Ordinal Regression with Multiple Output CNN for Age Estimation

Zhenxing Niu[1] Mo Zhou[1] Le Wang[2] Xinbo Gao[1] Gang Hua[3]
[1]Xidian University [2]Xi'an Jiaotong University [3]Microsoft Research Asia
{zhenxingniu,cdluminate}@gmail.com, lewang@mail.xjtu.edu.cn, xinbogao@mail.xidian.edu.cn
ganghua@gmail.com

Abstract

To address the non-stationary property of aging patterns, age estimation can be cast as an ordinal regression *problem. However, the processes of extracting features and learning a regression model are often separated and optimized independently in previous work. In this paper, we propose an End-to-End learning approach to address ordinal regression problems using deep Convolutional Neural Network, which could simultaneously conduct feature learning and regression modeling. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems. And we propose a multiple output CNN learning algorithm to collectively solve these classification sub-problems, so that the correlation between these tasks could be explored. In addition, we publish an Asian Face Age Dataset (AFAD) containing more than* $160K$ *facial images with precise age ground-truths, which is the largest public age dataset to date. To the best of our knowledge, this is the first work to address ordinal regression problems by using CNN, and achieves the state-of-the-art performance on both the MORPH and AFAD datasets.*

1. Introduction

Human age estimation from face images remains to be an active research topic, which has many applications, such as demographics analysis, commercial user management, visual surveillance [22, 23, 24, 30, 3], and even aging progression [28]. In previous methods, age estimation is often cast as a multi-class classification [12][15] or a metric regression problem [11][15][14]. In a multi-class classification problem, the class labels are assumed to be independent to one another. However, the age labels have a strong ordinal relationship since they form a well-ordered set, which is not exploited in these multi-class classification methods.

On the other hand, metric regression approaches treat the age labels as numerical values to utilize such ordinal information for age estimation. However, the human face matures in different ways depending on the person's age [25]. For example, facial aging effects appear as changes in the shape of the face during childhood and changes in skin texture during adulthood. This property makes the random process formed by the aging patterns *non-stationary* in general. As manifested in [6], learning non-stationary kernels for a regression problem is usually difficult since it will easily cause over-fitting in the training process.

Due to the fact that facial aging process is a non-stationary process, one reliable information we can use would be the *relative order* among the age labels in addition to their exact values. And hence the age estimation is cast as a *ordinal regression* problem [4][6][32][20]. For instance, Cao et al. [4] formulated age estimation as a ranking problem and proposed a novel method based on Rank-SVM [17].

Recently, to directly utilize the well-studied classification algorithms, the ordinal regression problem is transformed into a series of simpler binary classification sub-problems [10][21]. For example, a reduction framework is proposed in [21]. For each rank $k \in \{1, 2, \cdots, K-1\}$ a binary classifier is trained according to whether the rank of a sample is larger than k. Then, the rank of a sample is predicted based on the classification results of the $K-1$ binary classifiers on this sample. In [21], the well-tuned SVM classification algorithm is directly utilized to train those binary classifiers. A benefit of this kind of methods is that new generalization bounds for ordinal regression can be easily derived from known bounds for binary classification.

Inspired by it, we also transform ordinal regression as a series of binary classification sub-problems in this paper. In particular, the Convolutional Neural Network (CNN) is used to solve those binary classification sub-problems. Moreover, our CNN has multiple output layers where each output layer corresponds to a binary classification task, called *Multiple Output CNN* in this paper. Therefore, all the binary classifiers can be jointly trained in such a CNN. Since all the tasks share the same mid-level representations in the CNN, the correlation among distinct tasks could be explored, which is beneficial to improving the finial performance.

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.532

On the other hand, for either the metric regression based or the ordinal regression based approaches, the processes of extracting features and learning a regression model is separated and optimized independently. The most successful hand-crafted features for age estimation is the Bio-inspired Features (BIFs) [15]. Nevertheless, due to the unclear mechanism of how humane perceives the different aging pattern, it is still difficult to design good features for age estimation. On the contrary, in our approach we can conduct an *End-to-End learning* with CNN for age estimation, which could simultaneously optimize feature learning and regression modeling. As a result, we can automatically learn better features from facial images, and avoid directly designing hand-crafted features.

At last, the lack of a large scale age dataset is always a barrier for pushing the progress of research on age estimation. And most previous algorithmic evaluations were performed on relatively small datasets, mainly due to the difficulties in collecting a large dataset with precise human age ground-truths. The most popular public age datasets include FG-NET [1] (1002 face images), MORPH I (1690 face images), and MORPH II [26] ($55,608$ face images).

Even for the largest dataset MORPH II, its ethnic is very unbalanced, *i.e.*, more than 96% faces are African and European, but less than 1% faces from Asian. Thus, the performance of previous methods for age estimation on Asian faces is still unknown. Besides, a large scale age dataset is essential for introducing deep learning algorithm such as CNN to age estimation. In this paper, we publish a new age dataset called **Asian Face Age Dataset (AFAD)**, which includes more than $160K$ Asian facial images and age labels. Until now, this is the largest public age dataset. Briefly, our contributions are:

1. We propose to address ordinal regression problems using End-to-End deep learning methods.
2. We apply it to the task of age estimation, and achieve state-of-the-art results.
3. A new age dataset is released to the community for age estimation, which is the largest public dataset to date.

