Deep Age Estimation: From Classification to Ranking

Shixing Chen, Caojin Zhang, and Ming Dong, *Member, IEEE*

***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 set of basic CNNs, each of which is trained with ordinal age labels. Then, their binary outputs are aggregated for the final age prediction. From a theoretical perspective, we obtain an approximation for the final ranking error, show it is controlled by the maximum error produced among sub-ranking problems, and thus find a new error bound, which provides helpful guidance for the training and analysis of deep rankers. Based on the new error bound, we theoretically give an explicit formula for the learning of ranking-CNN and demonstrate its convergence using stochastic approximation method. Moreover, we rigorously prove that ranking-CNN, by considering ordinal relation between ages, is more likely to get smaller estimation errors when compared with multi-class classification approaches. Through extensive experiments, we show that ranking-CNN outperforms other state-of-the-art feature extractors and age estimators on benchmark datasets.**

***Index Terms*—Age estimation, Convolutional Neural Networks, Ranking algorithms, Error bound, Convergence.**

I. Introduction

HUMAN age is considered an important biometric trait for human identification or search. Relying on humans to supply age information from face images is often not feasible [1]. Thus, there has been a growing interest in the automatic determination of the specific age or age range of a subject based on a facial image. Some of the potential applications of automatic age estimation are in law enforcement, security control, and human computer interaction.

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. Specifically, simple geometry features (e.g., distances between eyes and nose) and texture features (e.g., skin wrinkles) were first adopted [2]. Later on, Biologically Inspired Features (BIF) [3] were proposed and widely adopted in age estimation applications. More recently, Scattering Transform (ST) [4] 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) [5].

S. Chen is with the Department of Computer Science, Wayne State University, Detroit, MI 48202. E-mail: schen@wayne.edu

C. Zhang is with the Department of Mathematics, Wayne State University, Detroit, MI 48202. E-mail: czhang@wayne.edu

Corresponding author. M. Dong is with the Department of Computer Science, Wayne State University, Detroit, MI 48202. E-mail: mdong@wayne.edu

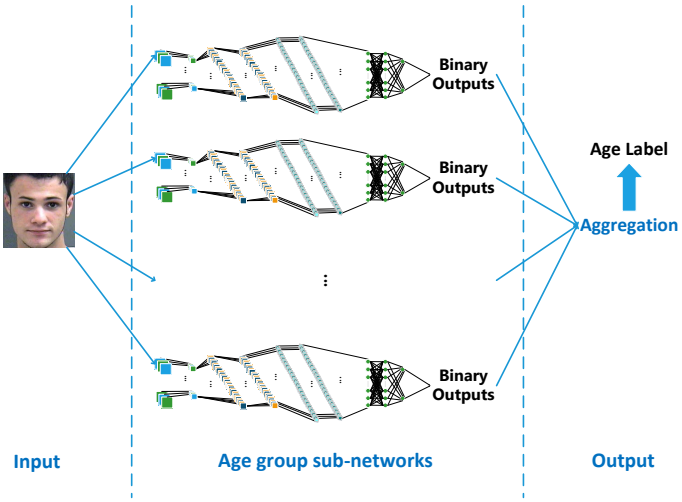


Fig. 1. Ranking-CNN for facial image-based age estimation.

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 [6], Multilayer Perceptrons [7], and the most commonly used Support Vector Machines (SVM) [3]. For regression methods, quadratic regression [5], Support Vector Regression (SVR) [3] and multi-instance regressor [8] 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 [9]. Experimental results show that the deeply-learned aging patterns lead to significant performance improvement on benchmark datasets [10] as well as unconstrained photos [11]. 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. When humans predict a person's age, it is usually easier to determine if a person is elder than a specific age than directly giving an exact age. Thus, cost-sensitive ranking techniques have recently been introduced to age estimation [4].

In this paper, we propose a novel age ranking approach based on CNN. Specifically, we propose a ranking-CNN model that contains a set 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. 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 set of basic CNNs, one for each age group. Different from the regression or the multi-class classification approaches, 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. Through extensive experiments, we show that ranking-CNN achieves superior results when compared with other state-of-the-art age estimation methods.
- From a theoretical point of view, we provide a tighter error bound for age ranking than prior work [4], which proved that the final ranking error is bounded by the sum of errors generated by all the classifiers. We divide the errors of sub-problems into two groups: overestimated errors (the sample's actual label is less than certain age classifier but was classified as older than that age) and underestimated errors (the sample's actual label is greater than that of certain age classifier but was classified as younger than that age). 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. From a technical perspective, the new error bound provides very helpful guidance for the training and analysis of ranking-CNN.
- Based on the new error bound, we give a Stochastic Gradient Descent (SGD) based scheme to train ranking-CNN in the context of GPU's high performance computing [12]. We employ stochastic approximation to assert the convergence, in which the parameters are updated as a stochastic process, leading to a limit of Ordinary Differential Equation (ODE) with stationary points that approximate the minimizers of the final ranking loss.
- Furthermore, we rigorously derive the expectation of prediction error of ranking-CNN and 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).

The abstract version of this paper has been published in [13]. The rest of the paper is arranged as follows. In Section II, we briefly review related work in age estimation, CNN, and the convergence analysis. In Section III, we first introduce ranking-CNN for age estimation. Then, we establish the theoretical error bound of ranking-CNN and show the convergence of learning ranking CNNs. Finally, we compare ranking-CNN with softmax-based multi-class CNNs and show that ranking method is preferred for age estimation. In Section IV, we present our age estimation results on the benchmark datasets. Finally, we conclude in Section V.

II. Related Work

A. Age Estimation

One of the earliest age estimation model can be traced back to [14], in which Active Appearance Model (AAM) [15] was employed to extract shape and appearance features from facial images. Based on these features, various classifiers such as shortest-distance classifier, quadratic function and neural networks were compared. Also, two assumptions were proposed: whether human aging process is age-specific or appearance-specific. That is, whether it is identical for everyone or only people with similar appearance would have similar aging processes.

Earlier works of age estimation usually follow the latter assumption and tend to cluster similar faces before estimation. In [16], 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. [17] proposed AGing pattErn Subspace (AGES). AGES is a person-specific age estimation method, which fulfills the estimation by projecting the facial image into the aging subspace with best reconstruction. However, a person's facial features might be almost identical in some age ranges. To resolve this issue, Zhang et al. [18] employed a warped Gaussian process to model a person's age, in which both person-specific and general aging information were adopted. In general, it is hard to obtain sufficient data to derive the long-term aging process for every individual. In [19], several short-term patterns, which usually are easier to get, were integrated to construct a long-term aging sequence. More recently, Shu et al. [20] aimed to automatically render aging faces in a personalized way by learning a set of age-group specific dictionaries.

Since the available images for a specific person are typically very limited, many researchers focus on developing non-personalized approaches instead. For instance, Yang and Ai [21] adopted a real AdaBoost algorithm to build a strong classifier from a series of weak ones using Local Binary Patterns [22]. Li et al. [23] proposed a method based on ordinal discriminative feature learning, which preserves locality ordinal information and removes redundancy features. Ni et al. [24] dealt with images with noisy labels through an outlier removal step using PCA and learned a multiple-instance regression estimator. In [3], BIF features were shown to be effective for age estimation on various datasets. Meanwhile, Guo et al. [25] investigated the influence of gender and race on age estimation while Lou et al. [26] introduced a graphical model to jointly learn age and facial expression labels. In [27], Eidinger et al. adopted dropout-SVM on the age estimation of unfiltered faces.

