We can also see that our Net_{CHL} with the proposed cumulative hidden layer obtains better results than Net_{LDL}. There are two reasons to explain these results. First, on the whole, our Net_{CHL} is a regression based age estimation method, while Net_{LDL} is a classification based method. Compared to the classification based formulation, the

Table 6
The age estimation results of Net_{CLC}, Net_{LDL} and Net_{CHL} on the Morph II dataset using the training and testing set split protocol in Table 1.

Method	S2+S3 MAE	S1+S3 MAE	Average MAE
Net_{CLC}	3.57	3.64	3.61
Net_{LDL}	3.36	3.40	3.38
Net_{CHL}	**3.15**	**3.16**	**3.16**

Table 7
The age estimation results of Net_{CLC}, Net_{LDL} and Net_{CHL} on the WebFace dataset using the four-fold cross validation protocol.

Method	Fold1 MAE	Fold2 MAE	Fold3 MAE	Fold4 MAE	Average MAE
Net_{CLC}	6.67	6.84	6.72	6.79	6.76
Net_{LDL}	6.46	6.47	6.34	6.35	6.41
Net_{CHL}	**6.13**	**6.14**	**6.07**	**6.14**	**6.12**

Table 8
The age estimation results of Net_{CHL} and Net_{D2C} on the Morph II dataset using the training and testing set split protocol in Table 1.

Method	S2+S3 MAE	S1+S3 MAE	Average MAE
Net_{CHL}	3.15	3.16	3.16
Net_{D2C}	**3.06**	**3.05**	**3.06**

regression based formulation is more favorable owing to the inherent characteristic of age estimation, i.e., the age of an individual is measured by the time passed from the individual's birth, and thus is a continuous process. Second, compared with Net_{LDL} using the Kullback-Leibler (KL) divergence as the loss function, our Net_{CHL} is an end-to-end framework using MAE as the loss function which can directly optimize the evaluation metric of age estimation.

4.3. Analyses of our novel comparative ranking layer

In this section we demonstrate the effectiveness of our proposed comparative ranking layer in improving age estimation performance. It is noted that our Net_{CHL} has already obtained state-of-the-art results on both datasets. A question arises: can the comparative ranking layer further improve age estimation? To answer this question, we train our D2C age estimation model Net_{D2C} by incorporating both the cumulative hidden layer and the comparative ranking layer (Fig. 6). The results are shown in Tables 8, 9. From these tables, we can see that Net_{D2C} is better than Net_{CHL} on both datasets. This shows that our proposed comparative ranking layer indeed can further improve the age estimation performance.

In order to better illustrate the role of our comparative ranking layer, we plot the age estimation MAE loss observed during training on the Morph II and the WebFace datasets in Figs. 9 and 10. We can see that the Net_{D2C}, which includes the comparative ranking layer, can find better minimum than Net_{CHL} without this layer. This validates our hypothesis that the comparative ranking layer can facilitate the aging feature learning process.

Some age estimation results obtained from Net_{CHL} and Net_{D2C} are shown in Fig. 11. We can see that even though the left face is younger than the right face in each pair by ground truth, Net_{CHL} predicts the opposite in these examples. In contrast, thanks to our proposed comparative ranking layer which explicitly consider the pair-wise information between faces during training, so the Net_{D2C} can learn discriminative aging feature from the subtle facial difference between face pairs with similar ages and thus makes more accurate predictions than Net_{CHL}. All the above results and analyses validate the effectiveness of our comparative ranking layer for human age estimation.

Table 9
The age estimation results of Net_{CHL} and Net_{D2C} on the WebFace dataset using the four-fold cross validation protocol.