2. Related Work

Age estimation: Most existing methods estimate the age of face image by two steps: local feature extraction and metric regression (or multi-class classification). Due to that geometry and texture features are helpful to distinguish baby, adult, and senior, many methods were proposed based on the AAM model [8] since it is a natural tool to simultaneously model the shape and texture of facial images [19][12]. The most successful hand-crafted features for age estimation is the Bio-inspired Features (BIFs) [15]. On the other hand, much attention were paid on the second step: age classification or regression based on these features. Fu and Huang [11] applied discriminative manifold learning and quadratic regression to age estimation. Guo *et al*. introduced many regression methods to predict the age of face image, such as SVR [15], PLS [13] and CCA [14].

Recently, to better handle the non-stationary property of the aging process, ordinal regression is employed for age estimation. Yang *et al*. [32] employ the RankBoost algorithm, which is a single hyperplane ranker in the feature space, for age estimation. Chang *et al*. [5] employ the parallel hyperplanes model OR-SVM [21] for age estimation, and further extended it to a more flexible scenario, *i.e.*, several possibly non-parallel hyperplanes [6].

Although deep learning has achieved success on many computer vision tasks (*e.g.*, image recognition, and object detection, etc), there are little works for introducing CNN to age estimation. Yi *et al*. [33] firstly proposed to use CNN for age estimation. However, the proposed CNN is very shallow, which only contains 4 layers (*i.e.*, 1 convolution + 1 pooling + 1 local layer + 1 full connection), and only a subset of MORPH II (about 10K facial images) is used to train the shallow CNN. Currently, in [31] a relatively deeper CNN has been proposed for age estimation. However, in their work, CNN is only used to extract features, which is then fed to another regressor for the final age estimation. But our approach conducts End-to-End learning which could better unleash the discriminative power of the CNN.

Ordinal regression: Most ordinal regression algorithms are modified from well-known classification algorithms [16][9][27]. For instance, Herbrich *et al*. [16] proposed a method of support vector learning for ordinal regression. In [9], the perceptron ranking (PRank) algorithm proposed by Crammer and Singer is to generalize the online perceptron algorithm with multiple thresholds for ordinal regression. In [27], Shashua and Levin proposed new support vector machine formulations to handle multiple thresholds.

On the other hand, to directly utilize the well-studied classification algorithms, the ordinal regression problem is transferred as a series of simpler binary classification subproblems [10][21]. For example, several decision trees are employed as binary classifiers for ordinal regression in [10]. Recently, Li *et al*. [21] proposed a framework to reduce an ordinal regression problem as a set of classification problems, and employ an SVM to solve the classification problems.

3. Ordinal Regression with CNN

3.1. Problem Formulation

Let us assume that the i-th image is represented in an input space $\boldsymbol{x}_i \in \mathcal{X}$, and there is an outcome space $y_i \in$

$\mathcal{Y} = \{r_1, r_2, \cdots, r_K\}$ with ordered ranks $r_K \succ r_{K-1} \succ \cdots \succ r_1$. The symbol $\succ$ denotes the ordering between different ranks. Given training samples $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, the ordinal regression is to find a mapping from images to ranks $h(\cdot) : \mathcal{X} \to \mathcal{Y}$ such that — using a predefined cost $c : \mathcal{X} \times \mathcal{Y} \to R$ — the risk functional $R(h)$ is minimized.

In this paper, the *cost matrix* $\mathcal{C}$ [21] is employed to measure the cost between predicted ranks and ground-truth ranks. In particular, $\mathcal{C}$ is a $K \times K$ matrix with $\mathcal{C}_{y,r}$ being the cost of predicting an example $(\boldsymbol{x}, y)$ as rank r. Naturally, it is assumed that $\mathcal{C}_{y,y} = 0$ and $\mathcal{C}_{y,r} > 0$ for $r \neq y$. Particularly, the *absolute cost matrix*, which is defined by $\mathcal{C}_{y,r} = |y - r|$, is a popular choice for general ordinal regression problems. Particularly, each age is often treated as a rank when applying ordinal regression algorithms to age estimation.

3.2. Our Approach

To directly utilize the well-studied classification algorithms, we transform ordinal regression as a series of binary classification sub-problems in this paper. In particular, an ordinal regression problem with K ranks is transformed into $K - 1$ simpler binary classification sub-problems. For each rank $r_k \in \{r_1, r_2, \cdots, r_{K-1}\}$, a binary classifier is constructed to predict whether the rank of a sample y_i is larger than r_k. And then the rank of an unseen sample is predicted based on the classification results of the $K - 1$ binary classifiers on this sample.

