Cumulative Attribute Relation Regularization Learning for Human Age Estimation

Qing Tian, Songcan Chen*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
{*tianqing, s.chen*}*@nuaa.edu.cn*

Abstract

In recent years, the problem of human face-based age estimation has attracted increasing attention due to its extensive applicability and motivated a variety of approaches being proposed, in which the method based on the coding of *cumulative attribute* (CA) achieves competitive performance by taking into account both the neighbor-similar and the ordinal characteristics of ages. However, in their learning, the inherent *mutual relations* between the CA codes have not been exploited, thus leaving us a performance space that can be improved. To this end, in this work we first derive such relations by performing the difference-like operation between the CA codes in certain order to construct so-called 0-order and 1-order relation matrices and then incorporate them as two corresponding regularization terms, coined as CA-oriented ordinal structure regularization (CAOSR) and CA-oriented adjacent difference orthogonal regularization (CAADOR), into the objective of the multi-output regressor. Consequently, corresponding CA-based regressors regularized with the mutual relations are developed. Finally, through extensive experiments on three human aging datasets, the FG-NET and the Morph Album 1 and Album 2, we demonstrate the effectiveness of our strategies in improving CA-based age estimation.

Keywords: Age Estimation; Cumulative Attribute; Mutual Relation; Regularization; LS-SVR; Ridge Regression

1. Introduction

In machine learning, large numbers of problems are related to human face due to that rich information is contained in it, such as facial expression, gender, race and age, in which the problem of human face-based age estimation has aroused increasing attention due to its wide applications such as web security control (Guo et al., 2008a; Lanitis et al., 2004), ancillary identity authentication (Jain

Corresponding author: s.chen@nuaa.edu.cn

et al., 2004), and advertisement recommendation (Romano J.R. and Fjermestad, 2006), etc.

In order to conduct age estimation based on human face, a variety of approaches have been proposed to date. Generally, they fall into three categories: *classification-based*, e.g., (Lanitis et al., 2004; Geng et al., 2013; Ueki et al., 2006; Alnajar et al., 2012; Sai et al., 2015), *regression-based*, e.g., (Lanitis et al., 2002; Fu et al., 2007; Luu et al., 2009; Yan et al., 2007b,a; Geng et al., 2007; Chang et al., 2011; Li et al., 2012a,b), and *their hybrid*, e.g., (Guo et al., 2008a,b; Kohli et al., 2013).

When we consider each age as a separate class, the age estimation can be made under ordinary classification framework. For example, Lanitis et al. (Lanitis et al., 2004) extracted AAM features from facial images and respectively applied the nearest neighbor classifier and artificial neural networks for age estimation and achieved comparable performance. Geng et al. (Geng et al., 2013) specially designed a three-layer conditional probability neural network (CPNN) to capture the age contribution information for age classification. Moreover, Ueki et al. (Ueki et al., 2006) conducted age group classification by building Gaussian mixture models after discriminative dimensionality-reduction and received promising results respectively for male and female on several famous age datasets. More recently, Alnajar et al. (Alnajar et al., 2012) employed the soft coding to extract codebooks for age group classification and received better estimation on an unconstrained real-life dataset than the hard coding approaches. And Sai et al. (Sai et al., 2015) even used the extreme learning machines to perform age group estimation and obtained competitive results.

Actually, the age estimation is more of a regression problem than a generic multi-class classification due to the continuity of aging. According to this characteristic, many attempts have been made. For instance, Lanitis et al. (Lanitis et al., 2002) established a quadratic function to fit the ages with facial images represented by AAM features. Fu et al. (Fu et al., 2007) borrowed the multiple linear regression to learn an aging prediction function in the manifold space. And Luu et al.(Luu et al., 2009) employed the off-the-shelf ξ-SVR (Vapnik, 1998) for aging function learning. Moreover, to handle the uncertainty of annotations of age labels, Yan et al. (Yan et al., 2007b) constructed a semi-definite programming (SDP) regression model to train an aging regressor. Although the SDP regressor can relatively model the age labels' uncertainty better, the learning is very time-consuming. To reduce the time complexity, they (Yan et al., 2007a) then proposed to speed up the SDP learning by using the Expectation-Maximization (EM). Furthermore, Geng et al. (Geng et al., 2007) proposed the aging pattern regressing (AGES) to generate age labels for missing patterns. Although the methods afore-mentioned can yield age estimation performance to different extents, they ignored the fact that there exists natural ordinality among ages (Chen et al., 2013; Chang et al., 2011). To this end, Chang et al. (Chang et al., 2011) specially designed an ordinal hyperplanes ranker (OHRank) for age estimation and on FG-NET dataset they obtained better performance than AGES. Later, Li et al. (Li et al., 2012a) presented a distance-based ordinal regressor for age estimation, in which the ordinal information of ages is

incorporated into the metric and on FG-NET they obtained competitive performance. Moreover, they (Li et al., 2012b) took the ordinality and local manifold structure preserving ability as a criterion to perform feature selection and conducted age regression with much competitive results. More recently, they (Li et al., 2014) presented an ordinal metric learning method for image ranking by preserving both the local geometry information and the ordinal relationship of the data.

Although the methods reviewed above can perform encouraging human age estimation with different performance, they have not exploited another essential characteristic of the ages that neighboring ages are generally more similar in facial appearance than those apart. For example, the facial appearance of 11-year-old is more similar to that of 13 compared to that of 30, as exhibited in Figure 2 (in Section 2). This characteristic is of help in estimating the ages, especially when the age distribution is imbalanced (Chen et al., 2013), because similar ages can be used to partially depict their neighboring ages that are absent in the learning and thus alleviate the imbalance. Therefore, such *neighbor-similarity* of ages should also be incorporated into the estimation. To simultaneously consider both the *ordinality* and the *neighbor-similarity* of the ages[1], Chen et al. (Chen et al., 2013) proposed the *cumulative attribute* (CA) coding to represent the age. Concretely, they first used the multivariate ridge regression (mRR) (An et al., 2007) to transform the instance from its original input feature to a CA code; and then applied a second-layer scalar-output regressor to map the CA code to a scalar age label. The flowchart of the two-layer regression is shown in Figure 1, and by this way they obtained competitive age estimations.

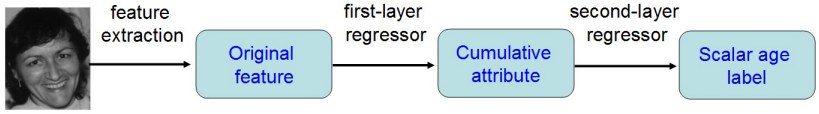


Figure 1: The flowchart of CA-based human age estimation.

Although the characteristics of ordinality and neighbor-similarity of the ages are considered in the CA coding, the inherent *mutual relations* explicitly or implicitly existing between the CA codes have not been exploited for learning, thus leaving us a room of promoting its performance. To this end, in this work we first derive such relations by performing difference-like operations on the

[1]Please note the difference between the ordinality and the neighbor-similarity of ages. The ordinality defines the global order relationship of the ages while the neighbor-similarity states the similarity of facial appearances at close ages. Moreover, the neighbor-similarity is also conceptually different from the local manifold geometry relationships (Li et al., 2012b, 2014) of the ages. The local manifold geometry structures describe the neighbor relationship of ages in manifold space, while the neighbor-similarity stated in this paper depicts the biological similarity of facial appearances of neighboring ages.

CA coding matrix[2] to construct so-called *0*-order and *1*-order *relation matrices* [3], respectively. Then, we formulate the *relation matrices* as two corresponding regularization terms, coined as *CA-oriented ordinal structure regularization* (CAOSR) and *CA-oriented adjacent difference orthogonal regularization* (CAADOR), respectively. And, in order to take the *mutual relations* into the CA learning, we regularize the first-layer regressor (as shown in Figure 1) by embedding the regularization terms, CAOSR and CAADOR, into its objective. Finally, through extensive experiments, we demonstrate the effectiveness of our strategies in improving CA learning on human age estimation.

The rest of this paper is organized as follows. In Section 2, we briefly review related work on CA coding. In Section 3, We derive two types of regularization terms, coined as *CAOSR* and *CAADOR*, to depict the mutual relations among the CA codes, and embed them into the objectives of the mRR and mLS-SVR, both of which act as the first-layer regressor in the CA learning, in Section 4. In Section 5, we conduct experiments to evaluate our strategies. Finally, we conclude the paper in Section 6.

2. Related Work

Following the spirit of literature such as (Mahajan et al., 2011), Chen et al. (Chen et al., 2013) presented the *cumulative attribute* (CA) coding for learning in such scenarios as human age estimation. Concretely, given a set of N training samples $\{x_i,\ l_i\} \in \Re^D \times \Re,\ l_i \in \{1, 2, ..., K\},\ i$=1,2,...,N, where x_i denotes the i-th instance and l_i is its corresponding scalar label, D denotes the feature dimensionality of x_i and K is the number of classes (e.g., the scale of the aging range). Here for the i-th sample x_i, its scalar label value l_i, e.g., the age value, is transformed into a K-dimensional vector y_i, named as *cumulative attribute* (CA) code, whose j-th element is defined as

$$y_i^j = \begin{cases} 1, & j \leq l_i \\ 0, & j > l_i \end{cases}$$

where j=1,2,...,K. As a comparison, the *non-cumulative attribute* (NCA) is given as well with the j-th element defined as

$$y_i^j = \begin{cases} 1, & j = l_i \\ 0, & j \neq l_i \end{cases}$$

As argued for age estimation, the CA coding can relatively well capture the characteristic that the attribute values at neighboring ages should be more similar than those further apart. Moreover, it can alleviate the challenge of the

[2]The *CA coding matrix* refers to such a matrix in which each column corresponds to a CA code for an instance from some class, as shown in Figure 3 (a).

