 CMC, vol.56, no.3, pp.467-482, 2018

# Feature Relationships Learning Incorporated Age Estimation Assisted by Cumulative Attribute Encoding

**Qing Tian[1, 2, 3, *], Meng Cao[1, 2] and Tinghuai Ma[1, 2]**

**Abstract:** The research of human facial age estimation (AE) has attracted increasing attention for its wide applications. Up to date, a number of models have been constructed or employed to perform AE. Although the goal of AE can be achieved by either classification or regression, the latter based methods generally yield more promising results because the continuity and gradualness of human aging can naturally be preserved in age regression. However, the neighbor-similarity and ordinality of age labels are not taken into account yet. To overcome this issue, the cumulative attribute (CA) coding was introduced. Although such age label relationships can be parameterized via CA coding, the potential relationships behind age features are not incorporated to estimate age. To this end, in this paper we propose to perform AE by encoding the potential age feature relationships with CA coding via an implicit modeling strategy. Besides that, we further extend our model to gender-aware AE by taking into account gender variance in aging process. Finally, we experimentally validate the superiority of the proposed methodology.

**Keywords:** Age estimation, cumulative attribute, gender-aware age estimation, correlation relationship learning.

## 1 Introduction

Human facial age estimation (AE) is an important research topic and attracted a lot of attention because of its applications in such as age-oriented product and service recommendation [Fjermestad and Romano (2006)], safety monitoring [Guo, Fu, Dyer et al. (2008); Lanitis, Taylor and Cootes (2004)], identity recognition [Jain, Dass and Nandakumar (2004)], etc. For the task of predicting human facial age, it is to estimate the old degree of a face image given its appearance representations (i.e. feature representations). To achieve the goal of AE, great numbers of models have been developed or employed so far. Summing up, they can be grouped into classification based models [Lanitis, Taylor and Cootes (2004); Geng, Yin and Zhou (2013); Alnajar, Shan, Gevers et al. (2012); Sai, Wang and Teoh (2015)], regression-based methods [Lanitis, Taylor and Cootes (2002); Fu, Xu and Huang (2007); Luu, Ricanek, Bui et al. (2009);

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[2] Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[3] School of Electrical and Electronic Engineering, University of Manchester, Manchester, M13 9PL, UK.

[*] Corresponding Author: Qing Tian. Email: tianqing@nuaa.edu.cn; qing.tian@manchester.ac.uk.

Yan, Wang, Tang et al. (2007); Yan, Wang, Huang et al. (2007); Geng, Zhou and Smith-miles (2007); Chang, Chen and Huang (2011); Li, Liu, Liu et al. (2012a, 2012b); Li (2012); Tian, Xue and Qiao (2016); Tian and Chen (2017)] and hybrid-based models [Guo, Fu, Dyer et al. (2008); Guo, Fu, Dyer et al. (2008); Kohli, Prakash and Gupta (2013)]. In the framework of classification based AE, each age is treated as a separated class and by this way the facial age of a given face is typically predicted using the trained AE classifier. Along this line, the authors of Lanitis et al. [Lanitis, Draganova and Christodoulou (2004)] extracted coefficients of active appearance models as feature representations and then employed neural networks to classify the facial age. The authors of Ueki et al. [Ueki, Hayashida and Kobayashi (2006)] constructed a mixed Gaussian model to perform age-group classification. More recently, the conditional probability networks model (abbreviated as CPNN) was generated by the authors of Geng et al. [Geng, Yin and Zhou (2013)] to perform natural AE by incorporating the distributions information of neighboring ages. Further, the authors of Geng et al. [Geng, Yin and Zhou (2013)] proposed fuzzy-set based soft-AE by taking into account the similarities of neighboring ages. Besides, the ELM was also introduced to classify nonlinear age patterns [Sai, Wang and Teoh (2015)]. Motivated by the great success of deep learning paradigm, these deep models with layers of nonlinear representation learning were successively presented to more precisely classify the appearance age of a given face image [Niu, Zhou, Wang et al. (2016); Yang, Gao, Xing et al. (2016); Xing, Li, Hu et al. (2017)].

In fact, human facial appearance is aging gradually, implying that the problem of AE should be more regarded as a regression problem. To this end, the quadratic functions [Lanitis, Taylor and Cootes (2002)] were adopted to regress facial age by function fitting. And, the authors of Fu et al. [Fu, Xu and Huang (2007)] extended AE with multiple directions of linear regression. To eliminate the impact of outlier age patterns, the soft-version of SVR model [Luu, Ricanek, Bui et al. (2009)] was also employed for age regression, and more complex SDP models were later adopted to handle AE with annotation noises [Yan, Wang, Tang et al. (2007); Yan, Wang, Huang et al. (2007)]. Motivated by the fact that some age patterns are usually missed in most of current benchmark aging databases, component regression type model-AGES was constructed to regress missing age patterns to benefit AE accuracy [Geng, Zhou and Smith-Miles (2007)]. Although the reviewed methods above are reasonable themselves from their motivations, nearly all of these methods ignore the ordinal characteristic in the process of facial appearance aging, thus leading to non-optimal AE accuracy [Chen, Gong, Xiang et al. (2013); Chang, Chen and Huang (2011); Tian, Xue and Qiao (2016)]. To cater such a characteristic, the authors in Chang et al. [Chang, Chen and Hung (2011)] modeled an ordinal regressor called OHRank for age regression. Later, the authors in Li et al. [Li, Liu, Liu et al. (2012)] proposed to construct an ordinal metric space to benefit subsequent AE decision making. Then, Li et al. [Li, Liu, Liu et al. (2012)] extended feature selection learning to AE by removing redundant features. Beyond the work of Li et al. [Li, Liu, Liu et al. (2012)], the authors of Li et al. [Li, Liu, Liu et al. (2015)] modeled an ordinal distance space with taking into account both the ordinality and distribution structure of human aging sequence.