Specifically, our approach contains three steps: (a) given the original training data $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, for the k-th binary classification sub-problem a specific training data is constructed as $D^k = \{\boldsymbol{x}_i, y_i^k, w_i^k\}_{i=1}^N$, where the $y_i^k \in \{0, 1\}$ is a binary class label indicating whether the rank of the i-th sample y_i is larger than r_k as follows,

$$y_i^k = \begin{cases} 1, & \text{if}(y_i > r_k) \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

And the w_i^k is the weight for the i-th example, *i.e.*,

$$w_i^k = |\mathcal{C}_{y,k} - \mathcal{C}_{y,k+1}|. \tag{2}$$

Since absolute cost matrix is adopted in our approach, we have $\forall(i, k), w_i^k = 1$. (b) the $K - 1$ binary classifiers are trained with their corresponding training data. It is noticed that we adopt one CNN to collectively implement these binary classifiers in our approach. In particular, our CNN has a multiple-output structure where each output corresponds to a binary classifier. Thus, these binary classifiers are jointly trained in such a CNN (refers to Sec.3.4); (c) the rank for an unseen sample $\boldsymbol{x'}$ is predicted as follows,

$$\begin{aligned} h(\boldsymbol{x'}) &= r_q \\ q &= 1 + \sum_{k=1}^{K-1} f_k(\boldsymbol{x'}), \end{aligned} \tag{3}$$

where $f_k(\boldsymbol{x'}) \in \{0, 1\}$ is the classification result of the k-th binary classifier for the sample $\boldsymbol{x'}$ (*i.e.*, the k-th output of our multiple-output CNN). Ideally, these $f_k(\boldsymbol{x'})$ should be consistent. However, ensuring the consistency in the training phase would significantly increase the training complexity. Hence we just apply Eq. 3 without explicitly ensuring the consistency among the different classifiers as in [21].

The benefits of our approach is two-fold: (1) an ordinal regression problem is solved by using an End-to-End deep learning method, so that we can automatically learn better features from facial images and avoid directly designing hand-crafted features. (2) the $K - 1$ classification sub-problems are treated as $K - 1$ tasks, which are simultaneously solved with our multiple output CNN. Due to that all the tasks share the same mid-level representations in such a CNN, the correlation of distinct tasks could be explored, which is beneficial to improve the finial performance.

Input: training data $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ and testing images $D' = \{\boldsymbol{x}'_j\}_{j=1}^M$

- Loop for $k = 1, 2, \cdots, K - 1$:
  - Build a distinct training data $D^k = \{\boldsymbol{x}_i, y_i^k, w_i^k\}_{i=1}^N$ for the k-th binary classification task according to Eq.1 and Eq.2.
- The learning of the multiple output CNN:
  - The proposed multiple output CNN is trained with $D^k, (k = 1, 2, \cdots, K - 1)$ according to the proposed learning algorithm (refers to Sec.3.4).
- For each testing image $\boldsymbol{x}'_j \in D'$:
  - Forward $\boldsymbol{x}'_j$ to the trained CNN, and get the $K - 1$ binary labels $f_k(\boldsymbol{x}'_j) \in \{0, 1\}, (t = 1, 2, \cdots, K - 1)$;
  - Predict its rank $h(\boldsymbol{x}'_j)$ with previous binary labels $\{f_k(\boldsymbol{x}'_j)\}_{k=1}^{K-1}$ according to Eq.3.

Output: the predicted ranks for testing images $\{h(\boldsymbol{x}'_j)\}_{j=1}^M$.

Figure 1. Our approach for solving ordinal regression with the multiple output CNN.

3.3. Architecture of the Multiple Output CNN

As shown in Fig.2, our network consists of 3 convolutional, 3 local response normalization, and 2 max pooling layers followed by a fully connected layer with 80 neurons.

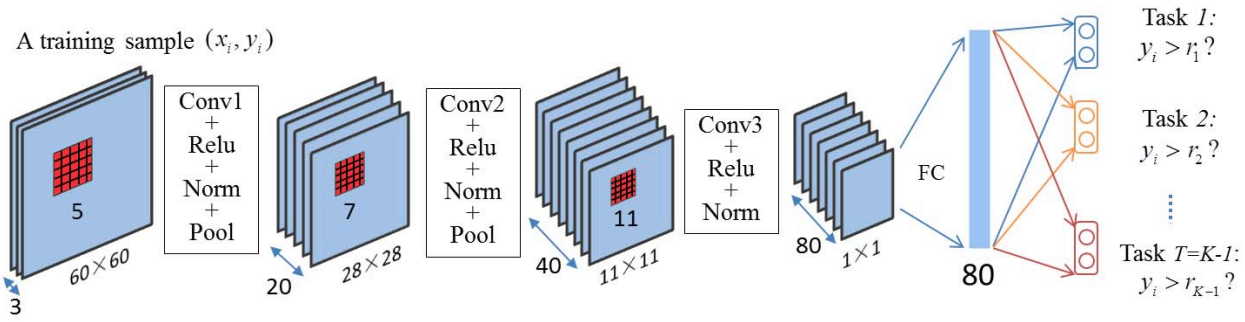


Figure 2. The architecture of the proposed Multiple Output CNN

At the input level, aligned face images of size $60 \times 60 \times 3$ are fed to the network as input. It is noted that color face images are used in this paper, which is different from the gray images used in [31]. At the first convolutional layer, 20 kernels of size $5 \times 5 \times 3$ with stride of 1 pixels is applied on the input images. And after local response normalization and max pooling operations, the feature maps of size $28 \times 28 \times 20$ are obtained.