[3]For why just extracting the *0*-order and *1*-order relation matrices, please refer to the *Remark* in Section 3.

insufficient and imbalanced sample distribution within the entire aging range, while the NCA coding cannot. The appealing characteristic of CA coding can be intuitively demonstrated in Figure 2.

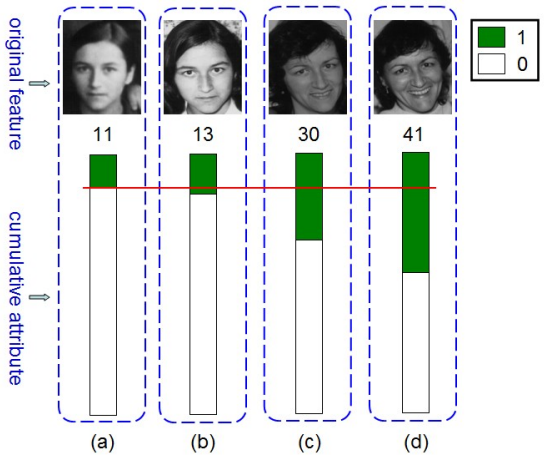


Figure 2: Demonstration of CA coding on age representation (the digit under each facial image represents the facial age).

Just as shown in Figure 2, the facial age appearances between (a) and (b) obviously are more similar than those between (a) and (c), which is consistent with the coding differences between their corresponding CA, i.e., the difference of 2 years between (a) and (b) is smaller than that of 19 between (a) and (c). However, using NCA, the age difference between (a) and (b) and that between (a) and (c) are the same and both equal to 1, by which the similarity of neighboring ages is not reflected at all, according to its definition above. Therefore, the coding way of CA is reasonable and desirable to depict human facial age.

Now given a human facial image, x_i, in order to learn a mapping $y_i = W^T x_i + B$ from its original feature x_i to the CA representation y_i, where $W = [w_1, .., w_K] \in \Re^{D \times K}$ is the weight matrix and $B \in \Re^K$ is the bias, Chen et al. (Chen et al., 2013) used the mRR because of its robustness in regression, formulated as

$$\min_{W,B} \frac{1}{2} \sum_{i=1}^{N} \|y_i - (W^T x_i + B)\|_F^2 + \frac{\lambda}{2} \|W\|_F^2, \tag{1}$$

where λ is a regularization pamameter. Eq. (1) is a non-constrained quadratic programming (QP) problem and the W and B can be solved with an analytical solution. Due to space limitation, we here omit the details, which can be referred to (Chen et al., 2013).

After obtaining the W and B by solving (1), we can predict the CA value for a given sample x_i by performing the first-layer regression as in (Chen et al.,

2013), and then adopt any off-the-shelf regressor or classifier as the second-layer estimator, such as Support Vector Regression (SVR) (Smola and Schölkopf, 2004), ridge regressor (RR) (Montgomery et al., 2012), and the nearest neighbor classifier (Garcia et al., 2012), to map the resulting CA vector y_i to its corresponding scalar value l_i. The whole flowchart is shown in Figure 1.

3. Two Types of CA-oriented Regularization

According to the definition of CA coding in Section 2, for the given K ordinal classes, we can demonstrate their CA codes together in a *CA coding matrix* as shown in Figure 3 (a).

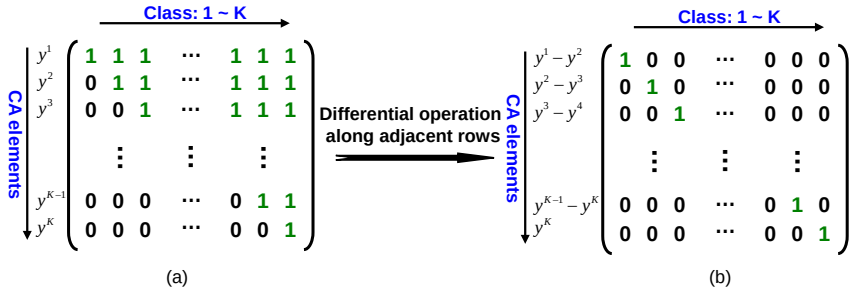


Figure 3: Demonstration of CA coding corresponding to K ordinal classes. (a): original *CA coding matrix* (b): CA-oriented adjacent difference matrix

3.1. Motivation

Although both the progressive, continuous aging and the ordinal characteristics of the ages are already reflected in the CA coding as reported in (Chen et al., 2013), the inherent mutual relations hidden between the CA codes have not been exploited into learning, thus leaving us an opportunity to promote its performance. By analyzing the structure of the CA coding matrix (as shown in Figure 3(a)), we can discover some *mutual relations* explicitly or implicitly hidden between the CA codes along the row direction, which motivates us to derive such *mutual relations* by performing difference-like operations along certain directions of the *CA coding matrix* to construct so-called *relation matrices*, which explicitly reveal some mutual relations among the original CA codes, as shown in Figure 3 (a) and (b). Then, we formulate the mutual relations as prior information and thus incorporate them into the CA learning.

3.2. Notations

Before presenting our work, first let us introduce some notations that will be used below. Let $X = [X_1, X_2, ..., X_K] \in \Re^{D \times N}$ denote the entire set of training instances, where X_k, $k = 1, 2, ..., K$, denotes the set of samples from the k-th class with size N_k and $\sum_{k=1}^{K} N_k = N$. In addition, let $Y = [Y_1, Y_2, .., Y_K] =$

$[(Y^1)^T, (Y^2)^T, ..., (Y^K)^T]^T \in \Re^{K\times N}$ denote the corresponding *CA coding matrix* (each column corresponds to one instance) and $L = [L_1, L_2, ..., L_K] \in \Re^{1\times N}$ the scalar class labels of the training set X, in which $Y_k \in \Re^{K\times N_k}$, $Y^k \in \Re^{1\times N}$ and $L_k = [l_1^k, l_2^k, ..., l_{N_k}^k] \in \Re^{1\times N_k}$, $k = 1, 2, ..., K$. In addition, $W = [w_1, w_2, ..., w_K] \in \Re^{D\times K}$ and $B = [b_1, b_2, ..., b_K]^T \in \Re^{K\times 1}$ respectively are the weight matrix and bias vector in the projection of

$$Y_k = W^T X_k + B{e_{N_k}}^T,\ k = 1, 2, ..., K, \tag{2}$$

where $e_{N_k} \in \Re^{N_k}$ is a vector with all elements equal to 1.

3.3. CA-oriented Ordinal Structure Regularization (CAOSR)

From Figure 3 (a), we can find that the original *CA coding matrix* (i.e., the *0-order relation matrix* [4]) of the K classes is an upper-triangle matrix. *Such a characteristic can be exploited to depict the explicit mutual relation between the CA codes and thus incorporated into the CA learning.* Inspired by this, we unfold Eq. (2) with respect to the W and X and get the detailed formulation[5] as

$$Y := \begin{pmatrix} Y^1: & (w_1^T X_1)^* & (w_1^T X_2)^* & (w_1^T X_3)^* & \dots & (w_1^T X_K)^* \\ Y^2: & (w_2^T X_1)^* & (w_2^T X_2)^* & (w_2^T X_3)^* & \dots & (w_2^T X_K)^* \\ Y^3: & (w_3^T X_1)^* & (w_3^T X_2)^* & (w_3^T X_3)^* & \dots & (w_3^T X_K)^* \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y^{K-1}: & (w_{K-1}^T X_1)^* & (w_{K-1}^T X_2)^* & (w_{K-1}^T X_3)^* & \dots & (w_{K-1}^T X_K)^* \\ Y^K: & (w_K^T X_1)^* & (w_K^T X_2)^* & (w_K^T X_3)^* & \dots & (w_K^T X_K)^* \end{pmatrix} \tag{3}$$

According to Figure 3 (a) and the analyses above, all elements of the block lower-triangle of Y in Eq. (3) should be equal to zeros, i.e.,

$$(w_i^T X_j)^* := \ w_i^T X_j + b_i {e_{N_j}}^T = \mathbf{0}_{N_j}^T,\ i, j = 1, 2, ..., K,\ (j < i) \tag{4}$$

where $\mathbf{0}_{N_j} \in \Re^{N_j}$ represents a zero vector with dimension N_j. Along this way, we can incorporate this prior knowledge as a regularization term, coined as *CA-oriented ordinal structure regularization* (CAOSR), into the objective of the first-layer multi-output regressor by *minimizing* the following term

$$\begin{aligned} \mathcal{L}_{CAOSR} &= \sum_{k=1}^{K}\sum_{c=1}^{k-1} \|[w_k; b_k]^T [X_c; e_c^T]\|^2 \\ &= \sum_{k=1}^{K} \|\widetilde{w}_k^T \widetilde{X} R^k\|^2, \end{aligned} \tag{5}$$

where $\widetilde{w}_k = [w_k; b_k]$ is an augmented weight vector with dimension D+1 for w_k, $R^k = diag(I_{\sum_{c=1}^{k-1} N_c}, \mathbf{0}_{\sum_{c'=k}^{K} N_{c'}})$ is an auxiliary square matrix of order N with $I_{\sum_{c=1}^{k-1} N_c}$ being an identity matrix of order $\sum_{c=1}^{k-1} N_c$ and $\mathbf{0}_{\sum_{c'=k}^{K} N_{c'}}$ an all-zero square matrix of order $\sum_{c'=k}^{K} N_{c'}$, $k = 1, 2, ...K$, and $\widetilde{X} = [X; e_N] \in \Re^{(D+1)\times N}$.

