Using Ranking-CNN for Age Estimation

Shixing Chen[1] Caojin Zhang[2] Ming Dong[1] Jialiang Le[3] Mike Rao[3]
[1]Department of Computer Science [2]Department of Mathematics [3]Research & Innovation Center
Wayne State University Wayne State University Ford Motor Company
{schen, czhang, mdong}@wayne.edu {jle1, mrao}@ford.com

Abstract

Human age is considered an important biometric trait for human identification or search. Recent research shows that the aging features deeply learned from large-scale data lead to significant performance improvement on facial image-based age estimation. However, age-related ordinal information is totally ignored in these approaches. In this paper, we propose a novel Convolutional Neural Network (CNN)-based framework, ranking-CNN, for age estimation. Ranking-CNN contains a series of basic CNNs, each of which is trained with ordinal age labels. Then, their binary outputs are aggregated for the final age prediction. We theoretically obtain a much tighter error bound for ranking-based age estimation. Moreover, we rigorously prove that ranking-CNN is more likely to get smaller estimation errors when compared with multi-class classification approaches. Through extensive experiments, we show that statistically, ranking-CNN significantly outperforms other state-of-the-art age estimation models on benchmark datasets.

1. Introduction

One major issue in age estimation models is how to extract effective aging features from a facial image. In the past decade, many efforts have been devoted to aging feature representations. Simple geometry features and texture features were first adopted in [20]. Later on, Biologically Inspired Features (BIF) [15] were proposed and widely adopted in age estimation applications. More recently, Scattering Transform (ST) [2] was also proposed as an improvement over BIF by adding filtering routes. Usually, these features can be further enhanced through manifold learning, e.g., Orthogonal Locality Preserving Projection (OLPP) [14].

The other important component in an age estimation model is the estimator. Commonly, age estimation is characterized to be a classification or regression problem. Classification models include k Nearest Neighbors [13], Multilayer Perceptrons [21], and Support Vector Machines (SVM) [15]. For regression methods, quadratic regression

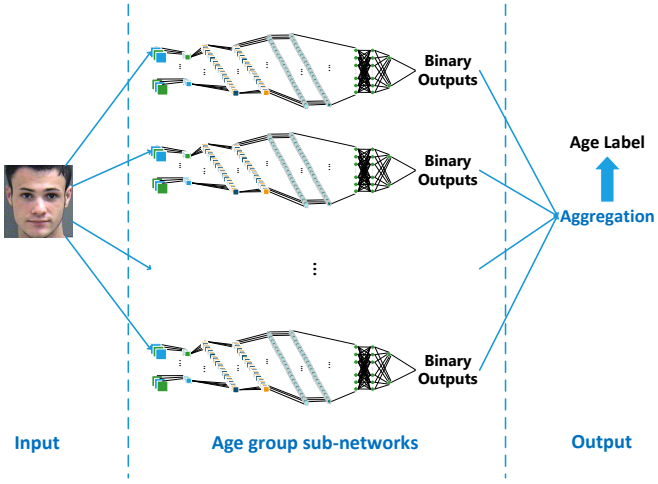

Figure 1. Ranking-CNN for facial image-based age estimation.

[14] and Support Vector Regression (SVR) [15] were considered in the literature. More recently, deep learning techniques such as Convolutional Neural Networks (CNN) have been applied to human age estimation to learn aging features directly from large-scale facial data [39]. Experimental results show that the deeply-learned aging patterns lead to significant performance improvement on benchmark datasets [37] as well as unconstrained photos [25]. However, multi-class classification completely ignores the ordinal information in age labels, and regression over-simplifies it to a linear model while human aging pattern is generally nonlinear. Recently, cost-sensitive ranking techniques have been introduced to age estimation [2].

In this paper, we propose a novel age ranking approach based on CNN. Specifically, we propose a ranking-CNN model that contains a series of basic CNNs, each of which has a sequence of convolutional layers, sub-sampling layers and fully connected layers. Basic CNNs are initialized with the weights of a pre-trained base CNN and fine-tuned with the ordinal age labels through supervised learning. Then, their binary outputs are aggregated to make the final age prediction. Fig. 1 shows an illustration of our model. Comparing with prior work where the same set of features was used for all age groups, in ranking-CNN, features are

learned independently in each age group to depict different aging patterns, which leads to significant performance gain. The major contribution of this work is summarized as follows:

- To the best of our knowledge, ranking-CNN is the first work that uses a deep ranking model for age estimation, in which binary ordinal age labels are used to train a series of basic CNNs, one for each age group. Each basic CNN in ranking-CNN can be trained using all the labeled data, leading to better performance of feature learning and also preventing overfitting.

- We provide a much tighter error bound for age ranking than that introduced in [2], which claimed that the final ranking error is bounded by the sum of errors generated by all the classifiers. We obtain the approximation for the final ranking error that is controlled by the maximum error produced among sub-problems. From a technical perspective, the tighter error bound provides several advantages for the training of ranking-CNN.

- We prove that ranking-CNN, by taking the ordinal relation between ages into consideration, is more likely to get smaller estimation errors when compared with multi-class classification approaches (i.e., CNNs using the softmax function). Moreover, through extensive experiments, we show that statistically, ranking-CNN significantly outperforms other state-of-the-art age estimation methods.