The similar operations are conducted at the second and third convolutional layers with different kernel size (refers Fig.2 for the details). And then a fully connected layer with 80 neurons is used to generate a mid-level representation.

After that, the network branches out $K-1$ output layers, where each output layer contains 2 neurons and corresponds to a binary classification task. The k-th task is to predict whether the age of the i-th facial image is larger than the rank r_k. For each task, the softmax normalized cross entropy loss is employed as loss function.

3.4. Learning the Multiple Output CNN

For a CNN with single output, we have N samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i$ denotes the i-th image and y_i denotes the corresponding class label. For binary class label $y_i \in \{0, 1\}$, it is reasonable to employ cross-entropy as the loss function,

$$E_s = -\frac{1}{N}\sum_{i=1}^{N} 1\{o_i = y_i\} w_i log(p(o_i|\boldsymbol{x}_i, \boldsymbol{W})), \quad (4)$$

where o_i indicates the output of the CNN for the i-th image, w_i indicates the weight of the i-th image, and $\boldsymbol{W}$ indicates the parameters of the entire CNN. Let $\boldsymbol{W}_l$ denotes the parameters of the l-th layer in the CNN, and hence we have $\boldsymbol{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_{L-1}\}$. The Boolean test $1\{\cdot\}$ is 1 if the inner condition is true, and 0 otherwise.

For a CNN with $K-1$ outputs, each output corresponds to a distinct task. All the $T = K-1$ tasks (outputs) share the same N input images $\{\boldsymbol{x}_i\}_{i=1}^N$, but have different class

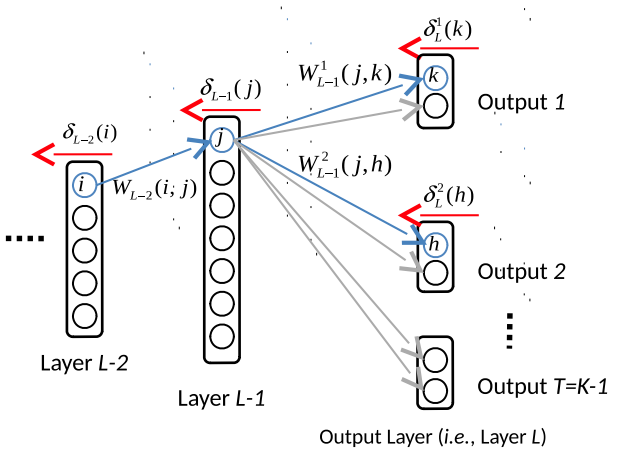


Figure 3. The back-propagation procedure for our Multiple Output CNN

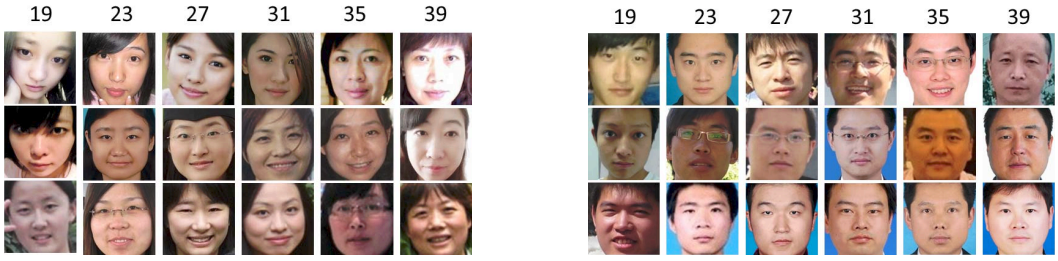
labels $\{\{y_i^t\}_{i=1}^N\}_{t=1}^T$. Let λ_t denotes the importance coefficient of the t-th task, the loss function of our multiple output CNN can be written as

$$E_m = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \lambda^t 1\{o_i^t = y_i^t\} w_i^t log(p(o_i^t|\boldsymbol{x}_i, \boldsymbol{W}^t)), \quad (5)$$

where o_i^t indicates the output of the t-th task for the i-th image, w_i^t indicates the weight of the i-th image for the t-th task, and $\boldsymbol{W}^t$ indicate the parameters of the t-th task.

According to the architecture of our multiple output CNN, each task has a distinct output layer, but all the tasks share the same intermediate layers. Thus, only the parameters of the output layers for distinct tasks $\boldsymbol{W}_{L-1}^t$ are different from each other, *i.e.*, we have $\boldsymbol{W}_{L-1} = \{\boldsymbol{W}_{L-1}^1, \boldsymbol{W}_{L-1}^2, \cdots, \boldsymbol{W}_{L-1}^T\}$. And the parameters of previous layers are the same as one another for all the tasks, *i.e.*, $\forall l \in \{1, \cdots, (L-2)\}$, $\boldsymbol{W}_l^1 = \boldsymbol{W}_l^2 = \cdots = \boldsymbol{W}_l^T = \boldsymbol{W}_l$.

