Moving Window Regression: A Novel Approach to Ordinal Regression

Nyeong-Ho Shin
Korea University
nhshin@mcl.korea.ac.kr

Seon-Ho Lee
Korea University
seonholee@mcl.korea.ac.kr

Chang-Su Kim
Korea University
changsukim@korea.ac.kr

Abstract

A novel ordinal regression algorithm, called moving window regression (MWR), is proposed in this paper. First, we propose the notion of relative rank (ρ-rank), which is a new order representation scheme for input and reference instances. Second, we develop global and local relative regressors (ρ-regressors) to predict ρ-ranks within entire and specific rank ranges, respectively. Third, we refine an initial rank estimate iteratively by selecting two reference instances to form a search window and then estimating the ρ-rank within the window. Extensive experiments results show that the proposed algorithm achieves the state-of-the-art performances on various benchmark datasets for facial age estimation and historical color image classification. The codes are available at https://github.com/nhshin-mcl/MWR.

1. Introduction

Ordinal regression aims to predict the rank of an object instance. It is widely used for computer vision tasks, including facial age estimation [35] and historical color image (HCI) classification [34]. Thus, various ordinal regression techniques have been developed.

Rank estimation, however, remains challenging because there is no clear distinction between ranks in many cases. For example, in facial age estimation, the aging process [2, 46], causing variations in facial shapes, sizes, and texture, has large individual differences due to numerous factors such as genes, diet, and lifestyle, and there are no clear aging characteristics in each age class. To address this issue, extensive researches have been carried out. Recently, Li *et al*. [26] used tens of local regressors, each of which learns aging characteristics within a specific age range. Also, Lim *et al*. [27] proposed the notion of order learning, and Lee and Kim [22] improved it based on the order-identity decomposition.

To measure the quantity of something, it would be easier and more accurate if some references are available [4]. For example, we can estimate the length of an object more accurately if we have a one-meter bar as a reference. In this case, our brains measure the ratio between the lengths of the object and the bar, instead of the absolute length of the object. Similarly, multiple references of different ranks can offer a useful ordinal scale in rank estimation, as a one-meter bar does in the length measurement. Therefore, rather than predicting the absolute rank of an instance directly, we attempt to estimate the relative rank that quantifies how much greater or smaller the instance is than the references.

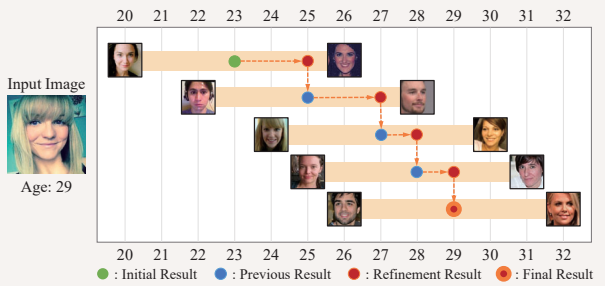


Figure 1. In the MWR algorithm for facial age estimation, an initial rank (age) estimate of the input image is refined iteratively by estimating the ρ-rank within a search window (orange bar). The window is bounded by the known ranks of two reference images, and the ρ-rank quantifies how much older or younger the input is than the references. Each window is centered around the previous result. It is strongly recommended to watch the accompanying video for an overview of MWR. Also, note that MWR is applicable to general ordinal regression tasks, as well as age estimation.

Based on the observations, we propose a novel ordinal regression algorithm, called moving window regression (MWR), which is illustrated in Figure 1. First, we propose the notion of relative rank (ρ-rank), which quantifies the ordinal relations among the ranks of input and reference instances: the ρ-rank measures how much greater the input is than the first reference and how much smaller it is than the second one. Second, to estimate the ρ-rank, we develop a relative regressor (ρ-regressor), composed of an encoder and a regression module. Third, we propose the MWR process in Figure 1. It obtains an initial rank estimate of an input instance based on the nearest neighbor (NN) criterion. Then, it refines the estimate iteratively by selecting two ref-

erence instances to form a search window and estimating the ρ-rank within the search window. Here, each window is centered around the previous estimate. This is iterated until the convergence. Last, to cope with diverse characteristics in different rank groups, we develop local ρ-regressors for those groups, as well as the global ρ-regressor for the entire rank range.

This work has the following major contributions:

- We first propose the notion of ρ-rank and design ρ-regressors for estimating ρ-ranks.
- We develop the novel MWR algorithm for accurate rank estimation, which iteratively predicts the ρ-rank within a moving search window.
- MWR provides the state-of-the-art performances on various benchmark datasets for facial age estimation and HCI classification. Especially, for facial age estimation, MWR performs the best in 17 out of 19 benchmark tests.

2. Related Work

2.1. Ordinal Regression

In ordinal regression, the rank of an object is estimated. In many cases, ordinal regression is converted into multiple binary classification problems [15, 24]. Liu *et al*. [30] developed a deep ordinal regressor for small datasets. They also proposed another ordinal regressor [31] to adopt the multi-class classification loss additionally. Fu *et al*. [16] addressed monocular depth estimation based on ordinal regression. Díaz *et al*. [13] used a soft ordinal label to train an ordinal regressor. Li *et al*. [25] developed a probabilistic embedding method for ordinal regression.

Order learning [27] learns ordering relations between instances. By comparing a test instance with references with known ranks, it can estimate the rank of the instance. In other words, it can perform ordinal regression. Lee and Kim [22] proposed the deep repulsive clustering and the order-identity decomposition for effective order learning. Their rank estimation algorithm was applied successfully to facial age estimation. The rationale behind order learning is that relative rank comparison of two instances is easier than absolute rank estimation of each instance. In this paper, we adopt this idea and propose a novel ordinal regression algorithm through relative comparisons. However, whereas order learning is based on ternary classification, the proposed MWR yields a continuous regression score, called ρ-rank, indicating the relative rank of an input instance between two references. Thus, unlike order learning, for example, in facial age estimation, given two references of ages 20 and 30, MWR can directly regress a continuous age of a test instance, which is older than 20 and younger than 30.