Recently, manifold learning algorithms were incorporated to achieve better performance of age estimation. In [5], Guo et al. proposed to use aging manifold with locally adjusted robust regressor. Dimension reduction approaches such as Principal

Component Analysis (PCA) [28], Locally Linear Embedding [29] and Orthogonal Locality Preserving Projections [30] were employed to learn a low-dimensional embedding. Then, SVR was used together with SVM for data approximation and local adjustment, respectively. Meanwhile, discriminative manifold learning was adopted for better visualization results in [31]. Later, Guo and Mu [32] proposed to use kernel partial least squares regression for simultaneous dimensionality reduction and age estimation.

More recently, CNN-based methods have been widely adopted for age estimation due to its superior performance over existing methods. Yi et al. [9] introduced a multi-task learning method with a relatively shallow CNN. Wang et al. [10] trained a deeper CNN for extracting features from different layers, and the features were then integrated by PCA. Based on these features, age estimation results from different regression and classification approaches were compared. In [33], Rothe et al. adopted the very deep VGG-16 architecture [34] for age estimation. In [35], Liu et al. used two large-scale deep neural networks, and fused the results from classification and regression for better performance. Zhu et al. [36] discussed an apparent age estimation problem with deeply learned features, in which the age labels are annotated by human assessors instead of the real chronological age. In both [11] and [37], CNN's performance on unconstrained facial images were validated. Hu et al. also considered to train the neural network from the age difference information [38].

Instead of multi-class classification and regression methods, ranking techniques derived from Ranking SVM [39], RankBoost [40], [41] and RankNet [42] were introduced to the problem of age estimation. With the ranking algorithms, the ordinal information of age labels is preserved, and the nature of human aging process is reflected. In [43], the method using ranking algorithms for age estimation was first introduced, in which multiple hyperplanes parallel to each other were used in a single kernel space. Later, a cost-sensitive ordinal ranking framework was proposed with ST features [4], where non-parallel hyperplanes were adopted to allow different kernel spaces for different binary classifiers. Most recently, Niu et al. [44] proposed to formulate age estimation as an ordinal regression problem with the use of multiple output CNN.

B. 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 [45] 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 tasks including language modeling, part-of-speech tagging and semantic role labeling with a convolutional architecture [46]. For text classification, CNN architectures have been widely adopted and achieved superior outcomes [47], [48].

In the computer vision field, CNN models have been applied to various tasks in the past decade. Great successes have been achieved in image classification [49], [50], object detection [51], [52], [53], face recognition [54], [55], [56] and image segmentation [57], [58]. Dating back to LeNet [59], [45], CNN was first introduced to solve the digit recognition problem using the MNIST database. The architecture of LeNet is relatively simple but effective. It contains two convolutional layers followed by two sub-sampling layers and two fully connected layers. The input is handwritten digits [60], and the output is the prediction from the network.

More recently, with the implementation using GPUs [49], [61], CNN models with deep architectures have achieved breakthroughs on object recognition problems in large-scale image datasets, e.g., the ImageNet dataset [62]. Furthermore, to build more effective CNN models, several new components were introduced: activation unit such as rectified linear unit (ReLU) [63] helps to accelerate the convergence during training and has a positive influence on the performance [49]; regularizer like dropout prevents overfitting by setting some activation units to zero in a specific layer [64]; and batch normalization allows the use of much higher learning rates to make training faster and to improve performance [65].

C. Convergence

Few theoretical results for the learning algorithm of CNNs is available even though it became one of the hottest topic for machine learning nowadays. Back Propagation (BP), a widely used algorithm for training neural networks, is shown to converge to a local minimum of the least squares error in [66], using an ODE approximation method. Detailed analysis has been gone through to prove the convergence theorem for a BP neural network with a hidden layer in [67]. BP with a momentum (BPM), a variation of BP, aims at improving its convergence speed. Phansalkar and Sastry analyzed the behavior of BPM for a one layer neural network with MAE type loss function in [68] and explains why BPM achieves a faster convergence. SGD is developed to avoid unnecessary work in computing the gradient over the entire dataset and deal with new data in an online setting.

As an online gradient method, convergence of SGD can be proved by stochastic approximation. It was first introduced by Robbins and Monro [69] in the early 1950s. Kushner discussed sufficient conditions for its convergence in his book [70], and then those criterion were adopted in [66] to study adaptive algorithms. Later, more general theory was presented in [71]. In recent years, it has been the subject of an enormous literature, both theoretical and applied, due to the large number of applications and the interesting theoretical issues in the analysis of "dynamically defined" stochastic processes.

III. 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 [27] to learn a nonlinear transformation of the

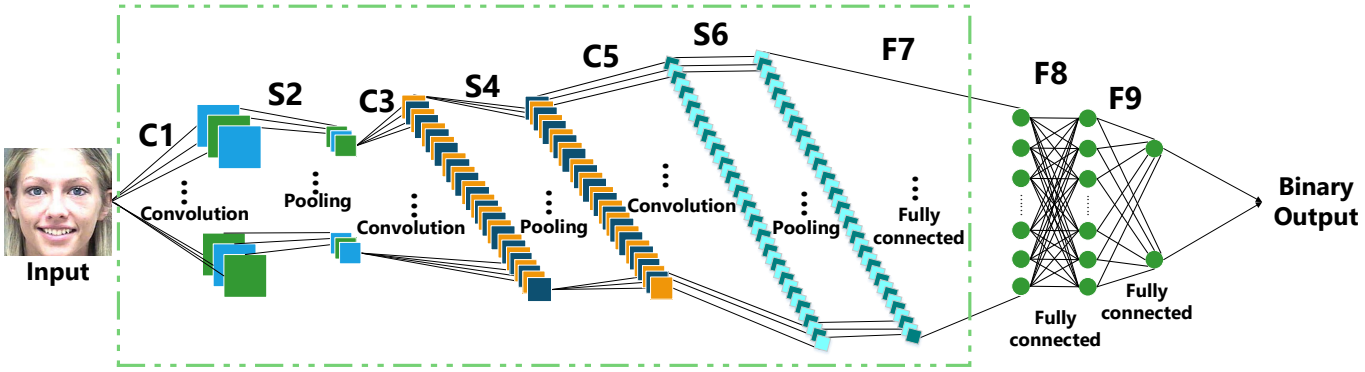


Fig. 2. Architecture of a Basic Binary CNN

input samples that captures their main variation. From the base network, we then train a set 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 SGD 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.

A. Basic Binary CNNs

1) Architecture and Algorithms: As shown in Fig. 2, a basic CNN has a classic architecture: 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) [63]. 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 [49].

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. We show the visualization of the feature maps after each layer later in Section III-A2.