As shown in Fig. 3, the learning procedure of the parameters for the output layer $\boldsymbol{W}_{L-1} = \{\boldsymbol{W}_{L-1}^t\}_{t=1}^T$ is similar to that of a single output CNN. For the t-th task, if the



(a) female (b) male

Figure 4. Some samples from our AFAD dataset including for 19-year, 23-year, 27-year, 31-year, 35-year, and 39-year old facial photos for (a) female and (b) male.

cross-entropy loss function is employed, the gradient of the weight from the j-th neuron in the layer $L-1$ to the k-th neuron in the layer L (*i.e.*, $W^t_{L-1}(j,k)$) is computed as,

$$\frac{\partial E_m}{\partial W^t_{L-1}(j,k)} = \delta^t_L(k)o(j) \quad (6)$$

$$\delta^t_L(k) = p(o^t(k)|\boldsymbol{x}_k, \boldsymbol{W}^t_{L-1}) - 1\{o^t = y^t\}, \quad (7)$$

where $o(j)$ is the output of the j-th neuron in the layer $L-1$, and $\delta^t_L(k)$ is the error of the k-th neuron in the output layer.

The key is the learning procedure of the parameters for the penultimate layer. Since each neuron in the penultimate layer (*i.e.*, layer $L-1$) is connected to all the neurons in the output layer (*i.e.*, layer L), the error of the penultimate layer δ_{L-1} is the integration of all the errors of output layers, as shown in Fig. 3. Specifically, the gradient of the weight from the i-th neuron in the layer $L-2$ to the j-th neuron in the layer $L-1$ (*i.e.*, $W_{L-2}(i,j)$) is computed as,

$$\frac{\partial E_m}{\partial W_{L-2}(i,j)} = \delta_{L-1}(k)o(i) \quad (8)$$

$$\delta_{L-1}(j) = \sum_{t=1}^{T} \lambda^t \Big(\sum_{k \in L^t} \delta^t_L(k) W^t_{L-1}(j,k) \Big), \quad (9)$$

where $o(i)$ is the output of i-th neuron in the layer $L-2$, and $\delta_{L-1}(j)$ is the error of the j-th neuron in the layer $L-1$. It is noticed that $\delta_{L-1}(j)$ is the weighted sum of the errors of output layers $\delta^t_L(k)$ over all the tasks, where the task-specific weights are $\lambda_t, (t = 1, \cdots, T)$.

The learning procedure of the weights for previous layers (*i.e.*, $W_{L-3}, W_{L-4}, \cdots$) is the same as the standard learning procedure of CNN.

4. The AFAD Dataset

We tested our method on the MORPH II Dataset [26]. This dataset contains $55,608$ face images, including $42,589$ African faces (77%), $10,559$ European faces (19%), $1,769$ Hispanic faces (3%), and only 154 Asian faces (0.2%).

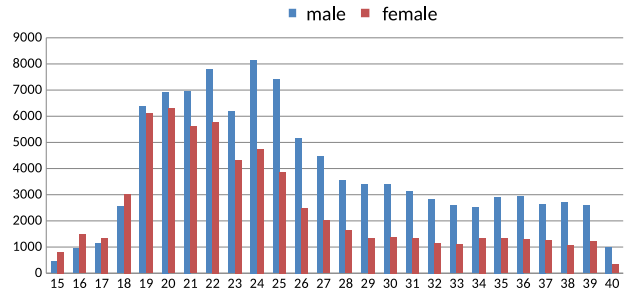


Figure 5. The distribution of age and gender for the AFAD dataset. There are $164,432$ well-labeled photos, including $63,680$ photos for female as well as $100,752$ photos for male. And the ages range from 15 to 40.

Ages range from 16 to 77 with a median age of 33. The average number of images per individual is 4.

Compared to other datasets such as FG-NET [1] (1002 face images) and MORPH I (1690 face images) dataset, the MORPH II dataset is the largest public dataset for age estimation. However, its ethnic is very unbalanced, *i.e.*, less than 1% Asian faces. Thus, the performance of previous methods for age estimation on Asian faces is not sufficiently studied. Therefore, in this paper we collect a new dataset called **Asian Face Age Dataset (AFAD)** for age estimation, which includes more than 160k images and aging labels for Asian.

Specifically, we build this dataset by collecting facial images on a particular social network, *i.e.*, RenRen Social Network (RSN) [2]. The RSN is a social network for students to connect with each other, upload photos, and make comments, etc. It is widely used by Asian students including middle school, high school, undergraduate, and graduate students. Even after leaving from school, some people still access their RSN account to connect with their old classmates. So, the age of the RSN user crosses a wide range from 15- to more than 40-years old, which is beneficial to building a dataset with a wide aging range.

Moreover, as a user creating an account on the RSN, he/she is required to provide his/her Data-of-Birth (DoB),

Table 1. The comparison of age estimation between metric regression and ordinal regression methods. The performance is measured by the Mean Absolute Error (MAE) metric.