2.2. Applications

Facial age estimation and HCI classification are the most representative vision applications, on which new ordinal regression methods are tested and compared. Let us briefly review conventional ordinal regression methods for these two tasks.

Facial age estimation: It is one of the most popular and the most challenging ordinal regression tasks, whose goal is to predict people's ages using their facial images. Various age estimators have been developed, which can be grouped into four categories: classification [40, 44], regression [26, 41], ranking [11, 35], and distribution learning [38, 45] methods.

OR-CNN [35] and Ranking-CNN [11] are the ranking (or ordinal regression) methods, which regard ages as ranks. To estimate a person's age, they dichotomize whether the person is older than each age or not. By combining these binary classification results, the age can be estimated [15, 24]. While OR-CNN uses a common feature for all binary classifiers, Ranking-CNN uses different features for different classifiers. However, based on binary classifiers, these methods disregard the continuity of the aging process.

HCI classification: It aims at estimating the shooting decade of a photograph. Palermo *et al*. [37] introduced this task and perform the classification by exploiting different characteristics of the color imaging process in each decade. Martin *et al*. [34] developed a binary classifier to predict whether a photograph was taken earlier or later than each decade and combined multiple classification results to yield a final estimate. Liu *et al*. [31] used an additional classification loss for training their ordinal regressor. Li *et al*. [25] proposed a new feature embedding scheme for ordinal regression.

3. Proposed Algorithm

We propose the MWR algorithm for ordinal regression. First, we introduce the notion of ρ-rank. Then, we develop two types of relative regression networks: global and local ρ-regressors designed for entire and specific rank ranges, respectively. Using them, we predict the ρ-rank of an instance by comparing it with selected reference instances. Lastly, we estimate the absolute rank via an iterative refinement process, called MWR.

We use facial age estimation examples for concrete description of MWR in this section. However, note that MWR can be applied to other ordinal regression tasks as well.

3.1. ρ-Rank

Given an object instance x, we aim at estimating the corresponding rank $\theta(x)$. However, in many cases, rank estimation is challenging. For example, in age estimation, the facial aging process has large individual variations. People

of the same age often look quite different from one another, which makes it difficult to train a reliable network for age estimation.

To deal with this issue, we propose the notion of ρ-rank. Note that it is easier to tell the age ordering relations among multiple people than to estimate the exact age of each person [8, 27, 48]. Hence, instead of estimating the absolute rank $\theta(x)$ of input x, we predict its ρ-rank

$$\rho(x, y_1, y_2) = \frac{\theta(x) - \mu(y_1, y_2)}{\tau(y_1, y_2)} \tag{1}$$

where y_1 and y_2 are two references with $\theta(y_1) < \theta(y_2)$. Also, $\mu(y_1, y_2) = \frac{1}{2}(\theta(y_1) + \theta(y_2))$ is the average rank of the two references, and $\tau(y_1, y_2) = \frac{1}{2}(\theta(y_2) - \theta(y_1))$ is half of the rank difference between them.

Suppose that $\theta(y_1) \leq \theta(x) \leq \theta(y_2)$. Then, the ρ-rank in (1) has the following properties. First, $\rho \in [-1, 1]$. Second, the sign of ρ represents whether the rank of input x is closer to that of y_1 or y_2. It is positive when x is closer to y_2 and negative otherwise. Third, the absolute value of ρ quantifies the closeness of x to y_1 or y_2. For instance, $|\rho| = 1$ only if $\theta(x) = \theta(y_1)$ or $\theta(y_2)$. Finally, the absolute rank θ can be reconstructed from the ρ-rank by

$$\theta(x) = \rho(x, y_1, y_2) \cdot \tau(y_1, y_2) + \mu(y_1, y_2). \tag{2}$$

Hence, we predict the ρ-rank and then reconstruct the absolute rank θ from the predicted ρ-rank via (2).

3.2. ρ-Regressor

We develop a regressor to predict the ρ-rank in (1). The proposed ρ-regressor in Figure 2 consists of an encoder $f(\cdot)$ and a regression module $g(\cdot)$.

We adopt VGG16 [43] as the encoder. Specifically, we use the output of the last pooling layer in VGG16 as the feature vector $f(x)$. Note that the ρ-rank of x is determined in the context of two references y_1 and y_2. Therefore, the features $f(y_1)$ and $f(y_2)$ are also extracted by the same encoder, as shown in Figure 2. Next, the regression module takes the triplet $(f(x), f(y_1), f(y_2))$ and yields an estimate of the ρ-rank in (1), which is given by

$$\hat{\rho}(x, y_1, y_2) = g(f(x), f(y_1), f(y_2)). \tag{3}$$

The regression module comprises three fully connected layers: the first two adopt the ReLU [1] activation, while the last one uses the $\tanh$ activation to yield a value in $[-1, 1]$.

The ρ-regressor is end-to-end trained. We form a triplet (x, y_1, y_2) by choosing the references with a fixed

$$\tau = \frac{1}{2}(\theta(y_2) - \theta(y_1)). \tag{4}$$

In other words, the rank difference between y_1 and y_2 is constrained to be a constant. This is to lower the learning difficulty of the ρ-regressor. By fixing τ, the ρ-regressor needs to consider a much smaller subset of $\{(x, y_1, y_2)\}$ and can achieve more reliable regression. Given the triplet (x, y_1, y_2), the ρ-regressor obtains the estimate $\hat{\rho}$ in (3). Then, its squared error from the ground-truth ρ in (1) is defined as the loss, and the ρ-regressor is trained to minimize such losses over numerous triplets. When $\theta(x) < \theta(y_1)$ or $\theta(x) > \theta(y_2)$, the ground-truth ρ is set to -1 or 1, respectively.