The rest of this paper is arranged as follows. In Section 2, we briefly review related work in age estimation and CNN. In Section 3, we introduce ranking-CNN for age estimation, establish its theoretical error bound, and compare it with softmax-based multi-class CNNs. In Section 4, we present our age estimation results on the benchmark datasets. Finally, we conclude in Section 5.

2. Related Work

2.1. Age Estimation

One of the earliest age estimation model can be traced back to [22], in which Active Appearance Model (AAM) [6] was employed to extract shape and appearance features from facial images. Later, in [10], the aging process was simulated using AAM for the same individual with a series of age-ascending facial images so that specific models associated with different people's aging processes can be constructed. Also, to interpret the long-term aging subspace of a person, Geng et al. [11] proposed AGing pattErn Subspace (AGES).

Since the available images for a specific person are typically very limited, many researchers focus on developing non-personalized approaches instead. Yang and Ai [38] adopted a real AdaBoost algorithm with Local Binary Patterns [1]. Li et al. [26] proposed a method based on ordinal discriminative feature learning. In [15], BIF features were shown to be effective for age estimation on various datasets. Meanwhile, manifold learning algorithms were incorporated to achieve better performance. In [14], Guo et al. proposed to use aging manifold with locally adjusted robust regressor.

More recently, CNN-based methods have been widely adopted for age estimation due to its superior performance over existing methods. Yi et al. [39] introduced a multi-task learning method with a relatively shallow CNN. Wang et al. [37] trained a deeper CNN for extracting features from different layers. Levi et al. validated CNN's performance on unconstrained facial images [25].

Instead of multi-class classification and regression methods, ranking techniques were introduced to the problem of age estimation. In [2], a cost-sensitive ordinal ranking framework was proposed with ST features. In [29], Niu et al. proposed to formulate age estimation as an ordinal regression problem with the use of multiple output CNN.

2.2. Convolutional Neural Networks

There are numerous kinds of CNN models developed in deep learning. The exact forms could vary, but the major components and computations are similar. CNN models derived from LeNet [24] consist of alternating convolutional and pooling layers followed by fully-connected layers with the input to successive layers being the feature maps from previous layers. Weights in layers are updated simultaneously for representative features and classification with a specific loss function through back propagation.

CNNs have been widely used on a variety of applications. In natural language processing, SENNA system has achieved state-of-the-art performance on various tasks [5]. In text classification, CNN architectures have been widely adopted and achieved superior outcomes [18]. In the computer vision field, great successes have been achieved in image classification [19], object detection [12], face recognition [34] and image segmentation [27].

Recently, with the implementation using GPUs [19, 17], CNN models with deep architectures have achieved breakthroughs on object recognition problems in large-scale image datasets, e.g., the ImageNet dataset [7]. To build more effective CNN models, several new components were introduced: activation unit such as rectified linear unit (ReLU) [28] helps to accelerate the convergence during training and has a positive influence on the performance [19]; regularizer like dropout prevents overfitting by setting some activation units to zero in a specific layer [33]; and batch normalization allows the use of much higher learning rates to make training faster and to improve performance [16].

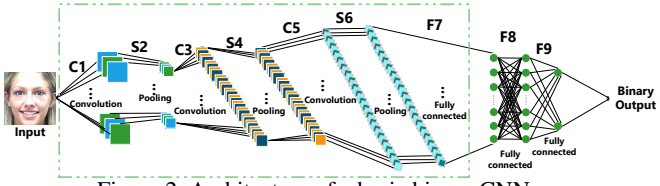

Figure 2. Architecture of a basic binary CNN

3. Ranking-CNN for Age Estimation

The training of ranking-CNN consists of two stages: pre-training with facial images and fine-tuning with age-labeled faces. First, a base network is pre-trained with unconstrained facial images [9] to learn a nonlinear transformation of the input samples that captures their main variation. From the base network, we then train a series of basic binary CNNs with ordinal age labels. Specifically, we categorize samples into two groups: with ordinal labels either higher or lower than a certain age, and then use them to train a corresponding binary CNN. The fully connected layers in the binary CNN first flatten the features obtained in the previous layers and then relate them to a binary prediction. The weights are updated through stochastic gradient descent by comparing the prediction with the given label. Finally, all the binary outputs are aggregated to make the final age prediction. In the following, we present our system in details.

3.1. Basic Binary CNNs

As shown in Fig. 2, a basic CNN has three convolutional and sub-sampling layers, and three fully connected layers. Specifically, C1 is the first convolutional layer with feature maps connected to a 5×5 neighboring area in the input. There are 96 filters applied to each of the 3 channels (RGB) of the input, followed by Rectified Linear Unit (ReLU) [28].

S2 is a sub-sampling layer with feature maps connected to corresponding feature maps in C1. In our case, we use max pooling on 3×3 regions with the stride of 2 to emphasize the most responsive points in the feature maps. S2 is followed by local response normalization (LRN) that can aid generalization [19] . C3 works in a similar way as C1 with 256 filters in 96 channels and 5×5 filter size followed by ReLU. Layer S4 functions similarly as S2, and is followed by LRN. Then, C5 is the third convolutional layer with 384 filters in 256 channels and smaller filter size 3×3, followed by the third max pooling layer S6.