Method	Fold1 MAE	Fold2 MAE	Fold3 MAE	Fold4 MAE	Average MAE
Net_{CHL}	6.13	6.14	6.07	6.14	6.12
Net_{D2C}	**6.03**	**6.07**	**5.99**	**6.06**	**6.04**

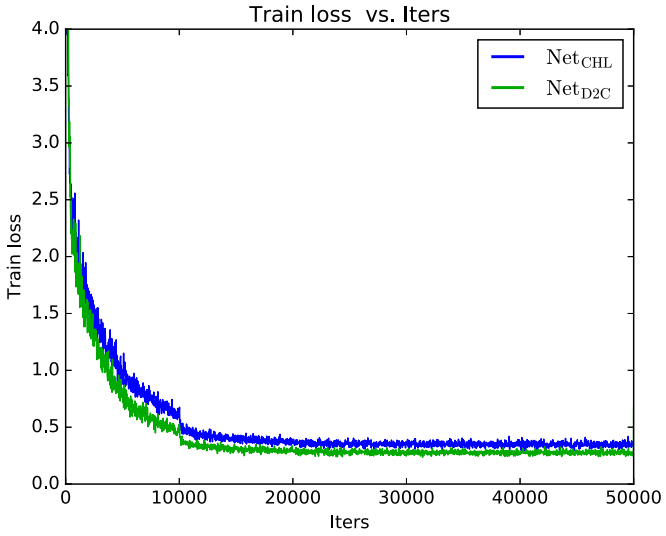

Fig. 9. Loss observed during training on the Morph II dataset.

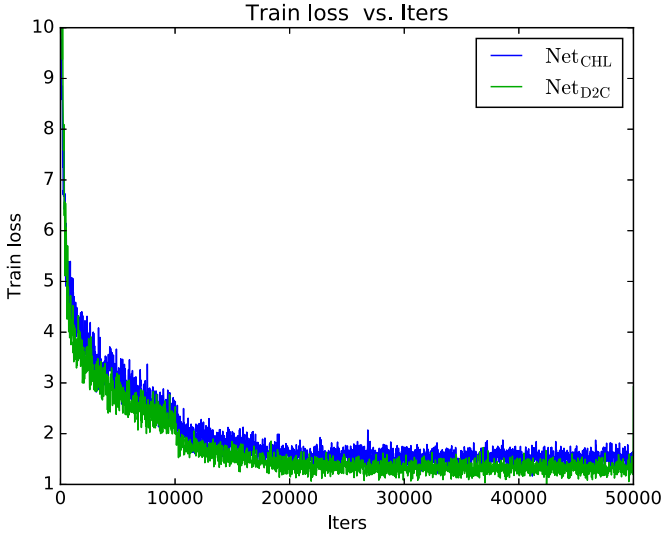

Fig. 10. Loss observed during training on the WebFace dataset.

4.3.1. Sensitiveness of the hyper-parameter β

As show in Eq. (9), the hyper parameter β is used to balance the auxiliary ranking loss and the main age estimation loss. It is known that adjusting hyper-parameters for hybrid loss terms are critical for heterogeneous learning goals. Based on this consideration, we conduct experiments to investigate the sensitiveness of β on the age estimation results. Specifically, we vary β from 0 to 1 to learn different models, the validation MAE of these models on both datasets are shown in Figs. 12 and 13. It is very clear that the models using the comparative ranking layer outperform the models without using it (in this case $\beta = 0$). We can also observe that the validation performance of our D2C model remains largely stable across a wide range of β. These experimental results and analyses demonstrate that our D2C age estimation model is insensitive to the value of β.

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

Tables 10, 11 compare our D2C age estimation model Net_{D2C} with several recently published methods on the Morph II and the WebFace datasets. Our D2C model outperforms all the other state-of-the-art

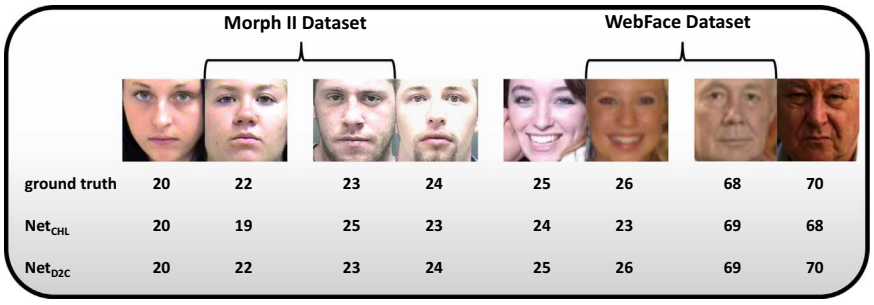

Fig. 11. Some age estimation results made by Net_{CHL} and Net_{D2C}. Net_{D2C} corrects some mistakes made by Net_{CHL} and makes more accurate predictions.