[4] Without losing the generality, here we also call the original *CA coding matrix* as the 0-*order relation matrix*.

[5] Due to space limitation, here $(w_i^T X_j)^*$ represents $w_i^T X_j + b_i {e_{N_j}}^T$.

3.4. CA-oriented Adjacent Difference Orthogonal Regularization (CAADOR)

Besides the explicit mutual relation depicted by the *CAOSR* above, performing the *1*-order *differentiating operation* on the original *CA coding matrix* along the row direction yields its *1*-order relation matrix, as shown in Figure 3(b). Interestingly, the *1*-order relation matrix reveals an exciting rule: *any two rows are mutually orthogonal to each other. It can be exploited to depict the implicit mutual relation between the CA codes.* For this purpose, we attempt to develop another variety of regularization scheme to take advantage of the implicit mutual relation for CA learning. Specifically, we perform the *1*-order differentiating operation over every two adjacent rows of the CA coding matrix Y, shown in Eq. (3), and consequently obtain the following derivation Y_{diff}[6], exhibited as

$$Y_{diff} := \begin{pmatrix} Y^1 - Y^2: & ((w_1 - w_2)^T X_1)^* & ((w_1 - w_2)^T X_2)^* & \dots & ((w_1 - w_2)^T X_K)^* \\ Y^2 - Y^3: & ((w_2 - w_3)^T X_1)^* & ((w_2 - w_3)^T X_2)^* & \dots & ((w_2 - w_3)^T X_K)^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y^{K-1} - Y^K: & ((w_{K-1} - w_K)^T X_1)^* & ((w_{K-1} - w_K)^T X_2)^* & \dots & ((w_{K-1} - w_K)^T X_K)^* \\ Y^K: & (w_K^T X_1)^* & (w_K^T X_2)^* & \dots & (w_K^T X_K)^* \end{pmatrix}. \tag{6}$$

As a result, any two rows of Y_{diff} should be orthogonal to each other, i.e., the inner product between any two of them is zero. According to the analysis, we come to propose the second CA-oriented regularization term, coined as *CA-oriented adjacent difference orthogonal regularization* (CAADOR), to express the implicit mutual relation between the CA codes. Similarly, the empirical risk $\mathcal{L}_{CAADOR}$ over the CAADOR should also be *minimized*

$$\begin{aligned} \mathcal{L}_{CAADOR} = & \sum_{k=1}^{K-1} \sum_{c \neq k}^{K-1} \|(\widetilde{w}_k - \widetilde{w}_{k+1})^T \widetilde{X} (Y^c - Y^{c+1})^T\|^2 \\ & + \sum_{k=1}^{K-1} \|(\widetilde{w}_k - \widetilde{w}_{k+1})^T \widetilde{X} (Y^K)^T\|^2 \\ & + \sum_{k=K}^{K} \sum_{c \neq k} \|\widetilde{w}_k^T \widetilde{X} (Y^c - Y^{c+1})^T\|^2. \end{aligned} \tag{7}$$

Note that the CAADOR is calculated in two cases: $k = 1, 2, ..., K-1$ and $k = K$ respectively corresponding to the first two terms and the last one in Eq. (7).

Remark

Theoretically, besides the CAOSR (on the *0*-order relation matrix) and the CAADOR (on the *1*-order relation matrix), other more types of relation regularization terms can be similarly developed based on higher-order relation matrices. However, in this work we consider just on the *0*-order and the *1*-order relation

[6] Due to space limitation as well, the $((w_i - w_{i+1})^T X_j)^*$ here represents $(w_i - w_{i+1})^T X_j + (b_i - b_{i+1}) e_{N_j}{}^T$, $i = 1, 2, ..., K-1; j = 1, 2, ..., K$.

matrices of the original CA coding matrix, due to that, 1) too many regularization terms usually come with proportional amount of hyper-parameters and thus heavily increase the computational complexity, 2) as mentioned in Sections 3.3 and 3.4, the CAOSR and CAADOR already can depict the explicit and implicit mutual relations between the CA codes, respectively, and more importantly 3) developing other much higher-order relation regularization terms, associated with corresponding order relation matrix, is trivial, because it can be achieved by referring to our strategies afore-mentioned.

4. Regularized mRR and mLS-SVR with the CAOSR and CAADOR

To validate the effectiveness of the two derived regularization terms, *CAOSR* and *CAADOR*, in incorporating the mutual relations between the CA codes into the learning and thus promoting the performance, we take the mRR as the first-layer regressor as in (Chen et al., 2013) by embedding them into the objective to regularize its learning. In addition, to evaluate the popularization ability of the terms, we also conduct evaluations in the framework of *large-margin learning* by taking the mLS-SVR as the first-layer regressor.

4.1. Regularized mRR

For the given augmented training set $\widetilde{X}$ (*cf.* Section 3.3) and its corresponding CA coding set Y, we can rewrite Eq. (1) of mRR as

$$\min_{\widetilde{w}_1,\widetilde{w}_2,\ldots,\widetilde{w}_K} \frac{1}{2}\sum_{k=1}^{K}\|Y^k-\widetilde{w}_k\widetilde{X}\|^2+\frac{\lambda}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2. \tag{8}$$

Then we can derive the regularized variant of mRR by adding the two regularization terms, $\mathcal{L}_{CAOSR}$ and $\mathcal{L}_{CAADOR}$, into (8) as

$$\min_{\widetilde{w}_1,\widetilde{w}_2,\ldots,\widetilde{w}_K} \frac{1}{2}\sum_{k=1}^{K}\|Y^k-\widetilde{w}_k\widetilde{X}\|^2+\frac{\lambda_1}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2+\frac{\lambda_2}{2}\mathcal{L}_{CAOSR}+\frac{\lambda_3}{2}\mathcal{L}_{CAADOR}, \tag{9}$$

where λ_1, λ_2 and λ_3 are regularization parameters. With the $\mathcal{L}_{CAOSR}$ and $\mathcal{L}_{CAADOR}$ being respectively substituted by Eqs. (5) and (7), Eq. (9) can then be rewritten as

$$\begin{aligned}\min_{\widetilde{w}_1,\widetilde{w}_2,\ldots,\widetilde{w}_K}\ & \frac{1}{2}\sum_{k=1}^{K}\|Y^k-\widetilde{w}_k\widetilde{X}\|^2+\frac{\lambda_1}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2+\frac{\lambda_2}{2}\sum_{k=1}^{K}\|\widetilde{w}_k^T\widetilde{X}R^k\|^2\\ &+\frac{\lambda_3}{2}\Big(\sum_{k=1}^{K-1}\sum_{c\neq k}^{K-1}\|(\widetilde{w}_k-\widetilde{w}_{k+1})^T\widetilde{X}(Y^c-Y^{c+1})^T\|^2\\ &+\sum_{k=1}^{K-1}\|(\widetilde{w}_k-\widetilde{w}_{k+1})^T\widetilde{X}(Y^K)^T\|^2\\ &+\sum_{k=K}^{K}\sum_{c\neq k}\|\widetilde{w}_k^T\widetilde{X}(Y^c-Y^{c+1})^T\|^2\Big).\end{aligned} \tag{10}$$

In order to solve for the $\widetilde{w}_k$, $k = 1, 2, ..., K$, we calculate the derivatives of Eq. (10) with respect to $\widetilde{w}_k$, $k = K, K-1, ..., 1$[7], and thus can obtain their analytical solutions as

$$\widetilde{w}_k = \begin{cases} \Big(\widetilde{X}\widetilde{X}^T + \lambda_1 I_{D+1} + \lambda_2 \widetilde{X} R^k (R^k)^T \widetilde{X}^T \\ +\lambda_3 \widetilde{X} \Upsilon_k \widetilde{X}^T\Big)^{-1} \Big(\lambda_3 \widetilde{X} \Upsilon_k \widetilde{X}^T \widetilde{w}_{k+1} + \widetilde{X}(Y^k)^T\Big), & k = 1, 2, ..., K-1 \\ \Big(\widetilde{X}\widetilde{X}^T + \lambda_1 I_{D+1} + \lambda_2 \widetilde{X} R^k (R^k)^T \widetilde{X}^T \\ +\lambda_3 \widetilde{X} \Upsilon_K \widetilde{X}^T\Big)^{-1} \Big(\widetilde{X}(Y^K)^T\Big), & k = K \end{cases} \tag{11}$$

where I_{D+1} is a (D+1)-order identity matrix, $\Upsilon_k = \sum_{c \neq k}^{K-1} (Y^c - Y^{c+1})^T (Y^c - Y^{c+1}) + (Y^K)^T (Y^K), k = 1, 2, ..., K-1$, and $\Upsilon_K = \sum_{c \neq K} (Y^c - Y^{c+1})^T (Y^c - Y^{c+1})$. And the complete procedure of solving the regularized mRR is summarized in Table 1.

Table 1: Algorithm for Regularized mRR.

Input:	Training data: $\widetilde{X}$ and Y with CA coding; Parameters: λ_1, λ_2, λ_3, and K.
Output:	$\widetilde{W} = [\widetilde{w}_1, ..., \widetilde{w}_K]$.
1. Compute $\widetilde{w}_K$ using (11) with the case of $k = K$;	
2. **for** $k = K-1, K-2, ..., 1$ **do**	
3. Compute $\widetilde{w}_k$ using (11) with the case of $k \neq K$.	
4. **end for**	

With obtained $\widetilde{W} = [\widetilde{w}_1, \widetilde{w}_2, ..., \widetilde{w}_K]$, for an unseen test instance $\widetilde{x_u}$, its CA coding vector y_u can be calculated through $y_u = \widetilde{W}^T \widetilde{x_u}$.