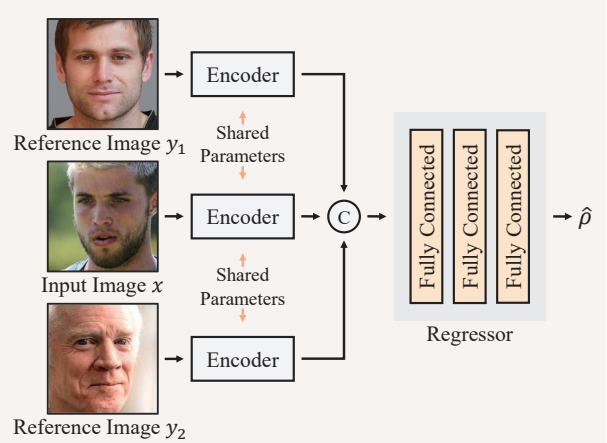


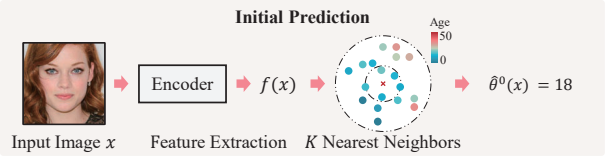
Figure 2. An overview of the ρ-regressor. Given an input instance x and two references y_1 and y_2, the encoder extracts their features separately. After concatenating those features, the regressor yields an estimate of the ρ-rank $\hat{\rho}(x, y_1, y_2)$. Here, © denotes concatenation.

3.3. Moving Window Regression

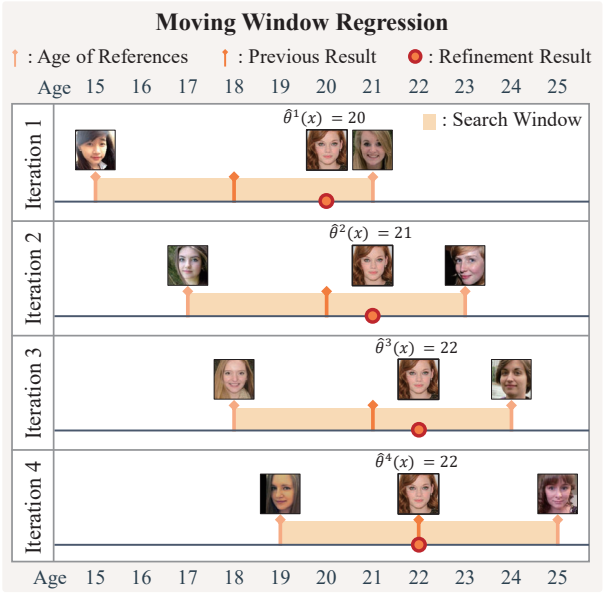
We estimate the absolute rank $\hat{\theta}(x)$ of an unseen test instance x by moving a window $[\theta(y_1), \theta(y_2)]$ and predicting the ρ-rank $\hat{\rho}(x, y_1, y_2)$ in (3) using the ρ-regressor. This MWR process is performed iteratively.

First, we obtain an initial estimate $\hat{\theta}^0(x)$, where the superscript denotes the iteration index. Figure 3(a) illustrates this initial estimation. The encoder extracts the feature vector $f(x)$. Then, in the feature space, we find the K NNs to x among all training instances in terms of the Euclidean distances. In this work, K is set to 5. The average rank of these neighbors is rounded to the nearest integer, which becomes the initial estimate $\hat{\theta}^0(x)$. This is reasonable since feature vectors are clustered according to the ranks, as visualized in Figure 4.

Next, at iteration step t, we refine $\hat{\theta}^{t-1}(x)$ to $\hat{\theta}^t(x)$, as in Figure 3(b). Among the training instances, we select a pair of references y_1^t and y_2^t, whose ranks are $\theta(y_1^t) = \hat{\theta}^{t-1}(x) - \tau$ and $\theta(y_2^t) = \hat{\theta}^{t-1}(x) + \tau$. The range $[\theta(y_1^t), \theta(y_2^t)]$ is referred to as the search window, which is centered around the previous estimate $\hat{\theta}^{t-1}(x)$. Within the search window,



(a)



(b)

Figure 3. An example of the MWR process in facial age estimation when the ground-truth age of input x is 22 and τ equals 3: (a) initial prediction and (b) iterative MWR refinement.

the ρ-regressor regresses the ρ-rank $\hat{\rho}(x, y_1^t, y_2^t)$, which is then converted to

$$\begin{aligned} \hat{\theta}^t(x) &= \text{round}\big(\hat{\rho}(x, y_1^t, y_2^t) \cdot \tau(y_1^t, y_2^t) + \mu(y_1^t, y_2^t)\big) \\ &= \text{round}\big(\hat{\rho}(x, y_1^t, y_2^t) \cdot \tau + \hat{\theta}^{t-1}(x)\big). \end{aligned} \tag{5}$$

The equality in (5) holds because τ is fixed and $\mu(y_1^t, y_2^t) = \hat{\theta}^{t-1}(x)$. This MWR is repeated until $\hat{\theta}^t(x) = \hat{\theta}^{t-1}(x)$ or a predefined number of iterations is reached. Note that the iteration terminates when the estimated rank is at the center of the search window, as shown in Iteration 4 in Figure 3(b).