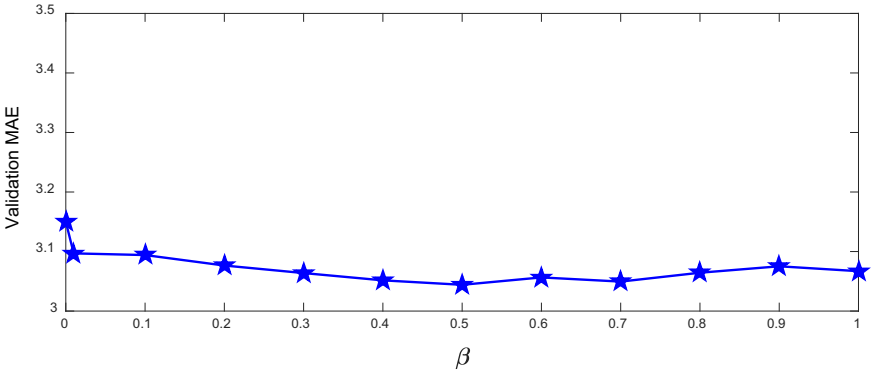

Fig. 12. The validation MAE of Net_{D2C} on the Morph II dataset with different β.

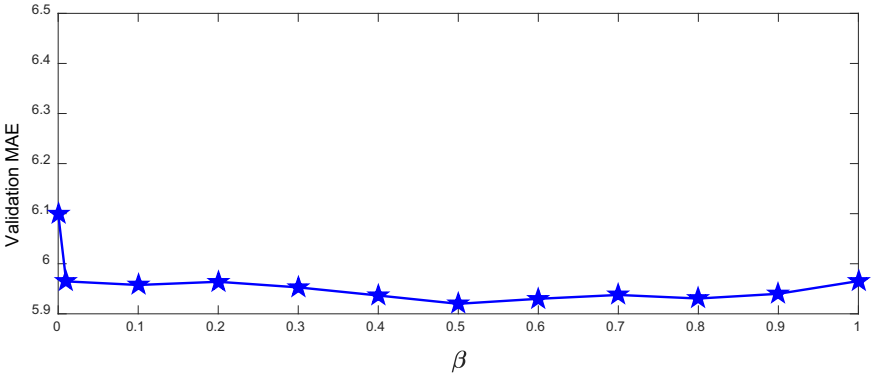

Fig. 13. The validation MAE of Net_{D2C} on the WebFace dataset with different β.

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

Methods	Age MAE
BIF [13]	5.09
KPLS [16]	4.18
KCCA [53]	3.98
Ridge [51]	4.80
Tree-a-CNN [33]	3.61
Multi-scale-CNN [32]	3.63
Our D2C model Net_{D2C}	**3.06**

methods on both datasets by a large margin. On the Morph II dataset, our D2C model reduces the age estimation MAE by 0.55 years which is a 15.2% relative improvement. To the best of our knowledge, this is the first time an MAE value near to 3 years has been obtained on this dataset.

Table 11
Comparison with the state-of-the-art methods on the WebFace dataset.

Methods	Age MAE
BIF [13]	10.65
RF [54]	9.38
Ridge [51]	9.75
Tree-a-CNN [33]	7.72
Our D2C model Net_{D2C}	**6.04**

On the WebFace dataset, our D2C model improves on the previous best results by 1.68 years which is about a 21.8% relative improvement. Since the WebFace dataset is compiled from faces in the wild, there have been fewer experiments on this challenging dataset. We compared the results from our model with all the published results that we could find for this dataset, including the latest in [33]. Our 21.8% relative improvement is significantly better than the state-of-the-art methods,

considering the difficulty of this dataset. The performance of our D2C model indicates the effectiveness of our proposed cumulative hidden layer and comparative ranking layer for human age estimation.