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 [64]. 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. Notice that we use the logistic function instead of softmax as the output of a basic CNN is binary. 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 SGD. Detailed analysis on learning and convergence will be given in Section III-B3.

2) Feature Maps: With a single trained CNN, given an input face, we can generate a set of feature maps after each of the convolutional and pooling layers. As our model has three convolutional layers and three pooling layers, we can generate six sets of feature maps in total. The number of feature maps in each set are determined by the number of filters in the corresponding layer.

Representative feature maps extracted from the base CNN are shown in Fig. 3. There are six sets of feature maps, i.e., CONV1, POOL1, CONV2, POOL2, CONV3, and POOL3, and we show nine feature maps in each set. Specifically, CONV1 is the set of feature maps after the first convolutional layer. In CONV1, there are 96 feature maps, showing the convolved results of the input image with 96 filters in layer C1. We can see that the shown nine feature maps are concentrating on different areas of the input face, some of which highlight the eyes and the mouth while others focus on the face contour. After max-pooling layer S2, we can get the corresponding set of feature maps POOL1. Feature maps in POOL1 generally have a higher contrast to pass more information to successive layers.

Then, after the second round convolution, we obtain 256 feature maps in CONV2. Clearly, these feature maps have more detailed information than CONV1 to further depict facial features. Again, the contrast in feature maps in POOL2 are enhanced to be more informative. With the filters in the third convolutional layer C3, 384 feature maps in CONV3 are generated. Now, each feature map in CONV3 concentrate on a certain area to describe the original image in a specific way. After the final pooling layer S6, the output POOL3 with 384 feature maps would be flatten in F7 as the vector to represent the face before age estimation. From these feature maps, we can generally get to know what information has been extracted by the network from the original image.

B. 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,

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

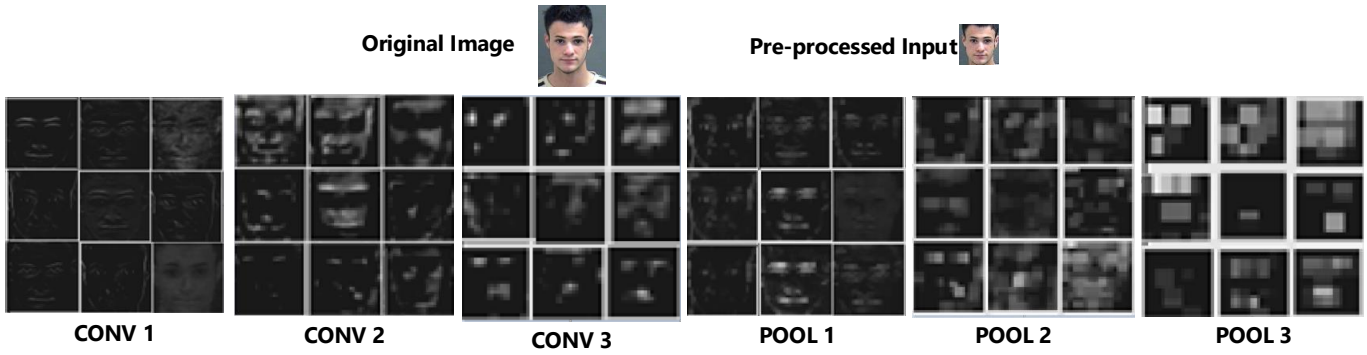


Fig. 3. Representative feature maps extracted from the base CNN: CONV1 for layer C1, POOL1 for layer S2, CONV2 for layer C3, POOL2 for layer S4, CONV3 for layer C5, and POOL3 for layer S6.

The binary ranking error $\varepsilon(x_i)$ is defined as,

$$\varepsilon_k(x_i) = [f_k(x_i) > 0][y_i \leq k\ ] + [f_k(x_i) \leq 0][y_i > k], \quad (2)$$

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. So, $\varepsilon_k(x_i) = 1$ if the ranking order is incorrect, and $\varepsilon_k(x_i) = 0$ otherwise.

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 $r(x_i)$,

$$r(x_i) = 1 + \sum_{k=1}^{K-1} [f_k(x_i) > 0]. \quad (3)$$

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 III-B1 followed by the convergence analysis and theoretical comparison between ranking and softmax-based multi-class classification.

In Algorithm 1, we provide the complete training and testing procedure of ranking-CNN.

Algorithm 1 Algorithm of Ranking-CNN

procedure TRAINING PROCEDURE
  pretrain *Base CNN*
  top:
  for *k = 1 to K-1* **do**
    $e_k \leftarrow k_{th}$ *Basic CNN*
  end for
  $k' \leftarrow$ **sort** e_k
  for k' *= 1 to K-1* **do**
    $D_k^+ = \{(x_i, +1) | y_i > k'\}$
    $D_k^- = \{(x_i, -1) | y_i \leq k'\}$
    fine-tune k'_{th} *Basic CNN* $\leftarrow e_{k'}$
  end for
  if *not converged*
    goto *top*
  end if
procedure TESTING PROCEDURE
  for *k = 1 to K-1* **do**
    $f_k(x_i) \leftarrow k_{th}$ *Basic CNN*
  end for
  final prediction $r(x_i) \leftarrow 1 + \sum_{k=1}^{K-1} [f_k(x_i) > 0]$

1) Error Bound: In ranking-CNN, we divide an age ranking estimation problem, ranging from $1, \cdots, K$, into a set 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 [4], which claims that the final ranking error is bounded by the sum of errors generated by all the classifiers. We adopt the idea in [4] 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^+ as the total number of sub-classifiers that misclassified when $y \leq k$. That is, $E^+ = \sum_{k=1}^{K-1} \gamma_k^+$, where $\gamma_k^+ = [f_k(x) > 0][y \leq k]$ and $[\cdot]$ is an indicator function taking value of 1 when the condition in $[\cdot]$ holds, 0 otherwise. Similarly, we denote $E^- = \sum_{k=1}^{K-1} \gamma_k^-$ for the case of $y > k$, where $\gamma_k^- = [f_k(x) \leq 0][y > k]$.

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

$$e_k(x) = \begin{cases} e_k^+ = (k - y + 1)\gamma_k^+ & y \leq k \\ e_k^- = (y - k)\gamma_k^- & y > k. \end{cases} \quad (4)$$

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:

$$|r(x) - y| \leq \max_k e_k(x), \quad (5)$$

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 (4) 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 (5) follows.

Firstly, we begin by definition:

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

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

$$\begin{aligned} & r(x) - (E^+ - E^-) \\ & = 1 + \textstyle\sum_{k=1}^{K-1} [f_k(x) > 0][y > k] + \sum_{k=1}^{K-1} [f_k(x) \leq 0][y > k] \\ & = 1 + \textstyle\sum_{k=1}^{K-1} ([f_k(x) > 0] + [f_k(x) \leq 0])[y > k] \\ & = 1 + \textstyle\sum_{k=1}^{K-1} [y > k] \\ & = y. \end{aligned} \tag{7}$$

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.