4.2. Regularized mLS-SVR

In accordance with (Bishop and Nasrabadi, 2006) and the notations afore-introduced, we write the mLS-SVR as

$$\begin{aligned} \min_{\widetilde{w}_1, \widetilde{w}_2, ..., \widetilde{w}_K} \quad & \frac{1}{2} \sum_{k=1}^{K} \|\widetilde{w}_k\|^2 + \frac{\lambda}{2} \sum_{k=1}^{K} \sum_{i=1}^{N} \xi_{k,i}^2 \\ s.t. \quad & y_{k,i} = \widetilde{w}_k^T \widetilde{x}_i + \xi_{k,i}, \; k \in \{1, 2, ..., K\}, i \in \{1, 2, ..., N\}. \end{aligned} \tag{12}$$

[7] From the following Eq. (11), it can be found that the analytical solution for $\widetilde{w}_k$ is involved with $\widetilde{w}_{k+1}$, except for the $\widetilde{w}_K$. Therefore, we can solve the $\widetilde{w}_k$ in descending order in terms of the subscript k, from K to 1.

Similar to (9), we introduce the two regularization terms, $\mathcal{L}_{CAOSR}$ and $\mathcal{L}_{CAADOR}$, into (12) as well and thus obtain its regularized variant as

$$\begin{aligned} \min_{\widetilde{w}_1,\widetilde{w}_2,...,\widetilde{w}_K} \ & \frac{1}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\xi_{k,i}^2 + \frac{\lambda_2}{2}\mathcal{L}_{CAOSR} + \frac{\lambda_3}{2}\mathcal{L}_{CAADOR} \\ s.t. \quad & y_{k,i} = \widetilde{w}_k^T\widetilde{x}_i + \xi_{k,i},\ k \in \{1,2,...,K\}, i \in \{1,2,...,N\}. \end{aligned} \tag{13}$$

After the $\mathcal{L}_{CAOSR}$ and $\mathcal{L}_{CAADOR}$ are substituted by (5) and (7), respectively, Eq. (13) then can be rewritten as

$$\begin{aligned} \min_{\widetilde{w}_1,\widetilde{w}_2,...,\widetilde{w}_K} \ & \frac{1}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\xi_{k,i}^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\|\widetilde{w}_k^T\widetilde{X}R^k\|^2 \\ & + \frac{\lambda_3}{2}\Big(\sum_{k=1}^{K-1}\sum_{c\neq k}^{K-1}\|(\widetilde{w}_k - \widetilde{w}_{k+1})^T\widetilde{X}(Y^c - Y^{c+1})^T\|^2 \\ & + \sum_{k=1}^{K-1}\|(\widetilde{w}_k - \widetilde{w}_{k+1})^T\widetilde{X}(Y^K)^T\|^2 \\ & + \sum_{k=K}^{K}\sum_{c\neq k}\|\widetilde{w}_k^T\widetilde{X}(Y^c - Y^{c+1})^T\|^2\Big). \\ s.t. \quad & y_{k,i} = \widetilde{w}_k^T\widetilde{x}_i + \xi_{k,i},\ k \in \{1,2,...,K\}, i \in \{1,2,...,N\}, \end{aligned} \tag{14}$$

in which $Y^k = [y_{k,1},\ y_{k,2},\ ...,\ y_{k,N}]$, $k = 1,2,...,K$. Eq. (14) is an equality-constrained convex optimization problem with analytical solution, which can be solved by means of Lagrangian multiplier theorem (Bishop and Nasrabadi, 2006). To this end, we introduce the Lagrangian multipliers $\alpha_{k,i}$, $k = 1,2,...,K$, $i = 1,2,...,N$, and consequently derive

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}\sum_{k=1}^{K}\|\widetilde{w}_k\|^2 + \frac{\lambda_1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\xi_{k,i}^2 + \frac{\lambda_2}{2}\sum_{k=1}^{K}\|\widetilde{w}_k^T\widetilde{X}R^k\|^2 \\ & + \frac{\lambda_3}{2}\Big(\sum_{k=1}^{K-1}\sum_{c\neq k}^{K-1}\|(\widetilde{w}_k - \widetilde{w}_{k+1})^T\widetilde{X}(Y^c - Y^{c+1})^T\|^2 \\ & + \sum_{k=1}^{K-1}\|(\widetilde{w}_k - \widetilde{w}_{k+1})^T\widetilde{X}(Y^K)^T\|^2 \\ & + \sum_{k=K}^{K}\sum_{c\neq k}\|\widetilde{w}_k^T\widetilde{X}(Y^c - Y^{c+1})^T\|^2\Big) \\ & - \sum_{k=1}^{K}\sum_{i=1}^{N}\alpha_{k,i}\Big(\widetilde{w}_k^T\widetilde{x}_i + \xi_{k,i} - y_{k,i}\Big). \end{aligned} \tag{15}$$

In accordance with the conditions of optimality of $\mathcal{L}$ with respect to w_k, $\xi_{k,i}$ and $\alpha_{k,i}$, respectively, *which are detailed in the Appendix*, we can obtain the

optimal solution as

$$\widetilde{w}_k = \begin{cases} \mathcal{H}_k^{-1}\Theta_k + \mathcal{H}_k^{-1}\widetilde{X}\mathcal{A}_k^{-1}\Big((Y^k)^T - \widetilde{X}^T\mathcal{H}_k^{-1}\Theta_k\Big), & k = 1, 2, ..., K-1 \\ \mathcal{H}_K^{-1}\widetilde{X}\mathcal{A}_K^{-1}(Y^K)^T, & k = K \end{cases} \tag{16}$$

where

$$\begin{aligned} \mathcal{H}_k &= I_{D+1} + \lambda_2\widetilde{X}R^k(R^k)^T\widetilde{X}^T + \lambda_3\widetilde{X}\sum_{c\neq k}^{K-1}\Big((Y^c - Y^{c+1})^T(Y^c - Y^{c+1})\Big)\widetilde{X}^T \\ &\quad + \lambda_3\widetilde{X}(Y^K)^T(Y^K)\widetilde{X}^T, \\ \Theta_k &= \lambda_3\widetilde{X}\sum_{c\neq k}^{K-1}\Big((Y^c - Y^{c+1})^T(Y^c - Y^{c+1})\Big)\widetilde{X}^T\widetilde{w}_{k+1} \\ &\quad + \lambda_3\widetilde{X}(Y^K)^T(Y^K)\widetilde{X}^T\widetilde{w}_{k+1}, \\ \mathcal{A}_k &= \widetilde{X}^T\mathcal{H}_k^{-1}\widetilde{X} + \frac{1}{\lambda_1}I_N, k = 1, 2, ..., K-1, \end{aligned}$$

$$\begin{aligned} \mathcal{H}_K &= I_{D+1} + \lambda_2\widetilde{X}R^K(R^K)^T\widetilde{X}^T + \lambda_3\widetilde{X}\sum_{c=1}^{K-1}\Big((Y^c - Y^{c+1})^T(Y^c - Y^{c+1})\Big)\widetilde{X}^T \\ \mathcal{A}_K &= \widetilde{X}^T\mathcal{H}_K^{-1}\widetilde{X} + \frac{1}{\lambda_1}I_N. \end{aligned}$$

As shown in (16), since the analytical solution of $\widetilde{w}_k$ is also involved with $\widetilde{w}_{k+1}$ (expect for the $\widetilde{w}_K$), therefore we compute $\widetilde{w}_k$ in a descending order in terms of the subscript k (from K to 1), as shown in Table 2.

Table 2: Algorithm for Regularized mLS-SVR.

Input:	Training data: $\widetilde{X}$ and Y with CA coding; Parameters: λ_1, λ_2, λ_3, and K.
Output:	$\widetilde{W} = [\widetilde{w}_1, ..., \widetilde{w}_K]$.
1. Compute $\widetilde{w}_K$ using (16) with the case of $k = K$;	
2. **for** $k = K-1, K-2, ..., 1$ **do**	
3. Compute $\widetilde{w}_k$ using (16) with the case of $k \neq K$.	
4. **end for**	

Then, for an unseen test instance $\widetilde{x_u}$, we can compute its CA coding vector y_u by performing $y_u = \widetilde{W}^T\widetilde{x_u}$ with $\widetilde{W} = [\widetilde{w}_1, \widetilde{w}_2, ..., \widetilde{w}_K]$.