Dataset	Metric Regression			Ordinal Regression		
	BIFs + LSVR [15]	BIFs + CCA [14]	CNN + LSVR [31]	BIFS + OR-SVM [5]	BIFS + OHRank [6]	Ours (OR-CNN)
MORPH II	4.31	4.73	5.13 (4.77 in [31])	4.21	3.82	**3.27**
AFAD	4.13	4.40	5.56	4.36	3.84	**3.34**

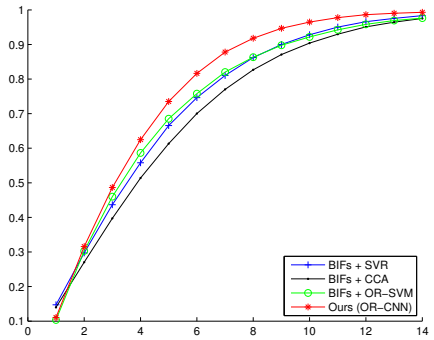


Figure 6. The comparison of age estimation with CS metric on MORPH II dataset

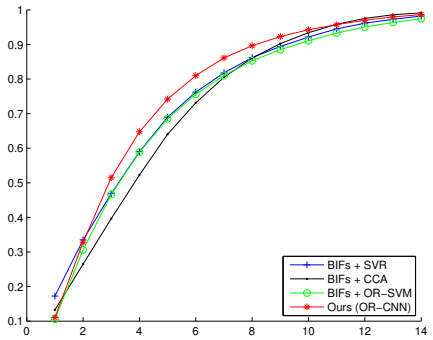


Figure 7. The comparison of age estimation with CS metric on AFAD dataset

which is utilized by the RSN to recommend classmates or friends to the user. In addition, there is a special photo album (*i.e.*, selfie album) for each user to upload his/her selfie photo to the RSN. Therefore, it is easy to get the ground-truth age of each selfie photo, *i.e.*, the difference between the uploading date of the selfie photo and the user's DoB.

Since the selfie photos are noisy, *e.g.*, some user may upload other photo as his/her selfie photo (*e.g.*, uploading a photo of a celebrity, an object, or even a landmark) or upload a photo taken long time ago, thus we employ some workers to *manually* filter out noisy data. At last, we collect a dataset with $164,432$ well-labeled photos. It consist of $63,680$ photos for female as well as $100,752$ photos for male, and the ages range from 15 to 40. The distribution of photo counts for distinct ages are illustrated in Fig. 5. Some samples are shown in Fig. 4.

5. Experiments

5.1. Preprocessing and Experimental Setting

The preprocessing of the facial image dataset is necessary. First, all images in the datasets are processed by a face detector [29]. Then the facial landmarks of face images are localized by AAM [8], and all the facial images are aligned that the two eyeballs stay at the same image position for all faces. After that, a region with the size of 64×64 is cropped from the aligned facial images to make the nose point stays at the center of the cropped images. Moreover, before being fed to the CNN the images patches with the size of 60×60 are randomly cropped from the 64×64 regions, which slightly reduced overfitting when training our CNN.

Following the experimental setting in [6][7][31], for both MORPH II and AFAD datasets, we randomly divide the whole dataset into two parts: one part (*i.e.*, 80% of the whole data) is used for training, and the other one (*i.e.*, 20% of the whole data) is used for testing. There is no overlap between the training and testing data. For statistical analysis, this procedure is done 100 times to evaluate the variance of MAE.

Besides, our algorithm is implemented based on Caffe [18], where a new layer is implemented according to Eq. 8 and Eq. 9. It is noted that the parameters for task importance $\lambda_t, (t = 1, \cdots, T)$ are set according to Eq. 11, which will be discussed in details at Section. 5.4.1

5.2. Age Estimation

In this section, the performance of age estimation is compared among several methods including metric regression based and ordinal regression based methods. For each kind of methods, they can be further categorized into two subcategories based on whether they used CNN learning algorithm.

The performance is measured by the Mean Absolute Error (MAE) metric and the Cumulative Score (CS). The MAE is calculated using the average of the absolute errors between the estimated result and the ground truth. The cu-

mulative score is calculated as follows:

$$CS(n) = -\frac{K_n}{K} \times 100\%, \tag{10}$$

where K is the total number of test images, and K_n is the number of testing facial images whose absolute error between the estimated age and the ground truth age is not greater than n years.

In particular, we compare with the method in [15] (denoted as 'BIFs + LSVR') which trains a linear Support Vector Regression (SVR) on the extracted BIFs features. For extracting BIFs features, we adopt the similar setting of 8 bands and 4 orientations, which produces features with more than 4000 dimensions. And the Liblinear software is used to train the linear SVR, where the L2-regularized L1-loss is adopted.

We also compare with the method in [14] (denoted as 'BIFs + CCA'). This is a multi-task learning approach which can simultaneously predict the age, gender, and race of a facial image. We evaluate this method to conduct the 3 tasks on the MORPH II dataset. Since the AFAD dataset only contains Asian faces, we evaluate this method to conduct only 2 tasks (*i.e.*, age estimation and gender classification) on our AFAD dataset. From Table 1, we can see that the 'BIFs + LSVR' achieves better performance than 'BIFs + CCA' on both datasets.

The third method [31] (denoted as 'CNN + LSVR') is the first work to develop a relatively deeper CNN (*i.e.*, 3 convolution + 2 pooling + full connection) to address the problem of age estimation. However, the proposed CNN is only used to extract features, which is then fed to a regressor (*i.e.*, a linear SVR regressor) for final age prediction. Thus, it is not an End-to-End learning method. We have re-implemented the method in [31], but find out that its MAE on MORPH is 5.13 instead of 4.77 reported in [31], which may be due to some differences in details.