3.4. Global and Local Regression

For accurate rank estimation, a network should be capable of learning various patterns. In facial age estimation, the facial aging process exhibits nonlinearity because each age group has different aging characteristics [2, 46]. During young ages, faces get bigger as they grow up. From middle to old ages, facial texture varies mainly due to skin aging. A global ρ-regressor, which is used for the entire age range, should learn these diverse aging patterns. On the other hand, a local ρ-regressor would be more effective for a specific age range if it were trained with images within that range only. The training also would be easier, because only the patterns in the smaller range should be learned [11].

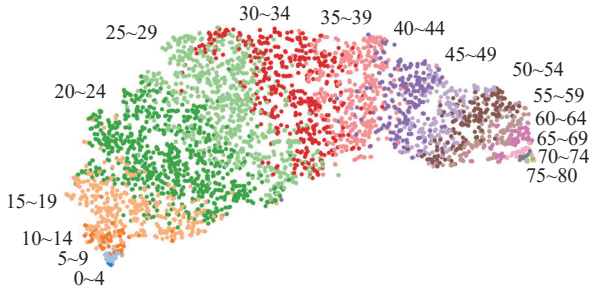
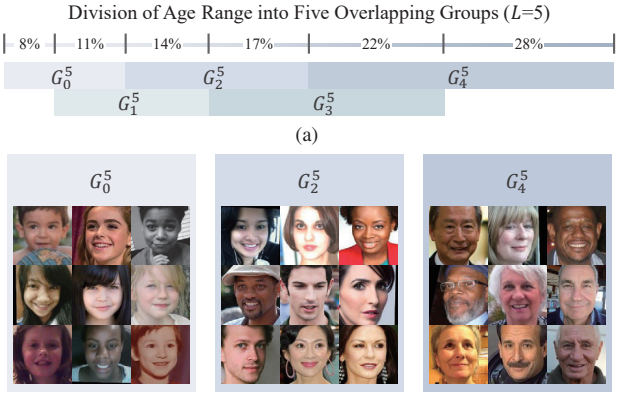


Figure 4. t-SNE visualization [33] of the feature space of the CLAP2015 dataset [14]. Note that feature vectors are aligned roughly according to the ages.



(b)

Figure 5. An example of the age range partition: (a) The entire age range is divided into five overlapping groups G_i^5, $i = 0, \ldots, 4$. (b) Images are sampled from groups G_0^5, G_2^5, and G_4^5 in the CLAP2015 dataset [14].

Hence, we employ a global ρ-regressor and L local ρ-regressors. To use local ρ-regressors, we divide the entire rank range into multiple groups. For example, in facial age estimation, the configuration $L = 5$ can be considered: First, the entire age range is partitioned into three age groups G_0^5, G_2^5, and G_4^5 in Figure 5(a). Then, in addition to the three groups, we use two more groups G_1^5 and G_3^5 in Figure 5(a). The lengths of the five overlapping groups form a geometric sequence, and G_4^5 covers the older half of the entire range. Thus, in the CLAP2015 dataset [14], the entire range $[3, 85]$ is divided into five groups $[3, 18]$, $[10, 29]$, $[19, 44]$, $[30, 62]$, and $[45, 85]$. Figure 5(b) shows some faces in group G_0^5, G_2^5, and G_4^5.

We can estimate the rank of an instance globally and then locally. First, we perform MWR using the global ρ-regressor to obtain the estimate $\hat{\theta}_{\text{global}}(x)$. Second, starting

Table 1. Comparison of facial age estimation results in the four evaluation settings (A, B, C, and D) of MORPH II and on FG-NET.

	Setting A		Setting B		Setting C		Setting D		FG-NET	
Algorithm	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
OR-CNN [35]	-	-	-	-	-	-	3.27	73.0	-	-
Ranking-CNN [11]	-	-	-	-	-	-	2.96	85.0	-	-
DMTL [18]	-	-	-	-	3.00	85.3	-	-	-	-
CMT [5]	-	-	-	-	2.91	-	-	-	-	-
DEX [40]	2.68	-	-	-	-	-	-	-	-	-
DRFs [41]	2.91	82.9	2.98	-	-	-	2.17	91.3	3.85	80.6
AGEn [44]	2.52	85.0	2.70	-	-	-	-	-	2.96	85.0
MV [38]	-	-	-	-	2.79	-	2.16	-	-	-
C3AE [9]	-	-	-	-	-	-	2.75	-	2.95	-
BridgeNet [26]	2.38	91.0	2.63	86.0	-	-	-	-	2.56	86.0
AVDL [45]	2.37	-	2.53	-	-	-	**1.94**	-	2.32	-
OL [27]	2.41	91.7	2.75	88.2	2.68	88.8	2.22	93.3	-	-
DRC-ORID [22]	2.26	93.8	**2.51**	89.7	2.58	89.5	2.16	93.5	-	-
Proposed	**2.13**	**94.2**	2.53	**90.4**	**2.53**	**90.5**	2.00	**95.0**	**2.23**	**91.1**

with $\hat{\theta}_{\text{global}}(x)$ for fast convergence, we iteratively refine the estimate using the L local ρ-regressors to yield $\hat{\theta}_{\text{local}}(x)$ eventually. At each iteration, due to the overlap, the previous estimate may belong to two rank groups. In such a case, both groups are selected and the estimated ranks from the corresponding local ρ-regressors are averaged.