Notice that $\{e_{(j)}^+, j = 1, 2, \cdots, E^+\}$ is a set of losses made by sub-classifiers with incorrect classification, where E^+ is the total number of sub-classifiers that misclassified when $y \leq k$. Next, based on the definition of the loss function in (4), when $y \leq k$, the loss associated with a sub-classifier must be greater than 1, i.e., $e_{(j)}^+ \geq 1$. Moreover, the difference of losses between two adjacent classifiers is at least 1, i.e., $e_{(j)}^+ - e_{(j-1)}^+ \geq 1$.

Then, we get:

$$\begin{aligned} e_{(E^+)}^+ = \; & e_{(1)}^+ + e_{(2)}^+ - e_{(1)}^+ + \cdots + e_{(E^+)}^+ - e_{(E^+-1)}^+ \\ & \geq \underbrace{1 + 1 + \cdots + 1}_{E^+} = E^+ \end{aligned} \tag{8}$$

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.

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:

- Theoretically, it was mentioned in both [4] and [44] 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.
- 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.
- Mathematically, based on the new error bound, we can theoretically give an explicit formula for the learning of ranking-CNN and demonstrate its convergence using stochastic approximation method. Moreover, 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 proofs are given in following sections.

3) Learning and Convergence of Ranking-CNN: For each ranker k, given a sample (x, y), consider binary target:

$$r(x) = \begin{cases} 1 & \text{If } y > k \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

Given the loss function for each ranker as:

$$\ell(w_k) = |y - k|(-r \log P(r = 1|x) - (1 - r) \log(1 - P(r = 1|x))) \tag{10}$$

where w_i denotes the parameters in k-th ranker.

In training ranking CNN, we implement the Back Propagation (5) using stochastic gradient decent as to minimize the maximum cross entropy loss as:

$$w_i^{n+1} = \begin{cases} w_i^n - \alpha_n \nabla \ell(w_i), & \text{If } \ell(w_i) = \max \ell(w_i) \\ w_i^n, & \text{Otherwise} \end{cases} \tag{11}$$

We let the learning rate satisfies:

$$\sum_n \alpha_n = \infty, \sum_n \alpha_n^2 < \infty, \lim \alpha_n = 0. \tag{12}$$

Denote

$$L(w_1, \cdots, w_{K-1}) = \max(\ell(w_1), \cdots, \ell(w_{K-1})) \tag{13}$$

We concatenate all parameters w_i, for $i = 1, \cdots, K-1$ into a vector W and interpolate the the updated parameters in each iterations as a sequence of stochastic process $W^n(\cdot)$ as follows:

$$t_n = \sum_{i=1}^{n-1} \alpha_i, \tag{14}$$

$$W^0(t) = W^n \quad \text{for } t \in [t_n, t_{n+1}) \tag{15}$$

$$W^n(t) = W^0(t + t_n). \tag{16}$$

Then using the stochastic approximation techniques provided in [56], or theorem 5.1 [60], we claim the sequence weakly

converges to a limit ODE (convergence in distribution defined as Section II [56]):

Theorem 2: Let $W^n(0)$ be fixed vectors or random vectors independent of α_n Then $W^n(\cdot)$ weakly converges to $W(\cdot)$, where $W(\cdot)$ satisfy the system of ODEs:

$$\frac{\partial W(t)}{\partial t} = \frac{\partial E_{x\in D}(L(W))}{\partial W} \tag{17}$$

Then the parameters converges to ODE's equilibrium point $(w_1^*, \cdots, w_{K-1}^*)$ (local minimum of the loss L) by Lyapunov condition [64]. Due to the error bound, we obtain an approximation of the local minimum of the aggregation loss: $E_{x\in D}|r(x)-y|$.

4) Ranking vs. Softmax: In this section, we theoretically show that our ranking-CNN outperforms softmax method because it is more likely to get smaller ranking errors $|r(x)-y|$. Thus, instead of a softmax classifier, ranking method is preferred for age estimation. The reason is that softmax failed to take the ordinal relation between ages into consideration.

A basic CNN in ranking-CNN differs from the softmax multi-class classification approach in the output layer. Suppose $z_1, \cdots, z_K$ are unnormalized outputs which explains the probability of a sample x belonging to each class. Denote weights $a_i = e^{z_i}$ and $\hat{y}$ as the estimated age label. For softmax, the posterior probability of each class is given by:

$$\begin{aligned} P(\hat{y}\in i|x) &= \frac{e^{z_i}}{\sum_{k=1}^K e^{z_k}} \\ &= \frac{a_i}{\sum_{k=1}^K a_k}, \end{aligned} \tag{18}$$

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

$$E(|r(x)-y||y) = \textstyle\sum_{i=1}^K |i-y|P(\hat{y}=i|x). \tag{19}$$

For the 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:

$$\begin{aligned} P(f_i(x)>0|x) &= \frac{e^{z_{i+1}}}{e^{z_i}+e^{z_{i+1}}} \\ &= \frac{a_{i+1}}{a_i+a_{i+1}}. \end{aligned} \tag{20}$$

The expected error for a given sample is

$$E(|r(x)-y||y) = \textstyle\sum_{i=1}^K |i-y|P(\hat{y}=i|x). \tag{21}$$

We present a theorem for a three ordinal class problem.

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

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

Proof. Given a sample with label 1, the expected error for ranking-CNN is

$$\begin{aligned} E(|r_1(x)-y||y=1) &= 2P(f_1(x)>0, f_2(x)>0|W,U,X) \\ &\quad +P(f_1(x)>0, f_2(x)<0|W,X) \\ &\quad +P(f_1(x)<0, f_2(x)>0|W,X) \\ &= \frac{2bc+b^2+ac}{(a+b)(b+c)}. \end{aligned} \tag{23}$$

For softmax,

$$\begin{aligned} E(|r_2(x)-y||y=1) = &\quad 2P(r_2(x)=2|W,X) \\ &+P(r_2(x)=3|W,X) \\ &= \frac{2c+b}{a+b+c}. \end{aligned} \tag{24}$$

Similarly, given $y=2$,

$$\begin{aligned} E(|r_1(x)-y||y=2) &= P(f_1(x)>0, f_2(x)>0|W,X) \\ &\quad +P(f_1(x)<0, f_2(x)<0|W,X) \\ &= \frac{ab+bc}{(a+b)(b+c)}. \end{aligned} \tag{25}$$

$$\begin{aligned} E(|r_2(x)-y||y=2) = &\quad P(r_2(x)=1|W,X) \\ &+P(r_2(x)=3|W,X) \\ &= \frac{a+c}{a+b+c}. \end{aligned} \tag{26}$$

Given $y=3$,

$$\begin{aligned} E(|r_1(x)-y||y=3) &= 2P(f_1(x)<0, f_2(x)<0|W,X) \\ &\quad +P(f_1(x)>0, f_2(x)<0|W,X) \\ &\quad +P(f_1(x)<0, f_2(x)>0|W,X) \\ &= \frac{2ab+b^2+ac}{(a+b)(b+c)}, \end{aligned} \tag{27}$$

and

$$\begin{aligned} E(|r_2(x)-y||y=3) = &\quad 2P(r_2(x)=1|W,X) \\ &+P(r_2(x)=2|W,X) \\ &= \frac{2a+b}{a+b+c}. \end{aligned} \tag{28}$$