We also compare with two typical ordinal regression based methods, denoted as the 'BIFs + OR-SVM' [5] and the 'BIFs + OHRank' [6] respectively. Both of them transform an ordinal regression problem as a series of binary classification sub-problems, which are solved by using a linear SVM. It is noted that the hyperplanes are restricted to parallel to each other in [5] but the OHRank [6] allows some possibly non-parallel hyperplanes.

From Table 1, we have the following conclusions: (1) the ordinal regression based methods outperform the metric regression based methods in general; (2) more importantly, the integration of ordinal regression and deep learning methods could boost the performance significantly. And our approach achieves the state-of-the-art on both MORPH II and AFAD datasets.

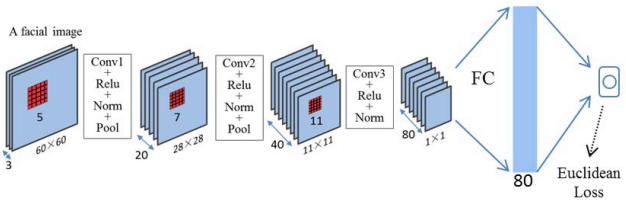


Figure 8. The architecture of network used in the method of Metric Regression with CNN (MR-CNN).

Table 2. The comparisons between metric regression and ordinal regression based methods. The performance is measured by the Mean Absolute Error (MAE) metric.

Methods	MORPH II	AFAD
BIFs + OHRank [6]	3.82	3.84
Proposed MR-CNN	3.42	3.51
Proposed OR-CNN	**3.27**	**3.34**

5.3. Comparing Metric and Ordinal Regression

Our approach presents an End-to-End CNN learning method, which transforms ordinal regression into a series of binary classification sub-problems. So it is still unclear which factor gives more contributions to the final improvement of performance. To take apart their distinctive contributions, we propose another baseline method, which only keeps the End-to-End CNN learning part and drops the part of transforming framework, *i.e.*, it casts age estimation as a metric regression problem instead of an ordinal regression problem, and addresses the metric regression problem with an End-to-End CNN learning algorithm. For clarity, the baseline method is called Metric Regression with CNN (MR-CNN), while the proposed approach is called Ordinal Regression with CNN (OR-CNN).

For the MR-CNN, we need to solve a general metric regression problem with CNN, so we make a little modification on the architecture of our multiple-output CNN. Specifically, the fully connected layer is directly connected to one (the only one) output layer, and previous layers (*i.e.*, three convolutional, local response normalization, max pooling layers) are kept. The output layer contains only one neurons, and its output indicates the regressed age for the input facial image. To train the MR-CNN, the L2-norm loss between ground-truth age and the regressed age is employed as the loss function. The architecture of the MR-CNN is shown in Fig.8.

The comparison between MR-CNN and OR-CNN is conducted on both the MORPH II and AFAD datasets. From Table 2, we can find that the OR-CNN outperforms MR-CNN on the two datasets. In addition, we have done some statistical analysis for our OR-CNN model. In particular, the dataset is randomly split into training and

Table 3. The analysis of the task importance. The performance is measured by the Mean Absolute Error (MAE) metric.

Datasets	Uniform	Data-specific Scheme
MORPH II	3.30	**3.27**
AFAD	3.41	**3.34**

Table 4. The analysis of the image color information. The performance is measured by the Mean Absolute Error (MAE) metric.

Datasets	Gray Image	Color Image
MORPH II	3.42	**3.27**
AFAD	3.44	**3.34**

testing subsets 100 times for cross-validation. We obtain MAE=$\mathbf{3.27 \pm 0.02}$ and MAE=$\mathbf{3.34 \pm 0.08}$ on the MORPH-II and AFAD dataset respectively. Compared with the *OHRank*, our analysis shows that the proposed algorithm significantly outperform it at significance level $\alpha = 0.01$.

Those promising results illustrate that (1) it is necessary to cast age estimation as ordinal regression rather than a general metric regression problem. Moreover, from the comparison between the 'BIFs + OHRank' and our 'OR-CNN', it is clear that (2) the integration of ordinal regression and deep learning methods could significantly boost the performance.

5.4. Discussion

5.4.1 Task importance analysis

In our approach, each task has a specific parameter λ_t, which indicates the relative importance among different tasks. These parameters will affect the error backward propagation for the network training (refers to Eq.9). In our approach, the importance parameters are set according to the reliability of classifiers corresponding to different tasks.

In particular, the k-th task corresponds to a binary classifier, which is trained to distinguish samples with rank larger than k from samples with rank smaller than k. Thus, for the k-th classifier, the number of samples with ranks nearby k, *e.g.*, samples with rank $\{(k-1), k, (k+1)\}$, is more important than other samples for the training of k-th classifier. In other words, if more samples are with rank close to k, we could better train the corresponding classifier, and hence it is better to give a relatively larger importance to this task.

In practice, we obtain the distribution of sample number over their ranks, and propose to set the importance parameters according to this distribution, which is called data-specific scheme. In other words, we set

$$\lambda_t = \frac{\sqrt{N_k}}{\sum_{k=1}^{K} \sqrt{N_k}}, \tag{11}$$

where N_k is the number of samples with rank k, and N is the total number of samples.