F7 is the first fully connected layer in which the feature maps are flattened into a feature vector. There are 512 neurons in F7 followed by ReLU and a dropout layer [33]. F8 is the second fully connected layer with 512 neurons that receives the output from F7 followed by ReLU and another dropout layer. F9 is the third fully connected layer and computes the probability that an input x (i.e., output after F8) belongs to class i using the logistic function. The optimal model parameters of a network are typically learned through minimizing a loss function. We use the negative log-likelihood as the loss function, and minimize it using stochastic gradient descent.

3.2. Ranking-CNN

Assume that x_i is the feature vector representing the ith sample and $y_i \in \{1,...,K\}$ is the corresponding ordinal label. To train the k-th binary CNN, the entire dataset D is split into two subsets, with age values higher or lower (or equal to) than k,

$$D_k^+ = \{(x_i,+1)|y_i > k\}, \quad D_k^- = \{(x_i,-1)|y_i \leq k\}. \quad (1)$$

Based on different splitting of D, $K-1$ basic networks can be trained from the base one. Note that in our model, each network is trained using the entire dataset, typically resulting in better ranking performance and also preventing overfitting. Given an unknown input x_i, we first use the basic networks to make a set of binary decisions and then aggregate them to make the final age prediction $r(x_i)$:

$$r(x_i) = 1 + \sum_{k=1}^{K-1} [f_k(x_i) > 0]. \quad (2)$$

where $f_k(x_i)$ is the output of the basic network and $[\cdot]$ denotes the truth-test operator, which is 1 if the inner condition is true, and 0 otherwise. It can be shown that the final ranking error is bounded by the maximum of the binary ranking errors. That is, the ranking-CNN results can be improved by optimizing the basic networks. We mathematically prove this in Section 3.2.1 followed by the theoretical comparison between ranking and softmax-based multi-class classification in Section 3.2.3.

3.2.1 Error Bound

In ranking-CNN, we divide an age ranking estimation problem, ranging from $1, \cdots, K$, into a series of binary classification sub-problems ($K-1$ classifiers). By aggregating the results of each sub-problem, we then obtain an estimated age $r(x)$. To assure a better overall performance of the model, a key issue is whether the ranking error can be reduced if we improve the accuracy of the binary classifiers. We rigorously address this issue with formal mathematical proof in this section.

Here, we provide a much tighter error bound for age ranking than that introduced in [2], which claims that the final ranking error is bounded by the sum of errors generated by all the classifiers. We adopt the idea in [2] that divides the errors of sub-problems into two groups: overestimated and underestimated errors. However, instead of simply aggregating errors, we rearrange them in an increasing order and go deep into the analysis of the underlying differences between any adjacent sub-classifier errors inside each group. By the accumulation of those differences, we theoretically obtain an approximation for the final ranking

error, which is controlled by the maximum error produced among sub-problems.

We denote $E^+ = \sum_{k=1}^{K-1} \gamma_k^+$ as the number of misclassifications $f_k(x) > 0$ when the actual value $y < k$, $k = 1, \cdots, K-1$. Similarly, we denote $E^- = \sum_{k=1}^{K-1} \gamma_k^-$ as the opposite case, where $\gamma_k^+ = [f_k(x) > 0][y \leq k]$ and $\gamma_k^- = [f_k(x) < 0][y > k]$, and $[\cdot]$ is an indicator function taking value of 1 when the condition in $[\cdot]$ holds, 0 otherwise.

For any observation (x, y), we define the cost function (error) for each classifier as:

$$e_k(x) = \begin{cases} e_k^+ = (k - y + 1)\gamma_k^+ & y \leq k \\ e_k^- = (y - k)\gamma_k^- & y > k. \end{cases} \tag{3}$$

Thus, we have a theorem for the error bound of final ranking error:

Theorem 1 *For any observation (x, y), in which $y > 0$ is the actual label (integer), then the following inequality holds:*

$$|r(x) - y| \leq \max_k e_k(x), \tag{4}$$

where $r(x)$ is the estimated rank of age, $k = 1, \cdots, K-1$. That is, we can diminish the final ranking error by minimizing the greatest binary error.

Proof

Denote $e_k(x)$ in (3) as e_k for simplicity. We split the proof into two parts. Firstly, we show $|E^+ - E^-| = |r(x) - y|$. Secondly, we demonstrate $\max_k e_k \geq \max\{E^+, E^-\}$. By $|E^+ - E^-| < \max\{E^+, E^-\}$ for E^+ and E^- nonnegative, the inequality (4) follows.

Firstly, we begin by definition:

$$\begin{aligned} & r(x) = 1 + \textstyle\sum_{k=1}^{K-1} [f_k(x)] \\ & = 1 + \textstyle\sum_{k=1}^{K-1} ([f_k(x) > 0][y \leq k] + [f_k(x) > 0][y > k]) \\ & = 1 + E^+ + \textstyle\sum_{k=1}^{K-1} [f_k(x) > 0][y > k]. \end{aligned} \tag{5}$$

Subtracting $(E^+ - E^-)$ on both sides, we get

$$\begin{aligned} & r(x) - (E^+ - E^-) \\ & = 1 + \textstyle\sum_{k=1}^{K-1} [f_k(x) > 0][y > k] + \sum_{k=1}^{K-1} [f_k(x) \leq 0][y > k] \\ & = 1 + \textstyle\sum_{k=1}^{K-1} ([f_k(x) > 0] + [f_k(x) \leq 0])[y > k] \\ & = 1 + \textstyle\sum_{k=1}^{K-1} [y > k] \\ & = y. \end{aligned} \tag{6}$$

Thus $|r(x) - y| = |E^+ - E^-|$ holds.