For ranking-CNN, it follows

$$\begin{aligned} E(|r_1(x)-y|) &= \textstyle\sum_{i=1}^3 E(|r_1(x)-i||y=i) \\ &2+\frac{ab+bc}{(a+b)(b+c)}. \end{aligned} \tag{29}$$

Similarly, for softmax,

$$\begin{aligned} E(|r_2-y|) &= \textstyle\sum_{i=1}^3 E(|r_2(x)-i||y=i) \\ &= 2+\frac{a+c}{a+b+c}. \end{aligned} \tag{30}$$

Since

$$\begin{aligned} &\frac{a+c}{a+b+c} - \frac{ab+bc}{(a+b)(b+c)} \\ &= \frac{a^2c+c^2a}{(a+b)(b+c)(a+b+c)} > 0, \end{aligned} \tag{31}$$

then we conclude

$$E(|r_1(x)-y|) < E(|r_2(x)-y|). \tag{32}$$

Furthermore, the case 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 $i = 1,2\cdots K$, becomes

$$P(\hat{y}=i|y) = \sum_{A\in\mathscr{F}_i}\prod_{j\in A} p_j \prod_{j\in A^c}(1-p_j), \tag{33}$$

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 hopeless since listing all the probability out as we did in theorem 3 looks impractical. Though Le Cam, L. [72] gives an approximation of Poisson-Binomial by a Poisson distribution, the computation for the

$$E(|r_1(x)-y|)=\sum_{y=1}^{K}\sum_{r=1}^{K}|r-y|P(\hat{y}=r) \tag{34}$$

is not an easy task. To overcome this, statistics provides us a powerful tool with no need for knowing the actual distributions. To further strengthen that ranking-CNN wins over softmax in age estimation, we propose a t-test with hypothesis that compared with softmax, our ranking-CNN does reduce the ranking error in the sense of statistical significance. The details will be discussed later in the experiment section.

C. 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.

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 [73]. Then, face alignment is conducted based on the location of eyes. Finally, the image is resized to a standard size of 256×256×3 for network training and age estimation.

IV. Experiments

In this section, we demonstrate the performance of ranking-CNN through extensive experiments. We first choose the appropriate architecture for the basic CNN by evaluating it on binary age ranking problems. Then, we move to multiple age estimation problems and evaluate ranking-CNN.

For multiple age estimation, we compared the features learned by ranking-CNN with the ones obtained through BIF+OLPP [3], ST[4], 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 [11], [9], 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 one. As for the age estimators, SVM is selected for comparison due to its proved performance [3]. In ranking-based approach (Ranking-SVM), following [4], SVM is used as the binary classifier for each age label and the results are aggregated to give the final output. Finally, we also directly compare age estimation results obtained by ranking-CNN with the ones reported in the literature by leading deep learners on benchmark datasets.

The comparison and evaluation of different methods in our experiments are reported in terms of precision of each age group, accuracy of each binary ranker as well as two widely adopted performance measures [44], [4]: Mean Absolute Error (MAE) and Cumulative Score (CS). MAE computes the absolute costs between the exact and the predicted ages (the lower the better):

$$MAE=\sum_{i=1}^{M}e_i/M, \tag{35}$$

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):

$$CS(L)=\sum_{i=1}^{M}[e_i\le L]/M, \tag{36}$$

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. Suppose $\{\varepsilon_i\}_{i=1}^N$ are the errors obtained through the test set $\{(x_i,y_i)\}_{i=1}^N$ by ranking-CNN, and $\{\tau_i\}_{i=1}^N$ are errors in testing by another method. We employ paired t-test to determine if the former significantly outperforms the latter. A two-sample t-statistic with unknown but equal variance is computed as:

$$t=\frac{\mu_1-\mu_2}{S_{1,2}\sqrt{\frac{2}{n}}}, \tag{37}$$

where μ_1 and μ_2 are the mean of two sets of errors respectively, $S_{1,2}=\sqrt{\frac{S_1^2+S_2^2}{2}}$, and S_1, S_2 are unbiased estimators of variances of two samples, where

$$\begin{aligned} S_1^2&=\frac{1}{N-1}\sum_{n=1}^{N}(\varepsilon_n-\mu_1)^2\\ S_2^2&=\frac{1}{N-1}\sum_{n=1}^{N}(\tau_n-\mu_2)^2 \end{aligned} \tag{38}$$

Define H_0: $\mu_1-\mu_2=0$ (the performance of ranking-CNN is not significantly improved), H_1: $\mu_2-\mu_1>0$ (otherwise). In the hypothesis test, we compute the p value at 1% significance level. If the p value is small enough, we reject the hypothesis H_0.

A. Basic CNN on Binary Age Ranking

We implemented two architectures for the basic CNN in the GPU mode with Caffe [61], namely, $2+2$ and $3+3$. For the $2+2$ architecture, it is derived from LeNet [45]. It contains

TABLE I
BASIC CNNS FOR BINARY AGE RANKING: ARCHITECTURE AND INITIALIZATION. THE HIGHEST ACCURACY IS HIGHLIGHTED IN **BOLD**.

	20-29 vs. 40-49				<20 vs. >50			
STRUCTURE	2+2	2+2	3+3	3+3	2+2	2+2	3+3	3+3
WEIGHT INITIALIZATION	XAVIER	GAUSSIAN	XAVIER	GAUSSIAN	XAVIER	GAUSSIAN	XAVIER	GAUSSIAN
# OF SAMPLES	3000	3000	3000	3000	1500	1500	1500	1500
ACCURACY	89.20% ±0.21%	88.13% ±0.15%	93.95% ±0.13%	**96.32% ±0.18%**	95.35% ±0.19%	94.98% ±0.17%	96.28% ±0.14%	**98.72% ±0.12%**

two convolutional layers with 20 and 50 filters in each layer respectively, followed by max-pooling layers and two fully connected layers. In the first fully connected layer, there are 500 outputs, and the number of outputs in the second fully connected layer is decided by the number of categories. For the $3+3$ architecture, it is similar to our basic CNN shown in Fig. 2. It is derived from a simplified version of the ImageNet CNN [49] with fewer layers for higher efficiency [11].

The networks are initialized with random weights generated in two methods. For the weights following Gaussian distribution, the mean is 0, and standard deviation is 0.01. For the Xavier initialization [74], the weights $W \sim U(-scale, scale)$ follow a uniform distribution with the range inversely proportional to the number of incoming and outgoing nodes:

$$\begin{aligned} scale &= \sqrt{3/n} \\ n &= \frac{fan_{in} + fan_{out}}{2} \\ fan_{in} &= num_{channel} \times columns_{filter} \times rows_{filter} \\ fan_{out} &= num_{output} \times columns_{filter} \times rows_{filter} \end{aligned} \tag{39}$$

where in our case, for example, in the first convolutional layer C1, $num_{channel}$ is 3, num_{output} is 96, $columns_{filter}$ and $rows_{filter}$ are both 5.