4.3. Time Complexity Analysis

According to the algorithm for solving regularized mRR shown in Table 1, it can be analyzed that the time complexity of computing each $\widetilde{w}_k, k = 1, 2, ..., K-1$, is about $\mathcal{O}\big((D+1)^2(3N+D+3)+(D+1)(2N^2+N)+(K-1)N^2\big)$

$= \mathcal{O}\big(max(D^2N, D^3, DN^2, KN^2)\big)$ and it costs about $\mathcal{O}\big((D+1)^2(3N+D+2)+(D+1)(2N^2+N)+(K-1)N^2\big) = \mathcal{O}\big(max(D^2N, D^3, DN^2, KN^2)\big)$ for computing $\widetilde{w}_K$, where N is the total number of training samples, D denotes the feature dimension and K indicates the number of classes (e.g., the size of aging range). Therefore, the total complexity of solving regularized mRR is $\mathcal{O}\big(max(KD^2N, KD^3, KDN^2, K^2N^2)\big)$. As for regularized mLS-SVR, from Table 2 it can be analyzed that the time complexity of computing each $\widetilde{w}_k, k = 1, 2, ..., K-1$, is about $\mathcal{O}\big((D+1)^2(4N+D+2)+(D+1)(4N^2+2N)+(K-1)N^2+N^3\big) = \mathcal{O}\big(max(D^2N, D^3, DN^2, KN^2, N^3)\big)$ while it costs $\mathcal{O}\big((D+1)^2(4N+D+1)+(D+1)(4N^2+N)+(K-1)N^2+N^3\big) = \mathcal{O}\big(max(D^2N, D^3, DN^2, KN^2, N^3)\big)$ for computing $\widetilde{w}_K$. As a result, the total complexity of solving regularized mLS-SVR is $\mathcal{O}\big(max(KD^2N, KD^3, KDN^2, K^2N^2, KN^3)\big)$.

5. Experiments

In this section, we conduct experiments to validate the effectiveness of our regularization schemes in capturing the mutual relations among the CA codes to improve the CA-based human age estimation. Specifically, first we make a general comparison with related methods, by experimenting with data represented with high-level (i.e., Active Appearance Model) and low-level (i.e., raw-pixel) features, respectively; Then, to detailedly evaluate the ability of our strategies in improving CA-based age estimation with increasing training samples, age classes, especially when the aging range is incomplete with some ages lost, we conduct experiments with raw-pixel representation to eliminate the effect caused by high-level feature representation. Finally, through experiments on a larger human age dataset, we evaluate the effectiveness of the proposed strategies in improving larger scale age estimation.

5.1. Datasets and Settings

Prior to reporting the experimental results, we first make an introduction on the datasets used and experimental settings, respectively.

Datasets: In this work we conduct experiments on three commonly used benchmark aging datasets, i.e., FG-NET (Guo et al., 2008a), Morph Album 1 (Ricanek and Tesafaye, 2006), and Morph Album 2 (Ricanek and Tesafaye, 2006; Mu et al., 2009). The FG-NET dataset consists of 1,002 facial images taken from 82 individuals of European, and the age ranges from 0 to 69 years. As for the Morph Album 1, it contains 1,690 images from about 631 individuals mainly of African and European, and the age of the images ranges from 15 to 68 years. In regard to the Morph Album 2, it consists of about 5,475 white individuals as well as other ethnic individuals, averagely with about 2 to 3 pictures per person aging from 16 to 77. Image examples from the three datasets are shown in Figure 4.

Settings: In the experiments, all the hyper-parameters involved are tuned by using the *cross-validation* scheme depending on the specific dataset. And, we adopt the *mean absolute error* (MAE) as performance measure, in which

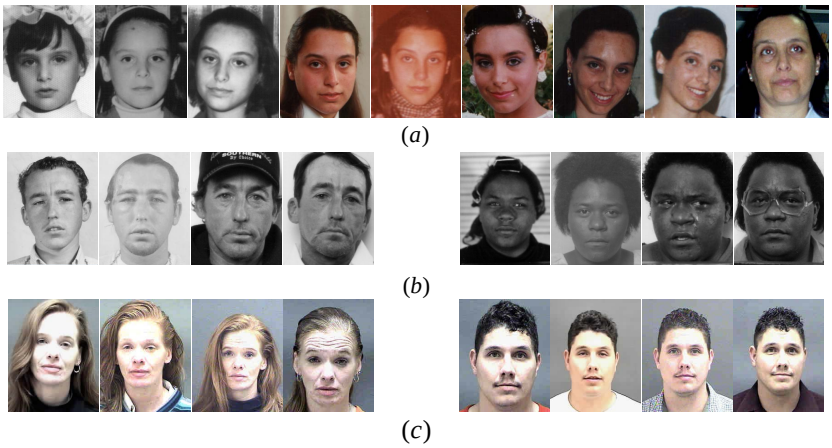


Figure 4: Image examples from the FG-NET (a), the Morph Album 1 (b) and the Morph Album 2 (c).

$MAE := \frac{1}{N}\sum_{i=1}^{N}|\widehat{l_i} - l_i|$ with l_i and $\widehat{l_i}$ denoting the ground-true and predicted values, respectively.

5.2. Evaluation on the Whole Aging Range

Firstly, to generally evaluate the effectiveness of the proposed regularization strategies as well as explore the effect of feature representation to human age estimation, we conduct comparative experiments with state-of-the-art methods. Specifically, we extract Active Appearance Model (AAM) feature representation due to its favourable performance on age estimation (Chen et al., 2013; Geng et al., 2007; Chang et al., 2011; Geng et al., 2013), as well as the raw-pixels from the FG-NET and the Morph Album 1, respectively and follow the same leave-one-person-out setting as in (Chen et al., 2013). Due to the state-of-the-art performance of CA-SVR proposed in (Chen et al., 2013), so here we just compare with it. Moreover, without loss of generality, we also adopt the SVR as the second-layer regressor. And, we report the experimental results in Tables 3 and 4, respectively. From the results, we can find that,

Table 3: Comparison of age estimation performance (MAE in years) with AAM representations.

(a) On FG-NET

first-layer regressor	mRR		mLS-SVR	
kernel type[1]	RBF	linear	RBF	linear
CA-SVR	4.42	5.62	4.40	5.59
Regularized CA-SVR	**4.38**	**5.52**	**4.37**	**5.52**

(b) On Morph Album 1

first-layer regressor	mRR		mLS-SVR	
kernel type	RBF	linear	RBF	linear
CA-SVR	4.73	4.98	4.72	4.96
Regularized CA-SVR	**4.71**	**4.93**	**4.69**	**4.92**

Table 4: Comparison of age estimation performance (MAE in years) with raw-pixel representations.

(a) On FG-NET

first-layer regressor	mRR		mLS-SVR	
kernel type	RBF	linear	RBF	linear
CA-SVR	6.54	7.26	6.53	7.24
Regularized CA-SVR	**6.41**	**7.08**	**6.42**	**7.03**

(b) On Morph Album 1

first-layer regressor	mRR		mLS-SVR	
kernel type	RBF	linear	RBF	linear
CA-SVR	5.82	5.92	5.83	5.90
Regularized CA-SVR	**5.73**	**5.70**	**5.75**	**5.71**

• The CA-based human estimations regularized with the proposed regularization terms are better than those without, especially when linear kernel is employed in the second-layer regression (mapping the CA coding to a scalar age label). For example, the MAE value is reduced by about 4 percentage points from 5.92 to 5.70, as shown in Table 4(b), which demonstrates the effectiveness of the proposed regularization strategy in improving CA-based age estimation.

• By comparing the MAEs reported in Table 3 with those in Table 4, it can be found that the formers are much lower than the latter ones. It demonstrates that such high-level feature representations as the AAM can significantly help improve the performance of age estimation, compared with those low-level features such as the raw-pixel representation.

5.3. Evaluation on Selected Aging Range

From the age distributions of the FG-NET and the Morph Album 1[8], demonstrated in Figures 5 and 6, we can find that the distributions are imbalanced. More concretely, the samples of FG-NET mainly range from 0 to about 19 years, while those of Morph Album 1 from 16 to around 38 years.

As a result, besides the evaluations reported in Section 5.2, below we conduct more empirical study on these two aging ranges to provide a detailed evaluation on the proposed regularization strategies with raw-pixels as feature representation, eliminating the performance effect caused by high-level feature representations such as the AAM. Specifically, from the FG-NET we randomly select 23 samples from each age ranging from 0 to 19 years, accounting for totally 20 age classes. As for the Morph Album 1 dataset, we pick 23 age classes from 16 to 38 years, each containing 31 samples. Then, on the selected FG-NET and Morph sets, we uniformly crop the interested face regions from the raw images and normalize them into 32×32 pixels based on eyes' centers. After that, we directly extract raw-pixels from the cropped face regions and apply the

[1]The *kernel type* in Tables 3 and 4 specifically indicates which kind of kernel function is employed in the second-layer regression.

[8]Since the age distribution of Morph Album 2 is similar to that of Album 1, so here we make evaluations just on the FG-NET and the Morph Album 1.

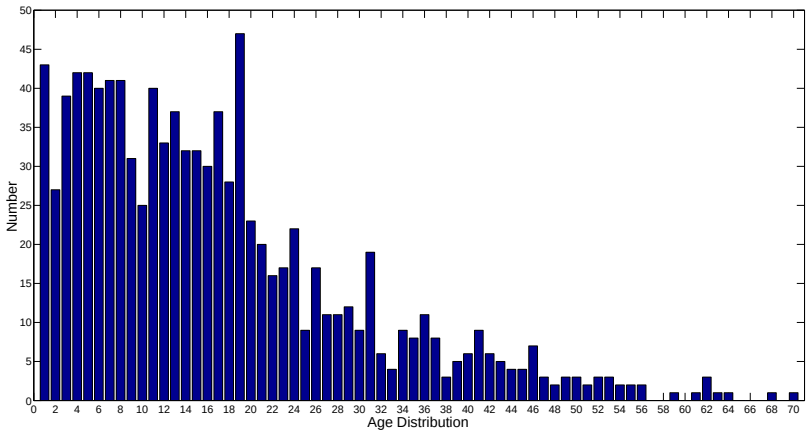


Figure 5: Sample distributions of the ages on FG-NET.