To facilitate the transition between local ρ-regressors during the iterative MWR process, we train each local ρ-regressor using instances not only in the corresponding rank group but also those in nearby rank groups. More specifically, let $G_i = [\theta_{\min}, \theta_{\max}]$ denote the ith rank group, where $\theta_{\min}$ and $\theta_{\max}$ are the minimum and maximum ranks in the group. Then, the i-th local ρ-regressor is trained using instances x whose ranks $\theta(x)$ are within an extended range $[\theta_{\min} - \alpha, \theta_{\max} + \alpha]$, where α is set to 6 for age estimation.

3.5. Reference Selection

To predict the rank of a test instance, we compare it with reference pairs. We select these pairs from the training set offline prior to testing, based on the regression error γ, given by

$$\gamma(y_1, y_2) = \frac{1}{|W|} \sum_{x \in W} |\hat{\rho}(x, y_1, y_2) - \rho(x, y_1, y_2)| \quad (6)$$

where $W = \{x \,|\, \theta(x) \in [\theta(y_1) - \alpha, \theta(y_2) + \alpha]\}$. The regression error $\gamma(y_1, y_2)$ represents the average estimation error of ρ-ranks, when (y_1, y_2) is used as the reference pair. Thus, at iteration t of the MWR process, we use the optimal pair (y_1, y_2) with the smallest $\gamma(y_1, y_2)$, which satisfies the constraints of $\theta(y_1) = \hat{\theta}^{t-1}(x) - \tau$ and $\theta(y_2) = \hat{\theta}^{t-1}(x) + \tau$. Alternative reference selection schemes will be compared in Section 4.4.

4. Experimental Results

We assess the performances of the proposed MWR on facial age estimation and HCI classification.

4.1. Implementation Details

To train each ρ-regressor, we initialize its encoder using VGG16 pre-trained on ILSVRC2012 [12] and its regressor with the Kaiming uniform method [19]. In facial age estimation, we do random horizontal flipping and random cropping to the image size of 224×224 for data augmentation. In HCI classification, we do random horizontal flipping only. We use the Adam optimizer [36] with a learning rate of 10^{-4}. The minibatch size is 18. We use a PC with an Intel i7 processor and an NVIDIA RTX 2080Ti GPU.

4.2. Facial Age Estimation

Datasets: For facial age estimation, we use seven datasets: MORPH II [39], FG-NET [21], CLAP2015 [14], UTK [49], CACD [10], Adience [23], and IMDB-WIKI [40]. We align all facial images, expect for Adience images, using landmarks detected by MTCNN [47], as done in [26]. For the Adience dataset, we use aligned images in [23]. Unless specified otherwise, we pre-train the ρ-regressors on the IMDB-WIKI dataset, as done in [22, 26, 27, 38, 40, 44, 45]. Details about the datasets and experimental settings are in the supplemental document.

Geometric scheme: Compared to young people, it is harder to estimate old people's ages; telling the difference between a 5-year-old and a 10-year-old is easier than that between a 65-year-old and a 70-year-old. Thus, for the relative age estimation of an old person, it is more effective to use a large search window. But, in (4), the size of a search window $[\theta(y_1), \theta(y_2)]$ is fixed by τ, regardless of the age $\theta(x)$ of input x. This is called an arithmetic τ and denoted by τ_{ari}, because the arithmetic difference is fixed. Instead, the geometric ratio between two reference ages can be fixed by

$$\tau_{\text{geo}} = \frac{1}{2}\big(\log \theta(y_2) - \log \theta(y_1)\big). \quad (7)$$

In this geometric scheme, the MWR process in (1)$\sim$ (5) is

Table 2. Comparison on the validation and test splits of CLAP2015.

Algorithm	Validataion		Test	
	MAE	ϵ-error	MAE	ϵ-error
AgeNet [29]	3.33	0.29	-	0.26
Zhu *et al*. [50]	-	0.31	-	0.29
DEX [40]	3.25	0.28	-	0.26
AGEn [44]	3.21	0.28	2.94	0.26
BridgeNet [26]	2.98	**0.26**	2.87	0.26
Proposed	**2.95**	**0.26**	**2.77**	**0.25**

modified by replacing each age $\theta(\cdot)$ with the logarithmic age $\log\theta(\cdot)$. Thus, a bigger search window is used for an older person. The default mode uses the geometric scheme with $\tau_{\text{geo}} = 0.1$. The arithmetic and geometric schemes will be compared in Section 4.4.

Comparison with the state-of-the-arts: Table 1 compares the proposed algorithm with conventional algorithms in the four evaluation settings of MORPH II [39] and also on FG-NET [21]. We use the mean absolute error (MAE) and cumulative score (CS) metrics. MAE is the average absolute error between predicted and ground-truth ages, and CS is the percentage of images whose absolute errors are less than or equal to a tolerance level l. As in [22, 27, 45], $l = 5$. For the local regression, we set $L = 5$, as shown in Figure 5.

In Table 1, the proposed algorithm provides better results than the conventional algorithms in most tests. Compared with the state-of-the-art DRC-ORID [22], the proposed algorithm yields better CS scores in all four settings of MORPH II and better MAE scores in three out of the four settings. The performance gaps are significant in many cases. For example, in setting D, the proposed algorithm reduces MAE by 7.4% (from 2.16 to 2.00) and improves the CS score by 1.5%. Similar to the proposed algorithm, the order learning methods in [22, 27] do relative comparisons. However, whereas these methods predict discrete order relations between instances, the proposed algorithm regresses continuous relative ages.

Table 1 also compares the results on FG-NET [21], which is a small dataset containing about 900 training images only in each fold. The proposed algorithm provides superior results by significant gaps of 0.09 in MAE and 5.1% in CS. These excellent results on FG-NET confirm that the proposed algorithm learns aging characteristics effectively even from a small number of training images.