We evaluated the architectures of the networks on two binary age ranking problems: age groups 20-29 vs. 40-49, and age groups <20 vs. >50 on MORPH dataset. MORPH contains 55,134 facial images with the age range from 16 to 77. It provides specific age, gender and ethnicity information for each individual. Based on the availability of samples, we randomly selected 6,000 and 3,000 images from MORPH, respectively, for the two problems. The selection is balanced over age groups. In our experiments, 80% of the data is used for training and the rest 20% for testing (no overlapping with training). The averaged accuracy is reported with standard deviations over 10 runs. In each run, the network is trained using supervised training, and the maximum number of iterations is set at 100,000. We consider the training converges when the change of training error between two adjacent iterations is less than 0.001.

As we can see in Table I, the $3+3$ architecture and Gaussian initialization $N(0, 0.01^2)$ gives the highest classification accuracy in both problems. For the same architecture, Xavier initialization generates comparable results better than all combinations with $2+2$ architecture. For $2+2$ CNNs, Xavier initialization actually gives higher accuracy than Gaussian. In "<20 vs. >50" problem, $2+2$ CNNs give close accuracy but when it comes to a more complex situation (i.e., "20-29 vs. 40-49"), the accuracy decreases dramatically. Since $2+2$ CNNs are generally trained faster than $3+3$ CNNs, we can infer that if the problem is not too complicated and computing resource is limited, then $2+2$ CNNs could still be considered. In our case, since we have to distinguish between adjacent ages, we select the $3+3$ architecture and Gaussian initialization for best performance. It is used for all the basic networks in ranking-CNN to complete the remaining experiments.

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.

B. Multiple Age Estimation

In this section, we evaluate the performance of Ranking-CNN on three benchmark datasets: MORPH Album 2 [75], FG-NET [76] and Adience Faces benchmark [27].

1) MORPH: To further demonstrate the performance improvement of ranking-CNN, we consider the age estimation problem in the range between 16 and 66 years old on the most commonly used age estimation benchmark dataset MORPH Album 2, and compare ranking-CNN with other state-of-the-art feature extractors and age estimators. First, we pre-train a base network with 26,580 image samples from the unfiltered faces dataset [27]. The age group labels for these images are used in training as surrogate labels [77]. Then, we fine-tune our ranking-CNN model on MORPH.

In our experiments, when fine-tuning from the pre-trained base CNN to basic CNNs, we set the learning rate for the last fully-connected layer 10 times of the one used in the previous layers. Thus, the majority of the weights in the basic CNNs has only a slight difference, all similar to the ones in the base CNN. In principle, this training procedure works similarly as weight sharing, but with the additional benefit of easier parallelization. That is, the 50 basic CNNs can be fine-tuned parallelly on a distributed computing platform, while traditional weight-sharing has to be done sequentially.

Following the settings used in some recent work on age estimation [44], [10], [78], [79], we randomly select 54,362 samples in the age range between 16 and 66 from the MORPH dataset. The age and gender information of the selected samples is shown in Table II. 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 repeat five independent runs to evaluate the performance during experiments.

As there are 51 age groups in this age range, 50 binary rankers are needed for ranking approaches (i.e., ranking-CNN

TABLE II
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

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 repeated with five independent runs.

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). CNN feature and ranking-CNN feature are the output after layer F8 of multi-class CNN and Ranking-CNN respectively. Also, we have 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 III. A dash in the table means that the selected feature set is not applicable to the selected estimator.

As shown in Table III, ranking-CNN with its features achieves the lowest MAE of 2.96±0.015 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±0.028. Note the multi-class CNN represents the commonly used CNN-based age estimation methods [11], [9]. Our experimental results strongly support the theoretical results (ranking vs. softmax) we presented before. 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±0.030 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.

The comparison in terms of CS of the eight combinations of features and estimators are given in Fig. 4. Clearly, ranking-CNN outperforms all others across the entire range of L (age error tolerance range) from 0 to 10. Specifically, 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.

In addition, in Fig. 5, we compare the estimation precision of each age category for the eight combinations. The precision is defined as below:

$$precision = \frac{samples_{correct} \cap samples_{all}}{samples_{all}} \quad (40)$$

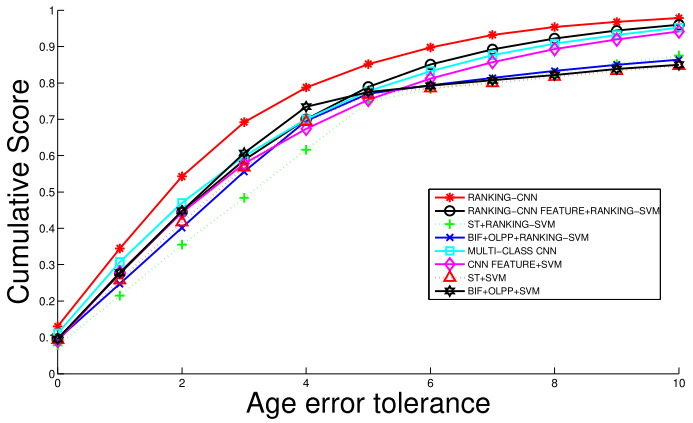


Fig. 4. Comparison on Cumulative Score with L in $[0, 10]$.

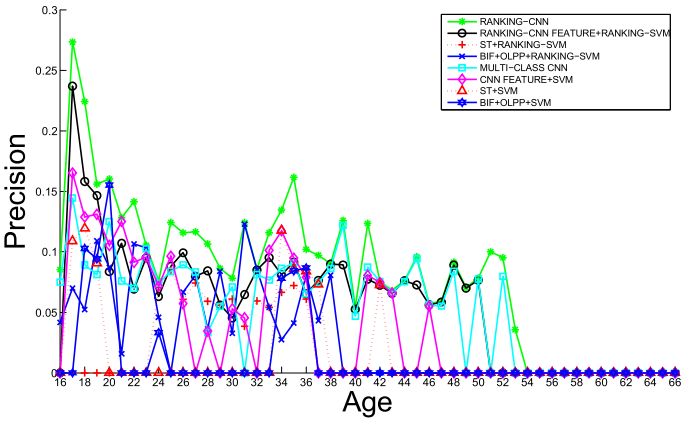


Fig. 5. Precision for each age group after aggregation.

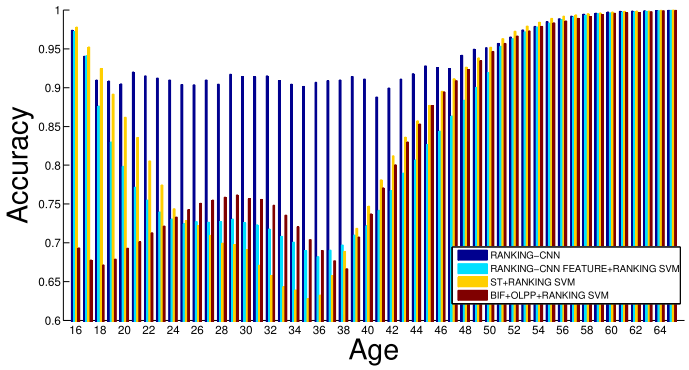


Fig. 6. Accuracy of each binary ranker in ranking models.

where $samples_{correct}$ denotes the samples correctly classified to a certain age category, and $samples_{all}$ denotes the total number of samples classified to this age category. It is obvious that ranking-CNN has a more consistent performance on each of the age groups. For methods like BIF+OLPP+SVM, there are many age categories with 0 precision. Taking a closer look, we found out that this is caused by the unbalanced classification of the multi-class estimators, where samples are mostly classified to certain categories instead of all the categories. The problem can be alleviated by ranking-CNN to some extent. In fact, none of the methods reach the age categories after 54. This is mainly because comparing with more than 50K samples in total, there are too few samples for age categories after 54 (averagely around 80 samples in each age category).