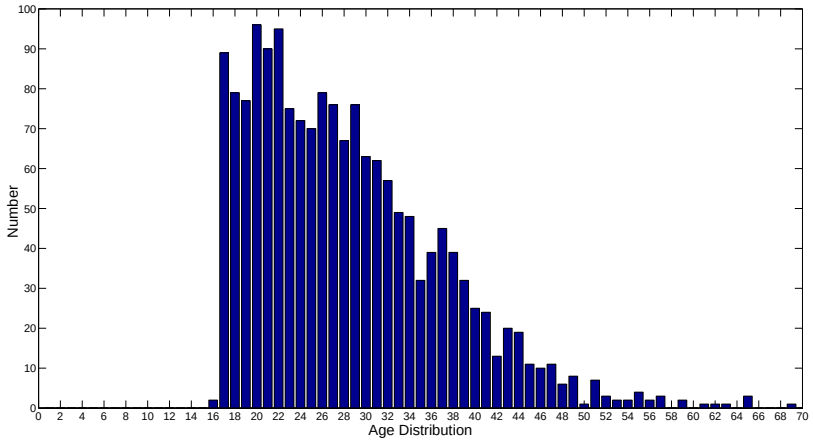


Figure 6: Sample distributions of the ages on Morph Album 1.

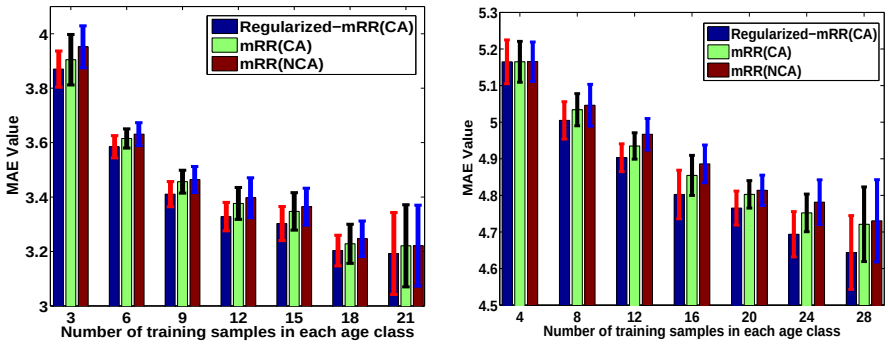
PCA (Jolliffe, 2005) to acquire about 95 percentage components as the feature representations.

With the extracted features, below we perform experiments to explore the ability of the proposed regularizations in improving the age estimation with increasing training samples, age classes, and especially when the aging range is incomplete with some ages lost randomly, respectively.

5.3.1. Evaluation with Increasing Training Samples From Each Age

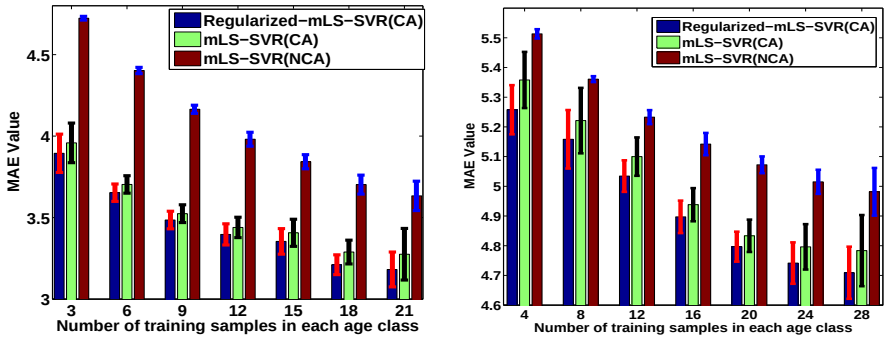
In this subsection we explore the performance improvement of the proposed regularizations with increasing training samples from each age. Specifically, we randomly select certain number of samples from each age class for training and the rest for test, and report the averaged results over 10 runs in Figures 7 and

8, respectively. From them[9], we can find that,



(a) on FG-NET (b) on Morph Album 1

Figure 7: Comparison between results of mRR and its variants.



(a) on FG-NET (b) on Morph Album 1

Figure 8: Comparison between results of mLS-SVR and its variants.

• With the number of training samples in each age class increasing, MAEs of all methods including the prototypes and their variants all decrease. It witnesses that an increasing number of training samples generally can improve the generalization ability of a regressor, which is consistent with statistical learning theorem (Vapnik, 1998).

• MAEs of the methods with CA coding, mRR(CA) and mLS-SVR(CA), are correspondingly higher than those of mRR(NCA) and mLS-SVR(NCA) with

[9]In the legends of Figures 7 and 8, the *CA* (or *NCA*) indicates that current method is associated with *cumulative attribute* (or *non-cumulative attribute*) coding output in the first-layer regression, and it refers to the same meaning in the legends of the following figures if any.

NCA coding. Again, it shows the superiority of CA coding in human age estimation, compared with NCA.

• The regularized variants of mRR and mLS-SVR with the CAOSR and CAADOR as well as the CA coding both can yield better age estimation results (with relatively lower MAEs) than corresponding non-regularized mRR(CA) and mLS-SVR(CA), respectively. It demonstrates the effectiveness of the CAOSR and CAADOR in optimizing CA learning and thus improving the age estimation. More importantly, it indicates that besides the regression framework, our regularization strategies can also work well in the large-margin learning since mLS-SVR is designed within it.

5.3.2. Evaluation with Increasing Age Classes

In this subsection, we make detailed evaluations on the two proposed CA-oriented regularization terms, CAOSR and CAADOR, with increasing age classes (i.e., increasing aging range). More concretely, we conduct experiments by selecting increasing sizes of aging range from the FG-NET and Morph with the size from 4 (i.e., 0 to 3 years on FG-NET and 16 to 19 years on Morph Album 1, respectively) to 16 (i.e., 0 to 15 years on FG-NET and 16 to 31 years on Morph Album 1, respectively), in which about 75 percentage of the samples randomly selected for training and the rest for test in each age, and show the averaged results over 5 runs from Tables 5 to 8 as below, in which the results written in bold followed by "•" show they have significant superiority in performance after the *t-test* with the *p-value* set at 0.05. From them, we can find that,

Table 5: Performance comparison (MAE ± STD, in years) with increasing size of aging range on FG-NET with the mRR as the first-layer regressor.

size of aging range	mRR	mRR (with CAOSR)	mRR (with CAADOR)	mRR (with CAOSR+CAADOR)
4	1.0858 ± 0.0568	**0.9840 ± 0.0681**	1.0265 ± 0.0633	**0.9561 ± 0.0739•**
6	1.2458 ± 0.0865	1.2302 ± 0.0699	1.2298 ± 0.0731	**1.1701 ± 0.0826**
8	1.4158 ± 0.0687	1.4100 ± 0.0702	1.4056 ± 0.0869	1.4032 ± 0.0965
10	1.7390 ± 0.0496	1.6857 ± 0.0562	1.6973 ± 0.0638	**1.6489 ± 0.0539•**
12	2.1008 ± 0.0635	**2.0265 ± 0.0701**	2.0761 ± 0.0639	**2.0169 ± 0.0890**
14	2.3132 ± 0.1032	2.2589 ± 0.0861	2.2690 ± 0.1005	**2.2188 ± 0.0981**
16	2.6460 ± 0.0967	2.6049 ± 0.0895	2.6051 ± 0.0868	2.5704 ± 0.0994

Table 6: Performance comparison (MAE ± STD, in years) with increasing size of aging range on FG-NET with the mLS-SVR as the first-layer regressor.

size of aging range	mLS-SVR	mLS-SVR (with CAOSR)	mLS-SVR (with CAADOR)	mLS-SVR (with CAOSR+CAADOR)
4	0.9686 ± 0.0824	0.9206 ± 0.0806	0.9358 ± 0.0807	**0.8589 ± 0.0795•**
6	1.2158 ± 0.1068	1.1968 ± 0.0908	1.2005 ± 0.0799	**1.1562 ± 0.0980**
8	1.4124 ± 0.0936	1.4105 ± 0.0936	1.3969 ± 0.0991	1.3941 ± 0.0856
10	1.7062 ± 0.0657	1.6755 ± 0.0821	1.6821 ± 0.0737	1.6598 ± 0.0902
12	2.0855 ± 0.1032	**2.0214 ± 0.0980**	2.0689 ± 0.1068	**2.0185 ± 0.1072**
14	2.3028 ± 0.1005	2.2668 ± 0.1162	2.2864 ± 0.1082	**2.2201 ± 0.1036**
16	2.6452 ± 0.0869	2.6088 ± 0.0989	2.6081 ± 0.1251	**2.5715 ± 0.1098**

Table 7: Performance comparison (MAE ± STD, in years) with increasing size of aging range on Morph Album 1 with the mRR as the first-layer regressor.

size of aging range	mRR	mRR (with CAOSR)	mRR (with CAADOR)	mRR (with CAOSR+CAADOR)
4	1.2068 ± 0.0486	**1.0760 ± 0.0504•**	1.1908 ± 0.0419	**1.0732 ± 0.0467•**
6	1.5189 ± 0.0425	**1.4509 ± 0.0436•**	1.5019 ± 0.0618	**1.4474 ± 0.0401•**
8	1.9559 ± 0.0268	**1.8608 ± 0.0360•**	1.9462 ± 0.0263	**1.8592 ± 0.0485•**
10	2.3356 ± 0.0492	2.3121 ± 0.0382	2.3268 ± 0.0401	2.3082 ± 0.0371
12	2.6825 ± 0.0901	2.6625 ± 0.0725	2.6726 ± 0.0913	**2.6330 ± 0.0604**
14	3.1062 ± 0.0682	3.0668 ± 0.0701	3.0852 ± 0.0565	3.0579 ±0.0895
16	3.4408 ± 0.0983	3.4095 ± 0.1030	3.4207 ± 0.892	**3.3765 ± 0.0916**

Table 8: Performance comparison (MAE ± STD, in years) with increasing size of aging range on Morph Album 1 with the mLS-SVR as the first-layer regressor.