Next, Table 2 compares the results on CLAP2015 [14] using the metrics of MAE and ϵ-error. For each image, CLAP2015 provides the standard deviation of age ratings by multiple annotators. The standard deviation represents the estimation difficulty. By taking this difficulty into account, ϵ-error is defined as $1 - \exp(-\frac{(\hat{\theta}(x)-\theta(x))^2}{2\sigma^2})$, where σ is the standard deviation of a test image x. The average ϵ-error over all test images is reported. Both validation and test splits of CLAP2015 are used. For evaluation on the test set, we use the validation set, as well as the training set, to train the global and local ρ-regressors, as in [26, 29, 44].

The proposed algorithm achieves the best performances on both splits. CLAP2015 is a challenging dataset, so many conventional algorithms adopt performance boosting schemes. For example, BridgeNet [26] improves its performance by averaging the predictions on 10 flipped and cropped images. AgeNet [29], DEX [40], and AGEn [44] employ eight or more networks. However, without using such schemes, the proposed algorithm outperforms all conventional algorithms. Especially, a significant MAE margin of 0.1 is achieved on the test split.

Table 3 lists the performances on the UTK dataset [49]. The proposed algorithm outperforms the conventional algorithms with meaningful margins. Notice that Gustafsson *et al*. [17] and Berg *et al*. [3] employ the deeper ResNet50 [20] as their backbone networks, whereas the proposed ρ-regressors use VGG16. Nevertheless, the proposed algorithm improves the MAE score by more than 0.18.

Table 3 also compares the results on CACD [10], which is a big dataset with many noisy data. The proposed algorithm also outperforms the conventional algorithms by 0.23

Table 3. Comparison on UTK and in the train and validation settings of CACD. For both datasets, IMDB-WIKI pre-training is not performed.

Algorithm	UTK	Train	Validation
	MAE	MAE	MAE
dLDLF [42]	-	4.73	6.77
AGEn [44]	-	4.68	-
DRFs [41]	-	4.64	5.77
CORAL [6]	5.47	-	-
Gustafsson *et al*. [17]	4.65	-	-
Berg *et al*. [3]	4.55	-	-
Proposed	**4.37**	**4.41**	**5.68**

Table 4. Accuracy and MAE comparison on the Adience and HCI datasets. For Adience, IMDB-WIKI pre-training is not performed.

Algorithm	Adience		HCI	
	Accuracy (%)	MAE	Accuracy (%)	MAE
Frank & Hall [15]	-	-	41.4	0.99
Cardoso *et al*. [7]	-	-	41.3	0.95
Palermo *et al*. [37]	-	-	44.9	0.93
RED-SVM [28]	-	-	35.9	0.96
Martin *et al*. [34]	-	-	42.8	0.87
OR-CNN [35]	56.7	0.54	38.7	0.95
CNNPOR [31]	57.4	0.55	50.1	0.82
GP-DNNOR [32]	57.4	0.54	46.6	0.76
SORD [13]	59.6	0.49	-	-
DRC-ORID [22]	-	-	44.7	0.80
Li *et al*. [25]	60.5	0.47	54.7	0.66
Proposed	**62.6**	**0.45**	**57.8**	**0.58**

Table 5. Comparison of reference selection schemes on the test split of CLAP2015. MAE/ϵ-error performances are compared. For the random scheme, the mean and standard deviation of MAE and ϵ-error of 5 evaluation results are reported.

Random	Min γ	Max γ
2.82±0.01/0.26±0.01	**2.77/0.25**	3.17/0.30

Table 6. Comparison of global and local ρ-regressors.

	MORPH II (MAE/CS)	CLAP2015 (MAE/ϵ-error)	UTK (MAE)
Global ρ-regressor	2.24/93.5	2.82/0.26	4.49
Local ρ-regressors	**2.13/94.2**	**2.77/0.25**	**4.37**

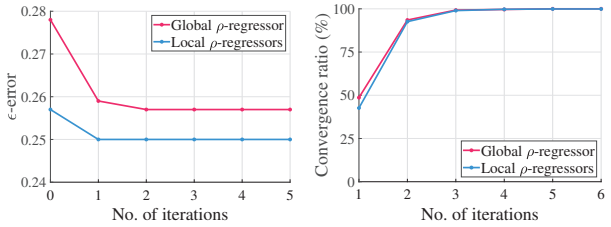


Figure 6. Plots of (a) ϵ-errors and (b) convergence ratios according to the number of iterations on the test split of CLAP2015.

and 0.09 in MAE on the train and validation splits, respectively.

Lastly, Table 4 compares the results on Adience [23], which is used for age group estimation. In this test, we adopt the arithmetic scheme with $\tau_{\text{ari}} = 2$. Also, we use three local ρ-regressors that cover three equal parts of the rank range, respectively. The proposed algorithm outperforms the conventional algorithms by significant gaps of 2.1% in accuracy and 0.02 in MAE.

4.3. HCI Classification

HCI [37] is a dataset for determining the decade when a photograph was taken. It contains images from five decades 1930s $\sim$ 1970s. There are 265 images in each decade. As done in [25, 31, 32, 37], we randomly split those 265 images into three subsets: 210 for training, 5 for validation, and 50 for testing. Then, the 10-fold cross-validation is performed. We use reference pairs satisfying $\tau_{\text{ari}} \geq 1$. For the local regression, three local ρ-regressors are employed that cover three equal parts of the rank range, respectively. Table 4 also compares the results on the HCI dataset. The proposed algorithm provides superior results by significant gaps of 3.1% in accuracy and 0.08 in MAE.