Secondly, we extract all $e_k^+ > 0$ and rearrange them in an increasing order denoted as a set $\{e_{(j)}^+, j = 1, 2, \cdots, E^+\}$. Similarly, we do the same operation on e_k^- and have the set $\{e_{(j)}^-, j = 1, 2, \cdots, E^-\}$, where for any random variable ξ, $\xi_{(\cdot)}$ denotes the order statistics.

Since y is an integer, by (3), $e_{(1)}^+ \geq 1$ and $|e_{(j)}^+ - e_{(j-1)}^+| \geq 1$ for any $j \in \{1, 2, \cdots, E^+\}$. We observe that:

$$e_{(E^+)}^+ \geq e_{(1)}^+ + |e_{(2)}^+ - e_{(1)}^+| + \cdots + |e_{(E^+)}^+ - e_{(E^+-1)}^+|. \tag{7}$$

It follows $e_{(E^+)}^+ \geq E^+$. Similarly, we can show $e_{(E^-)}^- \geq E^-$. Then, $\max_k e_k = \max\{e_{(E^+)}^+, e_{(E^-)}^-\} \geq \max\{E^+, E^-\}$, which completes the proof.

3.2.2 Technical Contribution of the New Error Bound

Ranking-CNN can be seen as an ensemble of CNNs, fused with aggregation. By showing that the final ranking error is bounded by the maximum error of the binary rankers, we make significant technical contribution in the following aspects:

1. Theoretically, it was mentioned in both [2] and [29] that the inconsistency issue of the binary outputs could not be resolved because that would make the training process significantly complicated. The aggregation was just carried out without explicit understanding of the inconsistency. With the tightened error bound, we can confidently demonstrate that the inconsistency doesn't actually matter because as long as the maximum binary error is decreased, the error produced by inconsistent labels can be ignored. It would neither influence the final estimation error nor complicate the training procedure.

2. Methodologically, the tightened bound provides extremely helpful guidance for the training of ranking-CNN. The training of an ensemble of deep learning models is typically very time consuming, especially when the number of sub-models is large. Based on our results, it is technically sound to focus on the sub-models with the largest errors. This training strategy will lead to more efficient training to achieve the desired performance gain. The training strategy can also be extended to ensemble learning with other decision fusion methods.

3. Mathematically, based on the new error bound, we can rigorously derive the expectation of prediction error of ranking-CNN and prove that ranking-CNN outperforms other softmax-based deep learning models. The detailed proof is given in the next section.

3.2.3 Ranking v.s Softmax

In this section, we focus on demonstrating that ranking-CNN outperforms softmax method because it is more likely to get smaller prediction error $|r(x) - y|$. The reason is that softmax failed to take the ordinal relation between ages into consideration. Thus, instead of a softmax classifier, ranking method is preferred for age estimation.

A basic CNN in ranking-CNN differs from the softmax multi-class classification approach in the output layer. Suppose after fully-connected layer, we get $z_1, z_2, \cdots z_K$ from K

networks. Denote $\hat{y}$ as the estimated age label, and $a_i = e^{z_i}$ where $e^{(\cdot)}$ is the natural exponential function. For softmax, the posterior probability of each class is given by:

$$P(\hat{y} \in i|x) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} = \frac{a_i}{\sum_{k=1}^{K} a_k}, \tag{8}$$

for $i = 1, \cdots, K$. Then, the expected error given the label of the observation (x, y) is

$$E(|r(x) - y||y) = \sum_{i=1}^{K} |i - y| P(\hat{y} = i|x). \tag{9}$$

For ranking-CNN, we use $K-1$ classifiers to determine ordinal relation between adjacent ages. The posterior probability for a prediction of age greater than a specific age i is given by:

$$P(f_i(x) > 0|x) = \frac{e^{z_{i+1}}}{e^{z_i} + e^{z_{i+1}}} = \frac{a_{i+1}}{a_i + a_{i+1}}. \tag{10}$$

Then, the expected error for a given sample is

$$E(|r(x) - y||y) = \sum_{i=1}^{K} |i - y| P(\hat{y} = i|x). \tag{11}$$

We present a theorem for a three ordinal class problem. In the theorem, we use a, b, c to represent a_1, a_2, a_3 respectively for better clarity.