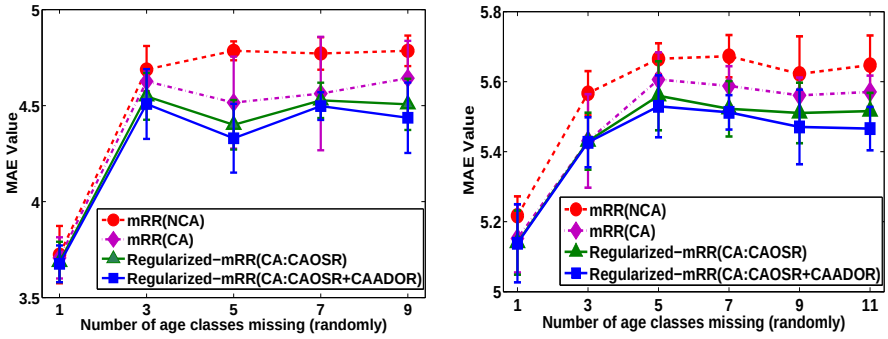
size of aging range	mLS-SVR	mLS-SVR (with CAOSR)	mLS-SVR (with CAADOR)	mLS-SVR (with CAOSR+CAADOR)
4	1.1986 ± 0.0435	**1.0757 ± 0.0551•**	1.1865 ± 0.0435	**1.0703 ± 0.0527•**
6	1.5135 ± 0.0504	**1.4520 ± 0.0403**	1.4989 ± 0.0568	**1.4389 ± 0.0404•**
8	1.9560 ± 0.0266	**1.8609 ± 0.0483•**	1.9353 ± 0.0259	**1.8590 ± 0.0481•**
10	2.3350 ± 0.0485	2.3117 ± 0.0382	2.3157 ± 0.0385	2.3082 ± 0.0357
12	2.6811 ± 0.0895	2.6625 ± 0.0768	2.6753 ± 0.0869	**2.6325 ± 0.0634**
14	3.1001 ± 0.1035	3.0659 ± 0.0977	3.0853 ± 0.1035	3.0571 ± 0.0980
16	3.4388 ± 0.1151	3.4101 ± 0.1032	3.4304 ± 0.0993	**3.3721 ± 0.1016**

• Generally, the MAEs (listed in the last column of the Tables 5 to 8) of the CA-based age estimation regularized with both the CAOSR and CAADOR are lower than those without the proposed regularization terms (listed in the first column of the four tables), and the improvement is significant when the size of aging range is relatively small, especially on the Morph dataset no matter which estimator is employed as the first-layer regressor, either the mRR or the mLS-SVR. For example, in all cases with the size of aging range equal to 4, the MAEs yielded by regularizing with both the CAOSR and the CAADOR are reduced by over 10 percentage points from those without the regularizations. It further demonstrates the effectiveness of our strategies in capturing the code-between mutual relations and thus improving the CA learning.

• By comparing between the MAEs yielded by regularizing just with the CAOSR (corresponding to $\lambda_3 = 0$ in Eqs. (11) and (16)) and those just with the CAADOR (corresponding to $\lambda_2 = 0$ in Eqs. (11) and (16)), it can be found that regularizing with the CAOSR can relatively improve the CA-based age estimation with much lower MAEs than that just with the CAADOR, although regularizing with both of them simultaneously can always yield the best performance. Therefore, the results show that the CAOSR is preferable to the CAADOR in terms of improving performance (additionally, the computational complexity of the former is also much lower than that of the latter, by comparing their respective formulations Eq. (5) and Eq. (7)).

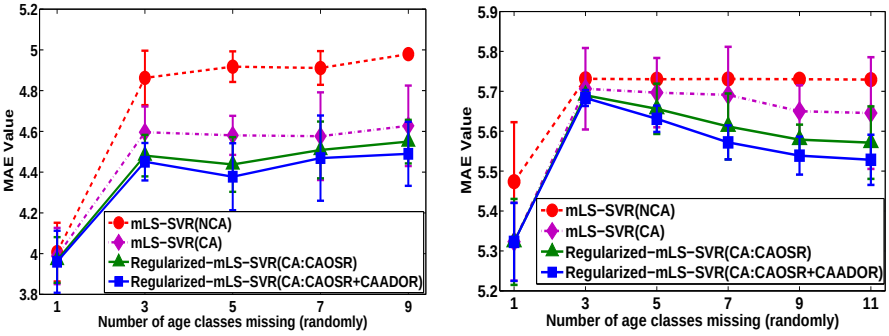
5.3.3. Evaluation on Incomplete Aging Range with Some Ages Lost

In real-world, usually the human age distribution of samples collected is imbalanced or say incomplete with some ages lost (i.e., missing), just as shown in Figures 5 and 6. So, next we evaluate the ability of the proposed regularizations in improving the CA-based age estimation in such cases. To this end, we conduct experiments by eliminating certain number of age classes randomly to simulate such a scenario and report the averaged results over 10 runs in Figures 9 and 10, with below analyses,



(a) on FG-NET (b) on Morph Album 1

Figure 9: Performance comparison between mRR and its variants with increasing number of age classes that are randomly eliminated.



(a) on FG-NET (b) on Morph Album 1

Figure 10: Performance comparison between mRR and its variants with increasing number of age classes that are randomly eliminated.

• With the increasing number of age classes lost (or say missing), the MAEs of all the methods generally grow higher, especially the ones using NCA coding, which implies that incomplete distribution of the aging range affects the age

estimation unfavorably. In spite of this, the methods with CA coding still generate relatively lower MAEs than those with NCA coding, which witnesses the superiority of CA coding over the NCA, in handling such age classes-incomplete cases.

• The regularized mRR and mLS-SVR just combined with the CAOSR (corresponding to $\lambda_3 = 0$ in the (11) and (16)), which we denote as *Regularized-mRR(CA:CAOSR)* and *Regularized-mLS-SVR(CA:CAOSR)*, respectively, can effectively promote age estimation with relatively lower MAEs than the non-regularized. It shows the effectiveness of the CAOSR in improving the CA learning even when the age classes are distributed incomplete.

• The regularized mRR and mLS-SVR incorporated with both the CAOSR and CAADOR, i.e., *Regularized-mRR(CA:CAOSR+CAADOR)* and *Regularized-mLS-SVR(CA:CAOSR+CAADOR)*, yield the lowest MAEs, which demonstrates that besides the CAOSR, the CAADOR can further help perform age estimation with incomplete age distributions, although its significance relatively is not so great as the CAOSR.

5.4. Evaluation on Larger Aging Dataset

Besides the above evaluations, next we conduct experiments on the relatively larger Morph Album 2 dataset to evaluate the effectiveness of the proposed regularization strategies. Specifically, we extract the bio-inspired features (BIF) (Mu et al., 2009) from face images of white individuals and then acquire about 95% principal components as feature representations. With the extracted representations, we randomly select 10 samples[10] from each age for training, and the remaining for test. In addition, besides the MAE criterion, here we also adopt the *Cumulative Score* (*CS*) (Geng et al., 2006) as another performance measure which is defined as $CS(j) = N_{e\leq j}/N \times 100\%$ with N being the total number of test samples while $N_{e\leq j}$ the number of test samples on which the age estimation makes an absolute error no higher than j years. And we report the average results over 10 random runs in Tables 9 and 10.

Table 9: Performance comparison $\big(\text{MAE} \pm \text{STD}, CS(j)\big)$ between mRR and its regularized variants (as the first-layer regressor) on Morph Album 2.

aging range (j of $CS(j)$)	mRR	mRR (with CAOSR)	mRR (with CAOSR+CAADOR)
16-19 (2)	1.1496 ± 0.0619 (71)	**0.9789 ± 0.0009 (79)•**	**0.9358 ± 0.0105 (83)•**
16-23 (4)	2.0374 ± 0.0818 (85)	**1.7780 ± 0.0039 (91)•**	**1.6959 ± 0.0078 (93)•**
16-27 (6)	2.7569 ± 0.0898 (90)	**2.6950 ± 0.0184 (95)**	**2.6387 ± 0.0151 (97)•**
16-35 (6)	4.2329 ± 0.1050 (70)	4.2036 ± 0.1104 (71)	**4.1789 ± 0.1039 (73)**
16-55 (6)	5.3037 ± 0.0785 (60)	5.2597 ± 0.0987 (61)	5.2209 ± 0.0753 (63)
16-77 (6)	5.7387 ± 0.0862 (56)	5.7085 ± 0.0806 (57)	5.6741 ± 0.0915 (58)

[10]If the total number of samples in some age is less than 10, we select all the samples for training.