5. Conclusion

In this paper, we have proposed a deep cumulatively and comparatively (D2C) supervised age estimation model. To combat the sample imbalance problem we proposed a novel cumulative hidden layer which is supervised by a point-wise cumulative signal. By incorporating this cumulative hidden layer, our model can not only learn from one face itself but also from faces with nearby ages. This alleviates the sample imbalance problem effectively. In order to learn more discriminative aging features, we further propose a novel comparative ranking layer which is supervised by a pair-wise comparative signal. This comparative ranking layer facilitates aging feature learning and further improves the age estimation performance. Our D2C age estimation model is evaluated on two of the largest benchmark datasets and outperforms the state-of-the-art by a large margin. The network used in this work is relatively shallow compared with modern very deep architectures. Future work will investigates the use of deeper networks to improve estimates of age.

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, 61672519, U1636218 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] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, IEEE Trans. Syst. Man, Cyber. B, Cyber. 34 (1) (2004) 621–628.
[2] 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.
[3] Z. Song, B. Ni, D. Guo, T. Sim, S. Yan, Learning universal multi-view age estimator using video context, in: Proceedings of IEEE International Conference Comput. Vis., 2011, pp. 241–248.
[4] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, H. Dai, Learning from facial aging patterns for automatic age estimation, in: Proceedings of ACM International Conference Multimed., 2006, pp. 307–316.
[5] Z. Yang, H. Ai, Demographic classification with local binary patterns, in: Proceedings of IEEE International Conference Comput. Bio., 2007, pp. 464–473.
[6] F. Gao, H. Ai, Face age classification on consumer images with gabor feature and fuzzy LDA method, in: Proceedings of IEEE International Conference Comput. Bio., 2009, pp. 132–141.
[7] Y.H. Kwon, N. da Vitoria Lobo, Age classification from facial images, Comput. Vis. Image Understand. 74 (1) (1999) 1–21.
[8] 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.
[9] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2007) 578–584.
[10] C.-C. Wang, Y.-C. Su, C.-T. Hsu, C.-W. Lin, H.M. Liao, Bayesian age estimation on face images, in: Proceedings of IEEE International Conference Multimed. Expo., 2009, pp. 282–285.
[11] B. Ni, Z. Song, S. Yan, Web image mining towards universal age estimator, in: Proceedings of ACM International Conference Multimed., 2009, pp. 85–94.
[12] G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Trans. Image Process 17 (7) (2008) 1178–1188.
[13] G. Guo, G. Mu, Y. Fu, T. Huang, Human age estimation using bio-inspired features, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2009, pp. 112–119.
[14] A. Gunay, V. Nabiyev, Automatic age classification with LBP, in: Proceedings of International Symp. Comput. Info. Sci., 2014, pp. 1–4.
[15] W. Gao, H. Ai, A probabilistic boosting tree for face gender classification on consumer images, in: Proceedings of IEEE International Conference Comput. Bio., 2009, pp. 169–178.
[16] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2011, pp. 657–664.
[17] S. Yan, H. Wang, T. Huang, X. Tang, Ranking with uncertain labels, in: Proceedings of IEEE Conference Multimed. Expo., 2007, pp. 96–99.
[18] S. Yan, H. Wang, X. Tang, T. Huang, Learning autostructured regressor from uncertain nonnegative labels, in: Proceedings of IEEE International Conference Comput. Vis., 2007, pp. 1–8.
[19] 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.
[20] X. Geng, Q. Wang, Y. Xia, Facial age estimation by adaptive label distribution learning, in: Proceedings of International Conference Pattern Recognit., 2014, pp. 4465–4470.
[21] K. Chen, J.-K. Kämäräinen, Z. Zhang, Facial age estimation using robust label distribution, in: Proceedings of ACM International Conference Multimed., 2016, pp. 77–81.
[22] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1734–1748.
[23] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[24] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of Adv. Neural Info. Process. Systems, 2012, pp. 1097–1105.
[25] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of IEEE International Conference Comput. Vis., 2015, pp. 1026–1034.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.
[27] G. Huang, H. Lee, E. Learned, Learning hierarchical representations for face verification with convolutional deep belief networks, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2012, pp. 2518–2525.