TABLE III
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 MODEL	SVM	4.99±0.035	5.15±0.040	3.95±0.020	-
	MULTI-CLASS CNN	-	-	3.65±0.028	-
RANKING MODEL	RANKING-SVM	5.03±0.028	4.88±0.030	-	3.63±0.019
	RANKING-CNN	-	-	-	**2.96±0.015**

TABLE IV
T TEST OUTCOMES OF ALL EIGHT COMBINATIONS OF FEATURES AND ESTIMATORS.

	RANKING-CNN	RANKING-CNN FEATURE +RANKING-SVM	ST+RANKING-SVM	BIF+OLPP +RANKING-SVM
RANKING-CNN	NAN	1	1	1
RANKING-CNN FEATURE + RANKING-SVM	$6.36e^{-148}$	NAN	1	1
ST+RANKING-SVM	0	0	NAN	1
BIF+OLPP+RANKING-SVM	0	0	$1.79e^{-135}$	NAN
MULTI-CLASS CNN	0	0.14	1	1
CNN FEATURE+SVM	$4.12e^{-276}$	$8.90e^{-184}$	1	1
ST+SVM	0	0	$1.94e^{-121}$	$2.00e^{-4}$
BIF+OLPP+SVM	0	0	$4.56e^{-90}$	0.18

	MULTI-CLASS CNN	CNN FEATURE+SVM	ST+SVM	BIF+OLPP+SVM
RANKING-CNN	1	1	1	1
RANKING-CNN FEATURE + RANKING-SVM	0.85	1	1	1
ST+RANKING-SVM	0	0	1	1
BIF+OLPP+RANKING-SVM	0	0	0.99	0.81
MULTI-CLASS CNN	NAN	1	1	1
CNN FEATURE+SVM	$5.43e^{-24}$	NAN	1	1
ST+SVM	0	0	NAN	$3.66e^{-6}$
BIF+OLPP+SVM	0	0	0.99	NAN

In Fig. 6, we further compared 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.

To demonstrate that the experimental results we obtained do not happen simply by chance, we report in Table IV the p-values from paired t-test. We report the p values of the paired t-test at significant level 1%. In Table IV, 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 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.

Furthermore, in Table V, we compare ranking-CNN with other deep learning-based age estimation models, i.e., Ordinal Regression with CNN (OR-CNN) [44], Metric Regression with CNN (MR-CNN) [44], Deep EXpectation (DEX) [33] and GoogLeNet in [38]. Since all the experiments are carried out on the MORPH dataset and we followed the same setting for data partition, we can directly compare the MAE of Ranking-CNN with the ones obtained by these deep learners. Notice that in order to make a fair comparison among all the deep learners, all existing results are reported without pre-training using additional facial images. For ranking-CNN, results with and without pre-training are both reported. Clearly, ranking-CNN outperforms these deep learning models in both cases.

Finally, we show the efficiency brought by the new error bound with a modified training strategy. According to [4] and our experiment results, the basic CNNs between age groups 36 and 45 get the largest training errors as they have more balanced training data (and thus lower baseline accuracy). For these basic CNNs, we train them until the change of the

training errors between two adjacent iterations is less than 0.001. For all other basic CNNs (16 to 35 and 46 to 65), we only train them until the change is less than 0.01. In this experiment setting, 80% of the basic CNNs are trained with dramatically less epochs (60% less on average), leading to much faster training. Yet, we still achieved very competitive results on age estimation, an MAE of 3.07±0.017.

TABLE V
COMPARISON WITH OTHER DEEP LEARNING MODELS ON THE MORPH DATASET. THE LOWEST MAE IS HIGHLIGHTED IN **BOLD**.

	MAE
Ranking-CNN	**2.96±0.015**
MR-CNN [44]	3.27±0.14
OR-CNN [44]	3.34±0.28
DEX [33]	3.25
GoogLeNet [38]	3.13
Ranking-CNN (without pre-training)	3.03±0.018

2) FG-NET: The FG-NET dataset is another benchmark dataset for age estimation. Since there are merely 1,002 photos in this dataset, it is not suitable for direct training of deep learners. Thus, we evaluate the performance on this dataset by fine-tuning the ranking-CNN model trained on the MORPH dataset. The age range we considered in Section IV-B1 is 16 to 66, so we select the 405 samples from FG-NET in the same range for this experiment. Similarly, we use 80% of these samples for training and 20% for testing and compare the MAE results with prior arts.

TABLE VI
AGE ESTIMATION RESULTS ON FG-NET DATASET. THE LOWEST MAE IS HIGHLIGHTED IN **BOLD**.

	MAE
Ranking-CNN	**4.13**
DEX [33]	4.63
CSOHR [4]	4.48
BIF+OLPP+SVM [3]	4.77
RankBoost [41]	5.67

As shown in Table VI, ranking-CNN outperforms other models on FG-NET dataset as well, and achieves the lowest MAE of 4.13. This further demonstrates the effectiveness and generalization ability of ranking-CNN. CSOHR achieves the second best MAE result of 4.48 while the MAE of DEX is 4.63. For BIF+OLPP+SVM and RankBoost, the MAE results are 4.77 and 5.67 respectively.

3) Adience: There are 26,580 photos in the Adience benchmark dataset of unfiltered faces. The samples are categorized into eight age groups with labels "0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53 and over 60", so we train seven basic CNNs for this task. Following the same settings in some recent work [11], [37], [27], we randomly select 80% of the samples as the training set and the rest 20% as the testing set.

TABLE VII
AGE ESTIMATION RESULTS ON THE ADIENCE BENCHMARK. THE HIGHEST ACCURACY IS HIGHLIGHTED IN **BOLD**.

	Accuracy
Ranking-CNN	**53.7±4.4**
CNN [11]	50.7±5.1
Cascaded CNN [37]	52.88±6
Dropout-SVM [27]	45.1±2.6

As shown in Table VII, the mean accuracy±standard error over all the age categories by ranking-CNN are compared with several results recently reported in the literature. It is obvious that ranking-CNN outperforms other methods and achieves the highest accuracy of 53.7 ± 4.4 for age categorization on Adience. Other CNN-based models also achieve good results. The accuracy of Cascaded CNN is 52.88 ± 6, and the multi-class CNN which has the architecture similar to the base CNN in ranking-CNN achieved 50.7 ± 5.1. The dropout-SVM method has the lowest accuracy of 45.1 ± 2.6 among the compared models.