Table 10: Performance comparison $\big($MAE $\pm$ STD, $CS(j)\big)$ between mLS-SVR and its regularized variants (as the first-layer regressor) on Morph Album 2.

aging range (j of $CS(j)$)	mLS-SVR	mLS-SVR (with CAOSR)	mLS-SVR (with CAOSR+CAADOR)
16-19 (2)	1.1302 ± 0.0507 (72)	**0.9708 ± 0.0008 (79)•**	**0.9438 ± 0.0021 (83)•**
16-23 (4)	1.9839 ± 0.0365 (86)	**1.7806 ± 0.0016 (90)•**	**1.7308 ± 0.0019 (92)•**
16-27 (6)	2.7052 ± 0.0693 (90)	2.6800 ± 0.0703 (91)	**2.6507 ± 0.0603 (92)**
16-35 (6)	4.2037 ± 0.1035 (70)	4.1803 ± 0.1134 (70)	4.1507 ± 0.1307 (72)
16-55 (6)	5.1893 ± 0.0708 (62)	5.1686 ± 0.0705 (62)	5.1039 ± 0.0547 (63)
16-77 (6)	5.5968 ± 0.0651 (58)	5.5639 ± 0.0572 (58)	5.5389 ± 0.0657 (59)

From the results shown in Tables 9 and 10, it can be found that the age estimations of regularized methods by our strategies are all superior to those of the original methods, either the mRR or the mLS-SVR as first-layer regressor. Particularly, when the size of aging range is less than 12 (e.g., from 16 to 27 years), the performance improvement is significant in MAE by about 13% to 18% reduction, or in CS by about 6% to 12% increase. Therefore, it demonstrates the effectiveness of the proposed strategies in capturing the mutual relations between CA codes and thus improving age estimation on the Morph Album 2.

6. Conclusions

In this work, in order to exploit and incorporate the mutual relations existing explicitly or implicitly between the CA codes into learning on human age estimation, we first derived the so-called *0*-order and *1*-order relation matrices by performing the difference-like operations on the CA codes, and formulated them correspondingly as two regularization terms, coined as the CAOSR and the CAADOR, respectively. Then, we embedded the two regularization terms into the objective of the mRR and mLS-SVR which act as the first-layer regressor in CA learning (as shown in Figure 1), by which the mutual relations between the CA codes are artfully taken into their learning. Finally, we experimentally evaluated the effectiveness of the regularization terms in depicting the mutual relations and thus improving the CA learning, with the conclusions that 1) both the CAOSR and CAADOR are effective in improving the CA-based age estimation, especially when the size of aging range is not large (e.g., not greater than 8 in our experiments), 2) the CAOSR is relatively preferable to the CAADOR in terms of improving performance, and 3) our regularization strategies can work as well in cases of aging range incomplete. More importantly, besides the CA coding, the proposed regularization strategies can be similarly extended to other types of codings such as ECOC (Ciompi et al., 2014), etc.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China under Grant 61472186, Funding of Jiangsu Innovation Program for

Graduate Education under Grant $CXLX13\_159$, Fundamental Research Funds for the Central Universities and Jiangsu Qinglan Project.

Appendix:

According to the Eq. (15) and Lagrangian optimization theorem (Bishop and Nasrabadi, 2006), the weight vectors $\widetilde{w}_1, \widetilde{w}_2, ..., \widetilde{w}_K$ here can be calculated separately in descending order w.r.t. the index k from $\widetilde{w}_K$ down to $\widetilde{w}_1$. Specifically, we solve for $\widetilde{w}_k, k = K, K-1, ..., 1$, by calculating the derivatives on Eq. (15) respectively w.r.t. $\widetilde{w}_k$, ξ_k and α_k and obtaining their saddle points as

a) For $k = K$,

$$\frac{\partial \mathcal{L}}{\partial \widetilde{w}_K} = \Big(I_{D+1} + \lambda_2 \widetilde{X} R^K (R^K)^T \widetilde{X}^T + \lambda_3 \widetilde{X} \textstyle\sum_{c \neq K} \big((Y^c - Y^{c+1})^T (Y^c - Y^{c+1})\big) \widetilde{X}^T\Big) \widetilde{w}_K - \widetilde{X} \alpha_K = 0;$$

$$\frac{\partial \mathcal{L}}{\partial \xi_K} = \lambda_1 \xi_K - \alpha_K = 0;$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_K} = \widetilde{w}_K^T \widetilde{X} + \xi_K^T - Y^K = 0.$$

Solving the three equations above together results in the solution for $\widetilde{w}_K$.

b) For $k = K-1, K-2, ..., 1$,

$$\frac{\partial \mathcal{L}}{\partial \widetilde{w}_k} = \Big(I_{D+1} + \lambda_2 \widetilde{X} R^k (R^k)^T \widetilde{X}^T + \lambda_3 \widetilde{X} \textstyle\sum_{c \neq k} \big((Y^c - Y^{c+1})^T (Y^c - Y^{c+1})\big) \widetilde{X}^T\Big) \widetilde{w}_k - \lambda_3 \widetilde{X} \textstyle\sum_{c \neq k} \big((Y^c - Y^{c+1})^T (Y^c - Y^{c+1})\big) \widetilde{X}^T \widetilde{w}_{k+1} - \widetilde{X} \alpha_k = 0;$$

$$\frac{\partial \mathcal{L}}{\partial \xi_k} = \lambda_1 \xi_k - \alpha_k = 0;$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = \widetilde{w}_k^T \widetilde{X} + \xi_k^T - Y^k = 0.$$

Similarly, solving the three equations above together results in the solution for $\widetilde{w}_k, k = K-1, K-2, ..., 1$.

References

Alnajar, F., Shan, C., Gevers, T., Geusebroek, J.M., 2012. Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions. Image and Vision Computing 30, 946–953.

An, S., Liu, W., Venkatesh, S., 2007. Face recognition using kernel ridge regression, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE. pp. 1–7.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. volume 1. springer New York.

Chang, K.Y., Chen, C.S., Hung, Y.P., 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 585–592.

Chen, K., Gong, S., Xiang, T., Mary, Q., Loy, C.C., 2013. Cumulative attribute space for age and crowd density estimation, in: Pattern Recognition, 2013. CVPR 2013. IEEE 26th International Conference on, IEEE. pp. 2467–2474.

Ciompi, F., Pujol, O., Radeva, P., 2014. Ecoc-drf: Discriminative random fields based on error correcting output codes. Pattern Recognition 47, 2193–2204.

Fu, Y., Xu, Y., Huang, T.S., 2007. Estimating human age by manifold analysis of face pictures and regression on aging features, in: Multimedia and Expo, 2007 IEEE International Conference on, IEEE. pp. 1383–1386.

Garcia, S., Derrac, J., Cano, J.R., Herrera, F., 2012. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 417–435.

Geng, X., Yin, C., Zhou, Z.H., 2013. Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2401–2412.

Geng, X., Zhou, Z.H., Smith-Miles, K., 2007. Automatic age estimation based on facial aging patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29, 2234–2240.

Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H., 2006. Learning from facial aging patterns for automatic age estimation, in: Proceedings of the 14th annual ACM international conference on Multimedia, ACM. pp. 307–316.

Guo, G., Fu, Y., Dyer, C.R., Huang, T.S., 2008a. Image-based human age estimation by manifold learning and locally adjusted robust regression. Image Processing, IEEE Transactions on 17, 1178–1188.

Guo, G., Fu, Y., Huang, T.S., Dyer, C.R., 2008b. Locally adjusted robust regression for human age estimation, in: Applications of Computer Vision, 2008. WACV 2008. IEEE Workshop on, IEEE. pp. 1–6.

Jain, A.K., Dass, S.C., Nandakumar, K., 2004. Soft biometric traits for personal recognition systems, in: Biometric Authentication. Springer, pp. 731–738.

Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.

Kohli, S., Prakash, S., Gupta, P., 2013. Hierarchical age estimation with dissimilarity-based classification. Neurocomputing 120, 164–176.

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

Lanitis, A., Taylor, C.J., Cootes, T.F., 2002. Toward automatic simulation of aging effects on face images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 442–455.

Li, C., Liu, Q., Liu, J., Lu, H., 2012a. Learning distance metric regression for facial age estimation, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE. pp. 2327–2330.

Li, C., Liu, Q., Liu, J., Lu, H., 2012b. Learning ordinal discriminative features for age estimation, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. pp. 2570–2577.

Li, C., Liu, Q., Liu, J., Lu, H., 2014. Ordinal distance metric learning for image ranking. IEEE Transactions on Neural Networks and Learning Systems , to appear.

Luu, K., Ricanek, K., Bui, T.D., Suen, C.Y., 2009. Age estimation using active appearance models and support vector machine regression, in: Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on, IEEE. pp. 1–5.

Mahajan, D., Sellamanickam, S., Nair, V., 2011. A joint learning framework for attribute models and object descriptions, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE. pp. 1227–1234.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to linear regression analysis. volume 821. Wiley.

Mu, G., Guo, G., Fu, Y., Huang, T.S., 2009. Human age estimation using bio-inspired features, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 112–119.

Ricanek, K., Tesafaye, T., 2006. Morph: A longitudinal image database of normal adult age-progression, in: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE. pp. 341–345.

Romano J.R., N.C., Fjermestad, J., 2006. Electronic customer relationship management. Electronic customer relationship management 3, 1.

Sai, P.K., Wang, J.G., Teoh, E.K., 2015. Facial age range estimation with extreme learning machines. Neurocomputing 149, 364–372.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14, 199–222.

Ueki, K., Hayashida, T., Kobayashi, T., 2006. Subspace-based age-group classification using facial images under various lighting conditions, in: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE. pp. 6–pp.

Vapnik, V.N., 1998. Statistical learning theory .

Yan, S., Wang, H., Huang, T.S., Yang, Q., Tang, X., 2007a. Ranking with uncertain labels, in: Multimedia and Expo, 2007 IEEE International Conference on, IEEE. pp. 96–99.

Yan, S., Wang, H., Tang, X., Huang, T.S., 2007b. Learning auto-structured regressor from uncertain nonnegative labels, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE. pp. 1–8.