Theorem 2 *Suppose we have classes* 1, 2 *and* 3 *with* $a, b, c > 0$ *respectively. There exists an ordinal relation:* $1 < 2 < 3$. *Denote the rank obtained by ranking-CNN as* $r_1(x)$ *and the result by softmax as* $r_2(x)$. *Then*

$$E(|r_1(x) - y|) < E(|r_2(x) - y|). \tag{12}$$

Proof. Given a sample with label 1, the expected errors for ranking-CNN and softmax are:

$$\begin{aligned} &E(|r_1(x) - y||y = 1) \\ &= 2P(f_1(x) > 0, f_2(x) > 0|x) + P(f_1(x) > 0, f_2(x) < 0|x) \\ &+ P(f_1(x) < 0, f_2(x) > 0|x) \\ &= 2\frac{b}{a+b}\frac{c}{b+c} + \frac{b}{a+b}\frac{b}{b+c} + \frac{a}{a+b}\frac{c}{b+c} \\ &= \frac{2bc + b^2 + ac}{(a+b)(b+c)}, \end{aligned} \tag{13}$$

and

$$\begin{aligned} E(|r_2(x) - y||y = 1) = &\quad 2P(r_2(x) = 2|x) + P(r_2(x) = 3|x) \\ = &\quad \frac{2c + b}{a + b + c}, \end{aligned} \tag{14}$$

respectively.

Similarly, given $y = 2$,

$$\begin{aligned} &E(|r_1(x) - y||y = 2) \\ &= P(f_1(x) > 0, f_2(x) > 0|x) + P(f_1(x) < 0, f_2(x) < 0|x) \\ &= \frac{ab + bc}{(a+b)(b+c)}, \end{aligned} \tag{15}$$

$$\begin{aligned} E(|r_2(x) - y||y = 2) = &\quad P(r_2(x) = 1|x) + P(r_2(x) = 3|x) \\ = &\quad \frac{a + c}{a + b + c}. \end{aligned} \tag{16}$$

Given $y = 3$,

$$\begin{aligned} &E(|r_1(x) - y||y = 3) \\ &= 2P(f_1(x) < 0, f_2(x) < 0|x) + P(f_1(x) > 0, f_2(x) < 0|x) \\ &+ P(f_1(x) < 0, f_2(x) > 0|x) \\ &= \frac{2ab + b^2 + ac}{(a+b)(b+c)}, \end{aligned} \tag{17}$$

$$\begin{aligned} E(|r_2(x) - y||y = 3) = &\quad 2P(r_2(x) = 1|x) + P(r_2(x) = 2|x) \\ = &\quad \frac{2a + b}{a + b + c}. \end{aligned} \tag{18}$$

Thus, for ranking-CNN, it follows

$$\begin{aligned} E(|r_1(x) - y|) = &\sum_{i=1}^{3} E(|r_1(x) - i||y = i) \\ &2 + \frac{ab + bc}{(a+b)(b+c)}. \end{aligned} \tag{19}$$

Similarly, for softmax,

$$E(|r_2 - y|) = \sum_{i=1}^{3} E(|r_2(x) - i||y = i) = 2 + \frac{a + c}{a + b + c}. \tag{20}$$

Since

$$\begin{aligned} &\frac{a + c}{a + b + c} - \frac{ab + bc}{(a+b)(b+c)} \\ &= \frac{a^2 c + c^2 a}{(a+b)(b+c)(a+b+c)} > 0, \end{aligned} \tag{21}$$

then we conclude

$$E(|r_1(x) - y|) < E(|r_2(x) - y|). \tag{22}$$

Furthermore, the cases for $K = 4, 5, \cdots$ could be shown in a similar way by induction. However, when the number of class K increases, the analytic expression of the distribution for each class $i = 1, \cdots, K$, becomes

$$P(\hat{y} = i|y) = \sum_{A \in \mathscr{F}_i} \prod_{j \in A} p_j \prod_{j \in A^c} (1 - p_j), \tag{23}$$

satisfying a Poisson-Binomial distribution, where $p_j = \frac{a_j}{a_{j-1} + a_j}$, $\mathscr{F}_i$ is the subset of i integers that could be selected from $\{1, 2, \cdots, K\}$ and A^c is the complement of A. Notice that $\mathscr{F}_i$ represents C_2^K possible cases. Then, to compute the expected value becomes dreadful since listing all the probability out as we did in theorem 2 seems impractical. Though Le Cam et al. [23] gave an approximation of Poisson-Binomial by a Poisson distribution, the computation for

$$E(|r_1(x) - y|) = \sum_{y=1}^{K} \sum_{r=1}^{K} |r - y| P(\hat{y} = r) \tag{24}$$

is still unrealistic. So, we generalize with the help of learning theory.

Theorem 3 *Suppose the VC dimension of each basic CNN classifier's hypothesis spaces $\mathcal{H}_i^2$ is d, the sample size for training is m. Then for any $\delta \in [0,1]$, with probability at least $1-\delta$, the expected error of the ranking-CNN is upper bounded as follows:*

$$E_D|r(x)-y| \leq \max_k \hat{e}_k(x) + 2\sqrt{\frac{d\log(2m)+\log(\frac{2}{\delta})}{m}} \quad (25)$$

where $\hat{e}_k(x)$ denotes the empirical values for $E_D e_k(x)$.

Proof. Taking expectation on both sides of Eq. (4), we get

$$E_D|r(x)-y| \leq E_D \max_k e_k(x) \quad (26)$$

Using Vapnik-Chervonenkis theory [35], the desired result follows.