On the contrary, we also evaluate the performance of our method when the importance parameters are set as a constant, which is called uniform scheme in this paper. From Table 3, we can find out that the performance could be slightly improved by choosing task importance parameters according to data-specific distributions.

5.4.2 The color information

It is noticed that color face images are directly fed into our CNN, which is different from previous methods that the color images are first converted into gray before fed into CNN [31][33]. In previous methods, it is believed that color information is unstable and useless for age estimation [33]. To investigate whether the color information is helpful for age estimation, in this section we try the case of converting color images into gray images. The performance is compared on both MORPH II and AFAD datasets.

From Table 4, we can find out that color information is helpful to improve performance of age estimation. It is possible that our convolutional neural network managed to extract useful information from color images.

6. Conclusion

In this paper, we have proposed to address ordinal regression problem by using End-to-End deep learning methods. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems, which are collectively solved with the proposed multiple output CNN learning algorithm. We apply it to the task of age estimation, and achieve better performance avoiding directly designing hand-crafted features. In addition, we publish an Asian Face Age Dataset (AFAD), which is the largest public age dataset until now. Our approach is evaluated on two large scale benchmark datasets and outperforms the state-of-the-art by a large margin. The promising performance of our approach demonstrates the potential of applying it towards other ordinal regression related applications.

7. Acknowledgements

This research was supported partly by the NSFC (Grant Nos. 61432014 and 61402348, and 61503296), the Fundamental Research Funds for the Central Universities (Grant Nos. BDZ021403 and XJJ2015066), the Program for Changjiang Scholars and Innovative Research Team in University of China (No.IRT13088), the Shaanxi Innovative Research Team for Key Science and Technology (No.2012KCT-02), and China Postdoctoral Science Foundation (Grant Nos.2014M562374 and 2015M572563). Dr. Gang Hua is partly supported by NSFC Grant No.61228303.

References

[1] The fg-net aging database. *http://sting.cycollege.ac.cy/alanitis/fgnetaging.html*.

[2] Renren social network. *http://www.renren.com/*.

[3] J. Cai, Z. Zha, W. Zhou, and Q. Tian. Attribute-assisted reranking for web image retrieval. *ACM Multimedia*, 2012.

[4] D. Cao, Z. Lei, Z. Zhang, J. Feng, and S. Li. Human age estimation using ranking svm. *CCBR*, pages 324–331, 2012.

[5] K. Chang, C. Chen, and Y. Hung. A ranking approach for human age estimation based on face images. *ICPR*, 2010.

[6] K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. *CVPR*, pages 585–592, 2011.

[7] K. Chen, S. Gong, T. Xiang, and C. Loy. Cumulative attribute space for age and crowd density estimation. *CVPR*, pages 2467–2474, 2013.

[8] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *ECCV*, pages 484–498, 1998.

[9] K. Crammer and Y. Singer. Pranking with ranking. *NIPS*, pages 641–647, 2002.

[10] E. Frank and M. Hall. A simple approach to ordinal classification. *Lecture Notes in Artificial Intelligence*, pages 145–156, 2001.

[11] Y. Fu and T. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, pages 578–584, 2008.

[12] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE T-PAMI*, pages 2234–2240, 2007.

[13] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. *CVPR*, pages 657–664, 2011.

[14] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. *FG*, pages 1–6, 2013.

[15] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. *CVPR*, pages 112–119, 2009.

[16] R. Herbrich, T. Graepel, and K. Obermayer. support vector learning for ordinal regression. *Proc. Int. Conf. Artif. Neural Netw*, pages 97–102, 1999.

[17] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *MIT Press, Cambridge*, 2000.

[18] 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.

[19] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE SMC-B*, pages 621–628, 2004.

[20] C. Li, Q. Liu, J. Liu, and H. Lu. Learning ordinal discriminative features for age estimation. *CVPR*, pages 2570–2577, 2012.

[21] L. Li and H. Lin. Ordinal regression by extended binary classification. *NIPS*, pages 865–872, 2006.

[22] Z. Niu, G. Hua, X. Gao, and Q.Tian. Context-aware topic model for scene recognition. *CVPR*, 2012.

[23] Z. Niu, G. Hua, X. Gao, and Q.Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. *CVPR*, 2014.

[24] Z. Niu, G. Hua, Q.Tian, and X. Gao. Visual topic network: Building better image representations for images in social media. *CVIU*, 2015.

[25] N. Ramanathan, R. Chellappa, and S. Biswas. Age progression in human faces: A survey. *JVLC*, 2009.

[26] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2015.

[27] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. *NIPS*, pages 961–968, 2003.

[28] I. Shlizerman, S. Suwajanakorn, and S. Seitz. Illumination-aware age progression. *CVPR*, 2014.

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001.

[30] L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue. *ICCV*, 2011.

[31] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. *WACV*, pages 534–541, 2015.

[32] P. Yang, L. Zhong, and D. Metaxas. Ranking model for facial age estimation. *ICPR*, 2010.

[33] D. Yi, Z. Lei, and S. Li. Age estimation by multi-scale convolutional network. *ACCV*, pages 144–158, 2014.