V. CONCLUSION

In this paper, we proposed ranking-CNN, a novel deep ranking framework for age estimation. Our model contains a set of basic CNNs, each of which is initialized with the pre-trained base CNN and fine-tuned with ordinal labels. The binary output of basic CNNs are aggregated to make the final age prediction. From a theoretical perspective, we established a much tighter error bound for ranking-based age estimation, based on which, we mathematically proved the convergence of SGD-based training of ranking-CNN using a novel stochastic approximation approach and rigorously showed 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 ranking-CNN outperforms other state-of-the-art age estimation methods on benchmark datasets.

ACKNOWLEDGMENT

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

REFERENCES

[1] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance," in *2013 International Conference on Biometrics (ICB)*. IEEE, 2013, pp. 1–8.
[2] 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)*, 1994, pp. 762–767.
[3] 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)*, 2009, pp. 112–119.
[4] 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*, vol. 24, no. 3, pp. 785–798, 2015.
[5] 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*, vol. 17, no. 7, pp. 1178–1188, 2008.
[6] A. Gunay and V. V. Nabiyev, "Automatic age classification with lbp," in *Computer and Information Sciences. ISCIS'08. 23rd International Symposium on*. IEEE, 2008, pp. 1–4.
[7] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, 2004.
[8] B. Ni, Z. Song, and S. Yan, "Web image and video mining towards universal and robust age estimator," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1217–1229, 2011.
[9] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Asian Conference on Computer Vision*. Springer, 2015, pp. 144–158.

[10] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 534–541.

[11] 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)*, 2015, pp. 34–42.

[12] J. Bolz, I. Farmer, E. Grinspun, and P. Schröoder, "Sparse matrix solvers on the gpu: conjugate gradients and multigrid," in *ACM Transactions on Graphics (TOG)*, vol. 22. ACM, 2003, pp. 917–924.

[13] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] 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*, vol. 24, no. 4, pp. 442–455, 2002.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[16] 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*, vol. 29, no. 12, pp. 2234–2240, 2007.

[17] 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*. ACM, 2006, pp. 307–316.

[18] Y. Zhang and D.-Y. Yeung, "Multi-task warped gaussian process for personalized age estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2622–2629.

[19] J. Suo, X. Chen, S. Shan, and W. Gao, "Learning long term face aging patterns from partially dense aging databases," in *IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 622–629.

[20] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3970–3978.

[21] Z. Yang and H. Ai, "Demographic classification with local binary patterns," in *International Conference on Biometrics*. Springer, 2007, pp. 464–473.

[22] 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*, vol. 28, no. 12, pp. 2037–2041, 2006.

[23] 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)*. IEEE, 2012, pp. 2570–2577.

[24] B. Ni, Z. Song, and S. Yan, "Web image mining towards universal age estimator," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 85–94.

[25] G. Guo and G. Mu, "Human age estimation: What is the influence across race and gender?" in *Computer Vision and Pattern Recognition Workshops, IEEE Computer Society Conference on*, 2010, pp. 71–78.

[26] Z. Lou, F. Alnajar, J. Alvarez, N. Hu, and T. Gevers, "Expression-invariant age estimation using structured learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[27] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[29] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[30] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3608–3614, 2006.

[31] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 578–584, 2008.

[32] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 657–664.

[33] 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*, pp. 1–14, 2016.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 16–24.

[36] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 25–31.

[37] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 1–8.

[38] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3087–3097, 2017.

[39] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Advances in Neural Information Processing Systems*, pp. 115–132, 1999.

[40] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 933–969, 2003.

[41] P. Yang, L. Zhong, and D. Metaxas, "Ranking model for facial age estimation," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3404–3407.

[42] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*. ACM, 2005, pp. 89–96.

[43] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Pattern Recognition (ICPR), 20th International Conference on*. IEEE, 2010, pp. 3396–3399.

[44] 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.

[45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

[47] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

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

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

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[51] 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)*, 2014, pp. 580–587.

[52] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[53] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, "Large-margin multimodal deep learning for rgb-d object recognition," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1887–1898, 2015.

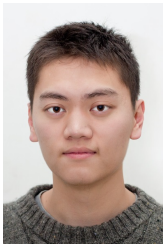
[54] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[55] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[56] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

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

[58] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.
[59] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
[60] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.
[61] 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.
[62] 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)*. IEEE, 2009, pp. 248–255.
[63] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.
[64] 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*, vol. 15, no. 1, pp. 1929–1958, 2014.
[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[66] C.-M. Kuan and K. Hornik, "Convergence of learning algorithms with constant learning rates," *IEEE Transactions on Neural Networks*, vol. 2, no. 5, pp. 484–489, 1991.
[67] W. Wu, N. Zhang, Z. Li, L. Li, and Y. Liu, "Convergence of gradient method with momentum for back-propagation neural networks," *Journal of Computational Mathematics*, pp. 613–623, 2008.
[68] V. Phansalkar and P. Sastry, "Analysis of the back-propagation algorithm with momentum," *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 505–506, 1994.
[69] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.
[70] H. J. Kushner, *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory*. MIT press, 1984, vol. 6.
[71] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.
[72] L. Le Cam *et al.*, "An approximation theorem for the poisson binomial distribution," *Pacific J. Math*, vol. 10, no. 4, pp. 1181–1197, 1960.
[73] 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)*, vol. 1. IEEE, 2001, pp. I–511.
[74] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.
[75] 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 (FGR06)*, 2006, pp. 341–345.
[76] "The fg-net aging database," http://www.fgnet.rsunit.com/.
[77] 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*, 2014, pp. 766–774.
[78] 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)*, 2013, pp. 2467–2474.
[79] 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)*. IEEE, 2011, pp. 585–592.



Shixing Chen received the BS degree from School of Telecommunications Engineering, Xidian University, P.R. China in 2013. He is currently a PhD candidate in the Machine Vision and Pattern Recognition Laboratory, Department of Computer Science, Wayne State University. His research interests include computer vision, machine learning, data mining, and multimedia analysis.

Caojin Zhang received his BS degree in Applied Statistics from School of Mathematics, Beijing Normal University, P.R. China, in 2013. He is currently working towards the PhD degree of Applied Math in Department of Mathematics, Wayne State University. His research interests lie in the areas of statistical models, stochastic control, machine learning and deep learning.

Ming Dong received his BS degree from Shanghai Jiao Tong University, Shanghai, P.R. China in 1995 and his PhD degree from the University of Cincinnati, Ohio, in 2001, both in electrical engineering. He is currently an associate professor of Computer Science and the Director of the Machine Vision and Pattern Recognition Laboratory. His research interests include pattern recognition, data mining, and multimedia analysis. His research is supported by the US federal and state governments, foundations, and industries. He has published over 100 technical articles, many in premium journals and conferences such as IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Multimedia, IEEE Trans. on Visualization and Computer Graphics, IEEE Trans. on Neural Networks, IEEE Trans. on Computers, IEEE Trans. on Knowledge and Data Engineering, IEEE CVPR, IEEE ICCV, IEEE ICDM, ACM Multimedia, and WWW. He was an associate editor of IEEE Trans. on Neural Networks, Pattern Analysis and Applications (Springer) and was on the editorial board of International Journal of Semantic Web and Information Systems. He currently serves as an associate editor of Smart Health journal (Elsevier) and as a program committee member for many related conferences. He is a member of the IEEE and ACM.