Remark 4 Notice the expected error for ranking-CNN is bounded by the maximum training error produced by its basic CNNs with binary output, adding a term associated with VC dimension. Since the VC dimension d of a softmax output CNN is greater than that of a basic CNN presented in Fig. 2 [32], if the weights of previous layers are fixed, it results in a greater second term on right hand side of Eq. (25) for a CNN with softmax output layer. It follows that given the same training samples, ranking-CNN is more likely to attain a smaller error by minimizing the training errors (the first term in Eq. (25)) than the one with a softmax output.

The error bound in Eq. (25) provides a solid support for our framework. We will further verify this conclusion in the sense of statistical significance by t-test later in the experiment section.

3.3. Age Estimation

When humans predict a person's age, it is generally easier to determine if a person is elder than a specific age than directly giving an exact age. With ranking-CNN, it provides a framework for simultaneous feature learning and age ranking based on facial images. The rationale of using ranking-CNN for age estimation is that the age labels are naturally ordinal, and ranking-CNN can keep the relative ordinal relationship among different age groups.

First, we pre-train a base network with 26,580 image samples from the unfiltered faces dataset [9]. The age group labels for these images are used in training as surrogate labels [8]. Then, we fine-tune our ranking-CNN model on the most commonly used age estimation benchmark dataset: MORPH Album 2 [30]. MORPH contains 55,134 facial images with the age range from 16 to 77. Following the settings used in some recent work on age estimation [29, 37, 4, 3], we randomly select 54,362 samples in the age range between 16 and 66 from MORPH dataset. The age and gender information of the selected samples is shown in Table 1. Note that these images are not used in the pre-training stage. All the selected samples are then divided into two sets: 80% of the samples are used for basic networks training and the rest 20% samples for testing. There is no overlapping between the training and testing sets, and we use 5-fold cross-validation to evaluate the performance during experiments.

Table 1. The age and gender information of the 54,362 samples randomly selected from MORPH Album 2.

	<20	20-29	30-39	40-49	>50	Total
Male	6543	13849	12322	9905	3321	45940
Female	829	2291	2886	1975	441	8422
Total	7372	16140	15208	11880	3762	54362

We adopt a general pre-processing procedure for face detection and alignment before feeding the raw data to the networks. Specifically, given an input color image, we first perform face detection using Harr-based cascade classifiers [36]. Then, face alignment is conducted based on the locations of eyes. Finally, the image is resized to a standard size of $256\times256\times3$ for network training and age estimation.

4. Experiments

In this section, we demonstrate the performance of ranking-CNN through extensive experiments. We implemented the architecture for ranking-CNN in the GPU mode with Caffe [17]. For the $3+3$ architecture of a basic CNN shown in Fig. 2, it is derived from a simplified version of the ImageNet CNN [19] with fewer layers for higher efficiency [25]. The network is initialized with random weights following Gaussian distribution, the mean is 0, and standard deviation is 0.01.

For our hardware settings, we use a single GTX 980 graphics card (including 2,048 CUDA cores), i7-4790K CPU, 32GB RAM, and 2TB hard disk drive. The training time for the base CNN with the selected $3+3$ architecture is around 6 hours. Fine-tuning takes about 20 to 30 minutes for each basic CNN. Totally, it takes about 30 hours to pre-train the base CNN and fine-tune 50 basic CNNs.

4.1. Evaluation Metrics

For multiple age estimation, we compared the features learned by ranking-CNN with the ones obtained through BIF+OLPP [15], ST[2], and multi-class CNN. BIF features are implemented with Gabor filters in 8 orientations and 8 scales and followed by max-pooling. In addition, OLPP is employed to learn the age manifold based on BIF features, in which the top 1,000 eigenvectors are used. In ST, the Gabor coefficients are scattered into 417 routes in two convolutional layers and pooled with Gaussian smoothing. Multi-class CNN is commonly used for age estimation [25, 39],

Table 2. Comparison of MAE among different combinations of features and estimators. The lowest MAE is highlighted in **bold**. A dash in the table means that the selected feature is not applicable to the selected estimator.

		Engineered Features		Learned Features	
		BIF+OLPP	ST	CNN Feature	Ranking-CNN Feature
Classification	SVM	4.99	5.15	3.95	-
Model	Multi-class CNN	-	-	3.65	-
Ranking	Ranking-SVM	5.03	4.88	-	3.63
Model	ranking-CNN	-	-	-	**2.96**

but it completely ignores the ordinal information in age labels. Its structure is similar to a basic CNN (three convolutional and pooling layers and three fully connected layers) with the exception that the last fully-connected layer contains multiple outputs corresponding to the number of ages to be classified instead of the binary ones. As for the age estimators, SVM is selected for comparison due to its proved performance [15]. In ranking-based approach (Ranking-SVM), following [2], SVM is used as the binary classifier for each age label and the results are aggregated to give the final output.

The comparison and evaluation of different methods in our experiments are reported in terms of accuracy of each binary ranker as well as two widely adopted performance measures [29, 2]: Mean Absolute Error (MAE) and Cumulative Score (CS). MAE computes the absolute costs between the exact and the predicted ages (the lower the better): $MAE = \sum_{i=1}^{M} e_i/M$, where $e_i = |\hat{l}_i - l_i|$ is the absolute cost of misclassifying true label l_i to $\hat{l}_i$, and M is the total amount of testing samples. CS indicates the percentage of data correctly classified in the range of $(l_i - L, l_i + L)$, a neighbor range of the exact age l_i (the larger the better): $CS(L) = \sum_{i=1}^{M}[e_i \leq L]/M$, where $[\cdot]$ is the truth-test operator and L is the parameter representing the tolerance range.