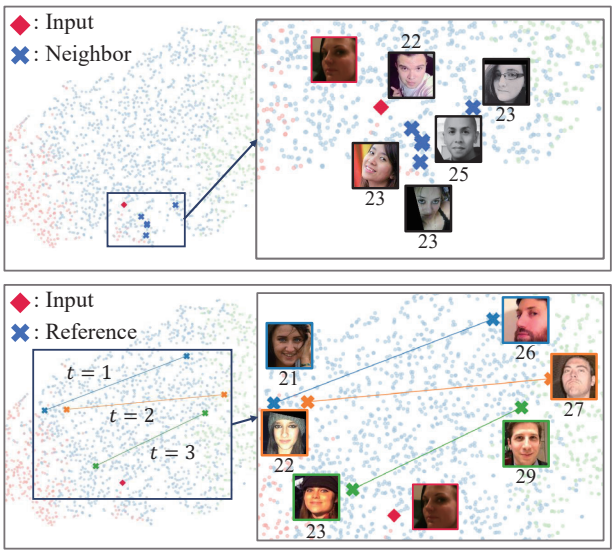


Figure 7. t-SNE visualization of MWR iterations. Top, the ground-truth age 27 of input x is initially estimated to $\hat{\theta}^0(x) = 23$ by finding the five NNs. Note that the 2D plot cannot perfectly preserve the order of all distances in the higher-dimensional feature space. Thus, the NNs in the feature space are not the nearest ones in this 2D plot. Above, the global estimate $\hat{\theta}_{\text{global}}(x) = 25$ is obtained in three iterations. The reference pair in each iteration is connected by an edge. In iterations 1, 2, and 3, the estimated ages are 24, 25, and 25, respectively.

4.4. Analysis

Next, we analyze the proposed MWR algorithm using facial age estimation datasets.

Reference selection: Table 5 compares three reference selection schemes on CLAP2015. At iteration t of the MWR process, a reference pair (y_1, y_2) should satisfy the constraints of $\theta(y_1) = \hat{\theta}^{t-1}(x) - \tau$ and $\theta(y_2) = \hat{\theta}^{t-1}(x) + \tau$. Among all pairs satisfying the constraints, the random scheme selects one randomly. On the other hand, the min γ or max γ scheme chooses the pair with the smallest or largest regression error γ in (6), respectively. The min γ scheme achieves the best results, so it is used as the default mode in this work. Examples of selected references are provided in the supplemental document.

Global vs. local regression: Table 6 compares the performances of global and local ρ-regressors in three test settings: MORPH II setting A, CLAP2015 test split, and UTK. First, only the global ρ-regressor is employed. In other words, MWR is performed without employing local ρ-regressors. Even the global regression outperforms the conventional state-of-the-art methods with meaningful margins in most cases. For example, on UTK, the global regression yields 0.06 lower MAE than the state-of-the-art Berg *et al*. [3]. Moreover, by employing local ρ-regressors,

Table 7. Comparison of the arithmetic and geometric schemes on the test split of CLAP2015. Only the global ρ-regressor is used. The scores are slightly poorer than Table 2. This is because the training is performed for 60 epochs, while it is done for 130 epochs in Table 2.

	τ_{ari} for arithmetic scheme			τ_{geo} for geometric scheme			
	3	5	7	0.05	0.075	0.1	0.125
MAE	3.21	3.33	3.24	3.03	2.95	**2.93**	2.95
ϵ-error	0.304	0.311	0.303	0.276	0.271	**0.266**	0.275

the proposed algorithm further improves the results. Especially, significant MAE improvements of 0.18 are achieved in comparison with Berg *et al*. [3] on UTK. More comparison results are presented in the supplemental document.

Iterative MWR process: To predict the rank of a test image, the proposed MWR iteratively compares it with references, which are selected from training images. The features of all references are extracted in advance for efficiency. Figure 6(a) plots how the age estimation performance varies as the MWR iteration goes on. In this test, we measure the ϵ-errors on the test split of CLAP2015. In both global and local regression, the errors are reduced significantly at the first iterations and then saturate after the second iterations. Note that the local regression starts from global regression results $\hat{\theta}_{\text{global}}(x)$. Nevertheless, MWR further improves the results meaningfully using the local ρ-regressors. Figure 6(b) shows the ratio of cases that are converged up until each iteration. Although the convergence of MWR is not guaranteed theoretically, the global and local regression, respectively, converges within 6 iterations for most of images. In most cases, 4 iterations are sufficient. The result from the last iteration becomes the local regression estimate $\hat{\theta}_{\text{local}}(x)$. Due to fast convergence, the total computing time for both global and local regression is only 0.007s (or equivalently 143fps) per image.

Figure 7 visualizes an example of the MWR process in the embedding space, where the proposed algorithm obtains a global regression result $\hat{\theta}_{\text{global}}(x)$ in three iterations.

Geometric vs. arithmetic schemes: While the arithmetic scheme processes the entire age range identically, the geometric scheme in (7) treats younger ages more finely using a smaller search range. Table 7 compares these two schemes with various τ's on CLAP2015. In this test, only the global ρ-regressor is used. In general, the geometric scheme is better than the arithmetic scheme. The best performance is achieved by the geometric scheme with $\tau_{\text{geo}} = 0.1$, which is used in the default mode. The size of a search window varies in the geometric scheme. For example, at $\tau_{\text{geo}} = 0.1$, the size $\theta^t(y_2) - \theta^t(y_1)$ of the window is 3 for $\hat{\theta}^{t-1}(x) = 13$, while it is 12 for $\hat{\theta}^{t-1}(x) = 60$. In other words, a fine search is carried out near a young age estimate, while a coarse search is done near an old one.