[28] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of Adv. Neural Info. Process. Systems., 2014, pp. 1988–1996.
[29] W. Ouyang, X. Wang, Joint deep learning for pedestrian detection, in: Proceedings of IEEE International Conference Comput. Vis., 2013, pp. 2056–2063.
[30] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: Proceedings of Eur. Conference Comput. Vis., 2014, pp. 94–108.
[31] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit. Workshops, 2015, pp. 34–42.
[32] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Proceedings of Asian Conference Comput. Vis., 2014, pp. 144–158.
[33] 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 IEEE International Conference Comput. Vis. Pattern Recognit., 2015, pp. 222–230.
[34] J. Wang, S. Kumar, S.-F. Chang, Sequential projection learning for hashing with compact codes, in: Proceedings of ACM International Conference Mach. Learn., 2010, pp. 1127–1134.
[35] M. Norouzi, D.M. Blei, Minimal loss hashing for compact binary codes, in: Proceedings of ACM International Conference Mach. Learn., 2011, pp. 353–360.
[36] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, S.-F. Chang, Supervised hashing with kernels, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2012, pp. 2074–2081.
[37] G. Lin, C. Shen, D. Suter, A. van den Hengel, A general two-step approach to learning-based hashing, in: Proceedings of IEEE International Conference Comput. Vis., 2013, pp. 2552–2559.
[38] G. Lin, C. Shen, Q. Shi, A. van den Hengel, D. Suter, Fast supervised hashing with decision trees for high-dimensional data, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2014, pp. 1963–1970.
[39] R. Xia, Y. Pan, H. Lai, C. Liu, S. Yan, Supervised hashing for image retrieval via image representation learning, in: Proceedings of AAAI Conference Artificial Intel., 2014, pp. 2156–2162.
[40] K. Chen, S. Gong, T. Xiang, C. Loy, Cumulative attribute space for age and crowd density estimation, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2013, pp. 2467–2474.
[41] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of IEEE International Conference Comput. Vis. Pattern Recognit., 2011, pp. 3337–3344.
[42] Y. Fu, T.M. Hospedales, T. Xiang, S. Gong, Attribute learning for understanding unstructured social activity, in: Proceedings of Eur. Conference Comput. Vis., 2012, pp. 530–543.
[43] R. Layne, T.M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes., in: Proceedings of British. Mach. Vis. Conference, 2012, pp. 8–19.
[44] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: Proceedings of ACM International Conference Mach. Learn., 2005, pp. 89–96.
[45] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of ACM International Conference Mach. Learn., 2009, pp. 41–48.
[46] C. Zhang, Z. Zhang, Improving multiview face detection with multi-task deep convolutional neural networks, in: WACV, 2014, pp. 1036–1041.
[47] 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 IEEE International Conference Comput. Vis. Pattern Recognit., 2015, pp. 1–9.
[48] 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.
[49] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensic Secur. 9 (12) (2014) 2170–2179.

[50] K. Ricanek, T. Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in: Proceedings of IEEE International Conference Face Gesture, 2006, pp. 341–345.
[51] Z. Song, Visual Image Recognition System With Object-level Image Representation, (Ph.D. thesis), National University of Singapore, 2012.
[52] 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 ACM International Conference Multimed., 2014, pp. 675–678.
[53] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: CCA vs. PLS, in: Proceedings of IEEE International Conference Face Gesture, 2013, pp. 1–6.
[54] S. Li, S. Shan, X. Chen, Relative forest for attribute prediction, in: Proceedings of Asian Conference Comput. Vis., 2013, pp. 316–327.

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

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 science from Tsinghua University, Beijing, China, in 2012. He is currently an associate professor 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 50 papers on international journals and conferences. His current research interests mainly focus on computer vision problems related to faces and humans.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University in 1998. From 1998–2000, 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.

Stephen J. Maybank received the BA degree in mathematics from Kings College Cambridge in 1976, and the Ph.D. degree in computer science from Birkbeck College, University of London in 1988. He is currently a professor in the Department of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, etc. He is a fellow of the IEEE.