Also, we used paired t-test to demonstrate the statistical significance of our empirical comparison. We employ paired t-test to determine if ranking-CNN significantly outperforms other methods. A two-sample t-statistic with unknown but equal variance is computed.

4.2. Age Estimation Results

In this section, we consider the age estimation problem in the range between 16 and 66 years old and compare ranking-CNN with other state-of-the-art feature extractors and age estimators. As there are 51 age groups in this age range, 50 binary rankers are needed for ranking approaches (i.e., ranking-CNN and ranking-SVM). In our experiments, 43,490 samples (80% of all the randomly selected samples) with binary labels are selected to train each basic network or SVM in ranking-CNN and ranking-SVM, respectively. The exactly same set of samples with multi-class labels are used to train multi-class CNN and SVM, respectively. The rest 10,872 samples were used for testing results. All experiments are carried out with 5-fold cross-validation.

Basically, we have three sets of features: engineered features (i.e., BIF+OLPP and ST), learned classification features (Multi-class CNN) and learned ranking features (ranking-CNN), and two sets of age estimators: classification methods (i.e., SVM and Multi-class CNN) and ranking methods (ranking-CNN and ranking-SVM). We report MAE of all possible combinations of feature extractors and age estimators (eight in total) in Table 2. A dash in the table means that the selected feature set is not applicable to the selected estimator.

As shown in Table 2, ranking-CNN with its features achieves the lowest MAE of 2.96 in all the combinations. Ranking-CNN features with Ranking-SVM achieves the second best MAE result, and this validates the effectiveness and generality of ranking-CNN features. In comparison, the lowest MAE achieved by the learned classification features is 3.65. Note the multi-class CNN represents the commonly used CNN-based age estimation methods [25, 39]. Our experimental results strongly support the theoretical results (ranking v.s. softmax) we presented in Section 3.2.3. Another fact we can see is that the performance of CNN-based features gets weakened when combined with SVM-based estimators. The lowest MAE achieved by engineered features is 4.88 by ST+ranking-SVM. Notice that ST works better with ranking-SVM, and BIF+OLPP works better with SVM. This could be caused by the fact that in the literature specific features were manually selected for certain estimators to achieve the best performance.

Table 3. Comparison with MR-CNN, OR-CNN and DEX on the MORPH dataset. The lowest MAE is highlighted in **bold**.

	Ranking-CNN	MR-CNN	OR-CNN	DEX
MAE	**2.96**	3.27	3.34	3.25

In Table 3, we compare ranking-CNN with the most recent age estimation models, i.e., Ordinal Regression with CNN (OR-CNN), Metric Regression with CNN (MR-CNN) [29] and Deep EXpectation (DEX) [31]. Since the experiments are all carried out on MORPH dataset and we followed the settings in [29] for data partition, we can directly compare the MAE of Ranking-CNN with the ones obtained by MR-CNN, OR-CNN and DEX. Clearly, ranking-CNN outperforms all MR-CNN, OR-CNN and DEX, and significantly improves the performance of age estimation.

The comparison in terms of CS of the eight combinations of features and estimators are given in Fig. 3. Clearly, ranking-CNN outperforms all others across the entire range of L (age error tolerance range) from 0 to 10. Specifically,

Table 4. T test outcomes of all eight combinations of features and estimators. Numbers #1 to #8 correspond to eight compared models in the sequence of: RANKING-CNN, RANKING-CNN FEATURE+RANKING-SVM, ST+RANKING-SVM, BIF+OLPP+RANKING-SVM, MULTI-CLASS CNN, CNN FEATURE+SVM, ST+SVM and BIF+OLPP+SVM.

	#1	#2	#3	#4	#5	#6	#7	#8
#1 RANKING-CNN	NAN	1	1	1	1	1	1	1
#2 RANKING-CNN FEATURE +RANKING-SVM	$6.36e^{-148}$	NAN	1	1	0.85	1	1	1
#3 ST+RANKING-SVM	0	0	NAN	1	0	0	1	1
#4 BIF+OLPP+RANKING-SVM	0	0	$1.79e^{-135}$	NAN	0	0	0.99	0.81
#5 MULTI-CLASS CNN	0	0.14	1	1	NAN	1	1	1
#6 CNN FEATURE+SVM	$4.12e^{-276}$	$8.90e^{-184}$	1	1	$5.43e^{-24}$	NAN	1	1
#7 ST+SVM	0	0	$1.94e^{-121}$	$2.00e^{-4}$	0	0	NAN	$3.66e^{-6}$
#8 BIF+OLPP+SVM	0	0	$4.56e^{-90}$	0.18	0	0	0.99	NAN

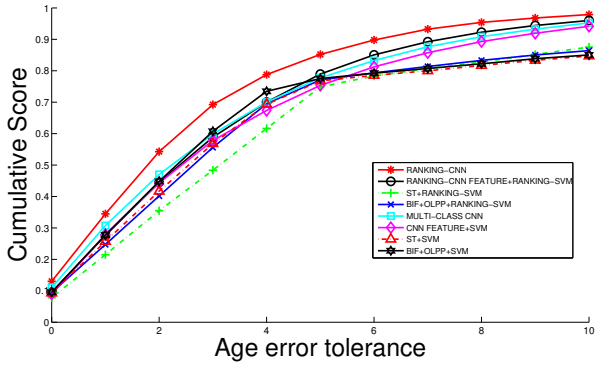

Figure 3. Comparison on Cumulative Score with L in $[0, 10]$.