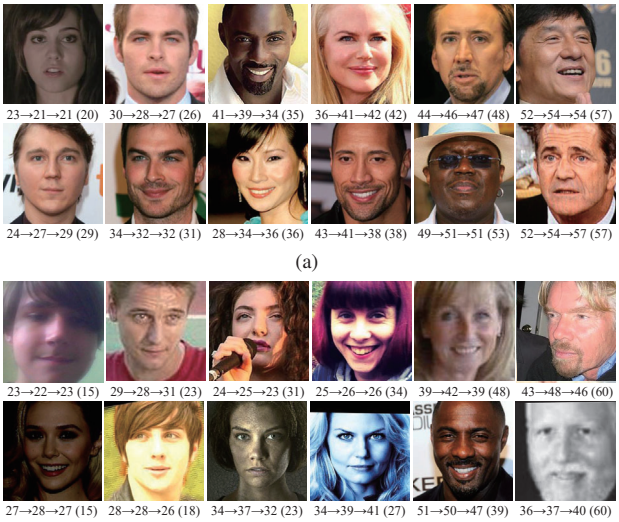


Figure 8. (a) Success and (b) failure cases of the proposed algorithm in facial age estimation. For each image x, the initial, global, and local estimates $\hat{\theta}^0(x) \rightarrow \hat{\theta}_{\text{global}}(x) \rightarrow \hat{\theta}_{\text{local}}(x)$ are reported with the ground-truth $(\theta(x))$ within the parentheses.

Success and failure cases: Figure 8 shows some success and failure cases of the proposed algorithm in facial age estimation. In (a), ages are estimated precisely with absolute errors less than 4. In (b), failure cases are shown, which are challenging examples due to various factors, such as low quality photographs, overexposure, and poor illumination.

5. Conclusions

We proposed a novel ordinal regression algorithm, called MWR. First, for relative ordinal regression, we designed global and local ρ-regressors. Then, we developed the MWR algorithm using these ρ-regressors. MWR first obtains an initial rank estimate based on the NN criterion. Then, it refines the estimate iteratively by selecting two reference instances to form a search window and estimating the ρ-rank within search window. Extensive experiments on various datasets showed that the proposed MWR algorithm provides outstanding rank estimation performances.

Acknowledgements

This work was conducted partly by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD) and supported partly by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2021R1A4A1031864 and No. NRF-2022R1A2B5B03002310).

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (ReLU). In *arXiv preprint arXiv:1803.08375*, 2018. 3

[2] A. Midori Albert, Karl Ricanek Jr., and Eric Patterson. A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Science International*, 172:1–9, 2007. 1, 4

[3] Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. In *Proc. IEEE ICPR*, 2021. 6, 7, 8

[4] Arthur L. Blumenthal. *The Process of Cognition.* Prentice Hall, 1977. 1

[5] Yoo ByungIn, Youngjun Kwak, Youngsung Kim, Changkyu Choi, and Junmo Kim. Deep facial age estimation using conditional multitask learning with weak label expansion. *IEEE Signal Process. Lett.*, 25:808–812, 2018. 5

[6] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank-consistent ordinal regression for neural networks. *Pattern Recog. Lett.*, 140:325–331, 2020. 6

[7] Jaime Cardoso and Joaquim Pinto da Costa. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8:1393–1429, 2007. 6

[8] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A ranking approach for human age estimation based on face images. In *Proc. IEEE ICPR*, 2010. 3

[9] Zhang Chao, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3AE: Exploring the limits of compact model for age estimation. In *CVPR*, 2019. 5

[10] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 17:804–815, 2015. 5, 6

[11] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *CVPR*, 2017. 2, 4, 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[13] Raúl Díaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 2, 6

[14] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV Workshops*, 2015. 4, 5, 6

[15] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML*, 2001. 2, 6

[16] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[17] Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B. Schön. Energy-based models for deep probabilistic regression. In *ECCV*, 2020. 6

[18] Hu Han, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40:2597–2609, 2017. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[21] Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:442–455, 2002. 5, 6

[22] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *ICLR*, 2021. 1, 2, 5, 6

[23] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshops*, 2015. 5, 7

[24] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *NIPS*, 2007. 2

[25] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 2, 6, 7

[26] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. BridgeNet: A continuity-aware probabilistic network for age estimation. In *CVPR*, 2019. 1, 2, 5, 6

[27] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020. 1, 2, 3, 5, 6

[28] Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24:1329–1367, 2012. 6

[29] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. AgeNet: Deeply learned regressor and classifier for robust apparent age estimation. In *ICCV Workshops*, 2015. 6

[30] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. Deep ordinal regression based on data relationship for small datasets. In *IJCAI*, 2017. 2

[31] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, 2018. 2, 6, 7

[32] Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *ICCV*, 2019. 6, 7

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 4

[34] Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *ICMR*, 2014. 1, 2, 6

[35] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *CVPR*, 2016. 1, 2, 5, 6

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[37] Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *ECCV*, 2012. 2, 6, 7

[38] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018. 2, 5

[39] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *FGR*, 2006. 5, 6

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126:144–157, 2018. 2, 5, 6

[41] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Yuille. Deep regression forests for age estimation. In *CVPR*, 2018. 2, 5, 6

[42] Wei Shen, Kai Zhao, Yilu Guo, and Alan Yuille. Label distribution learning forests. In *NIPS*, 2017. 6

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[44] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z. Li. Efficient group-n encoding and decoding for facial age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40:2610–2623, 2018. 2, 5, 6

[45] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, 2020. 2, 5, 6

[46] Leslie Zebrowitz. *Reading Faces: Window to the Soul?* Westview Press, 1997. 1, 4

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23:1499–1503, 2016. 5

[48] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *BMVC*, 2017. 3

[49] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017b. 5, 6

[50] Yu Zhu, Guowang Mu, and Guodong Guo. A study on apparent age estimation. In *ICCV Workshops*, 2015. 6