Ranking-CNN can reach the accuracy of 89.90% for $L = 6$, and 92.93% for $L = 7$. The other fact we notice is that four CNN-based methods reach a higher accuracy for $L = 10$ than the others.

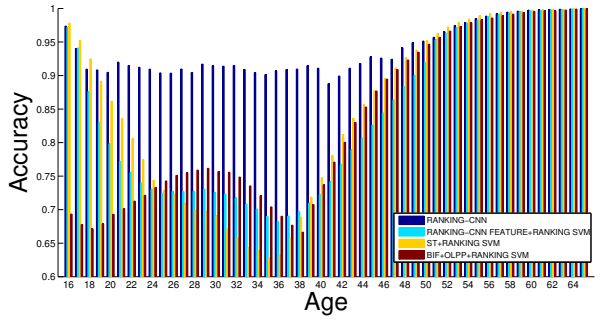

Figure 4. Accuracy of each binary ranker in ranking models.

In Fig. 4, we further compare the four ranking-based methods and report their performance on each binary ranker. Again, ranking-CNN demonstrates a consistent outstanding performance throughout all binary problems. Note that when the data for the binary rankers are not balanced (and thus higher baseline accuracy, e.g., age<20 and age>48), all rankers seem to perform quite well. However, when it comes to the age range with more balanced data (and thus lower baseline accuracy, age $20-48$), the superior performance of ranking-CNN is shown, and this would lead to better overall performance of age estimation. Again, our results clearly illustrated the remarkable improvement of using ranking-CNN for age estimation.

Last, to demonstrate that the experimental results we obtained do not happen simply by chance, we report in Table 4 the p-values from paired t-test at significant level 1%. In Table 4, if $p < 1\%$, we reject the null hypothesis. Otherwise, we don't. For example, when comparing "ranking-CNN" with "ranking-CNN feature+ranking SVM", the p-value $6.36e^{-148}$ is much less than 0.01, which means that we **reject** the null hypothesis that "the performance of ranking-CNN is not significantly improved". The "NaN" in the table means we could not compare a method with itself. As we can see, statistically, ranking-CNN significantly outperforms all other methods, which implies if we repeat the experiments for numerous times, then in 99% of those experiments, ranking-CNN would significantly outperform. From the table, Ranking-CNN Feature+Ranking SVM and the Multi-Class CNN tied for the second place, followed by CNN Feature+SVM. ST+Ranking SVM stands out among the engineered feature-based methods. Lastly, BIF+OLPP+Ranking-SVM ties with BIF+OLPP+SVM, and ST+SVM has no significant improvement than any other methods.

5. Conclusion

In this paper, we proposed ranking-CNN, a novel deep ranking framework for age estimation. We established a much tighter error bound for ranking-based age estimation and showed rigorously that ranking-CNN, by taking the ordinal relation between ages into consideration, is more likely to get smaller estimation errors when compared with multi-class classification approaches. Through extensive experiments, we show that statistically, ranking-CNN significantly outperforms other state-of-the-art age estimation methods on benchmark datasets.

Acknowledgment This work was partially supported by US National Science Foundation (NSF) under grant CNS-1637312, and by Ford Motor Company University Research Program under grant 2015-9186R.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] K.-Y. Chang and C.-S. Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, 24(3):785–798, 2015.

[3] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 585–592. IEEE, 2011.

[4] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2467–2474, 2013.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, Nov. 2011.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[8] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 766–774, 2014.

[9] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[10] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007.

[11] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th annual ACM International Conference on Multimedia*, pages 307–316. ACM, 2006.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[13] A. Gunay and V. V. Nabiyev. Automatic age classification with lbp. In *Computer and Information Sciences. ISCIS'08. 23rd International Symposium on*, pages 1–4. IEEE, 2008.

[14] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[15] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 112–119, 2009.

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[18] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[20] Y. H. Kwon and N. D. V. Lobo. Age classification from facial images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–767, 1994.

[21] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

[22] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[23] L. Le Cam et al. An approximation theorem for the poisson binomial distribution. *Pacific J. Math*, 10(4):1181–1197, 1960.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015.

[26] C. Li, Q. Liu, J. Liu, and H. Lu. Learning ordinal discriminative features for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2570–2577. IEEE, 2012.

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[28] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

[29] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[30] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[31] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14, 2016.

[32] E. D. Sontag. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[35] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[36] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511. IEEE, 2001.

[37] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *Applications of Computer Vision, 2015 IEEE Winter Conference on*, pages 534–541. IEEE, 2015.

[38] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *International Conference on Biometrics*, pages 464–473. Springer, 2007.

[39] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision*, pages 144–158. Springer, 2015.