**Figure 1:** Ordinality and similarity of neighboring ages in facial appearance. The digit under each image indicates its facial age

In fact, apart from the ordinal relationships of age classes, another characteristic of them is that neighboring facial appearances look more similar than those further distributed, just as shown in Fig. 1. As claimed in Tian et al. [Tian, Xue and Qiao (2016); Tian and Chen (2017)] that incorporating these two characteristics of ages in AE learning is beneficial to improve the generalization performance of learned AE estimator. To nobly model the ordinality and similarity of neighboring ages, the authors in Chen et al. [Chen, Gong, Xiang et al. (2013)] constructed the so-called cumulative attribute (CA) representation to encode the age labels. Concretely, they first mapped original feature representation of a face image into CA coding, and then regressed its corresponding scalar age value from the CA coding. The whole pipeline of cumulative attribute coding based age estimation is demonstrated in Fig. 2. Although such a coding scheme yielded better AE results than the previous ones, it just can model the relationships among age semantic labels but ignores the underlying relationships among the original feature representations. Actually, underlying correlations may exist among the entries of original feature representations due to that they are extracted from the same face image. Moreover, the distribution space of face images typically follows some mixed Gaussian distributions [Bocklet, Maier, Bauer et al. (2008)]. Motivated by this fact, in this paper, we propose to model the underlying relationships among age original feature representations nobly by correlation matrix scheme with desirable analytical solutions. By this way, the mutual relationships among the original features can be characterized and incorporated in AE learning. Moreover, we further extend the model to gender-aware scenarios by incorporating gender variance between male and female. Finally, through experiments, we demonstrate the effectiveness of the proposed method. The main contributions of the paper are four-fold as follows:

1. *Constructing a correlation learning model for age estimation (AE) to automatically exploit the underlying relationships among facial representations with CA coding.*
2. *Extending the proposed model to gender-aware AE scenario by incorporating gender variance.*
3. *Providing closed-form solutions via alternating optimization for the proposed models.*
4. *Experimentally validating the effectiveness of the proposed methods.*

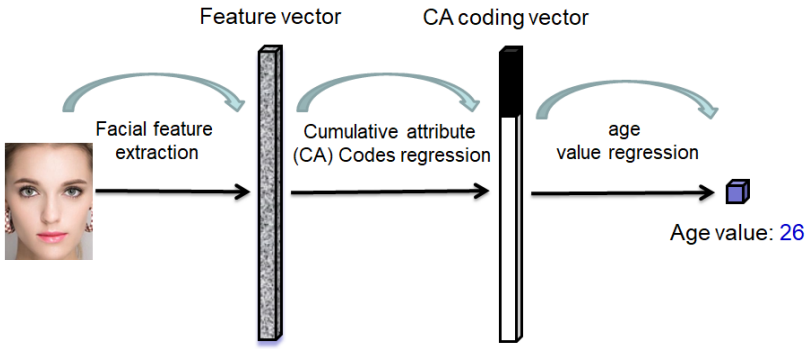


**Figure 2:** The pipeline of cumulative attribute coding based age estimation

The rest of this paper is organized as follows. Section 2 presents the background contents of CA coding based AE. Section 3 introduces the proposed method. Section 4 reports the validation results. Finally, Section 5 gives the conclusions and future works.

## 2 Background

For the problem of AE, assume we adopt the regression function $f(x) = w^T x + b$ to predict the age value for a given face image $x \in R$, with $w \in R$ and $b \in R$ being the regression coefficients vector and bias. To obtain the parameters $\{w, b\}$ of the above regression function, we need to optimize the following objective function on the given training instances and their age labels $\{x_i, l_i\}_{i=1}^N$ as follows:

$$\min_{w,b} \quad \frac{1}{2}\sum_{i=1}^{N}\left\| l_i - \left(w^T x_i + b\right)\right\|_2^2 + \frac{\lambda}{2}\|w\|_2^2, \tag{1}$$

where the first term indicates the empirical loss on the training set, the second term is dedicated to control the complexity of the model, and $\lambda$ are predefined regularization coefficient. However, as analyzed in the introduction section, Eq. (1) is just a general regression function without particularly preserving the ordinality and neighbor-similarity of ages. To overcome this issue, the concept of cumulative attribute (CA) coding was proposed in Chen et al. [Chen, Gong, Xiang et al. (2013)]. We assume the scale of human aging sequence is 100, then the length $K$ of CA coding is 100. According to Chen et al. [Chen, Gong, Xiang et al. (2013)], each element of CA coding vector $y_i \in R^K$ of an instance $x_i$ is defined as

$$y_i^j = \begin{cases} 1, & j \le l_i \\ 0, & j > l_i \end{cases} \tag{2}$$

Through incorporating Eq. (2) into Eq. (1), Eq. (1) becomes

$$\min_{W,b} \quad \frac{1}{2}\sum_{i=1}^{N}\left\| y_i - \left(W^T x_i + b\right)\right\|_F^2 + \frac{\lambda}{2}\|W\|_F^2, \tag{3}$$

where $W=\left[w_1,\ \ w_2,\ \ \cdots,\ \ w_d\right]^T \in R^{d\times K}$ is the regression coefficient matrix and $b\in R^K$ stands for the bias vector. After solving the parameters of Eq. (3), we can obtain the CA coding vector of an instance and then generate its age label by off-the-self regression models.

## 3 The proposed methodology

### *3.1 Feature relationships leaning incorporated AE with CA coding*

As claimed in the introduction section, although the CA coding can encode the ordinality and neighbor-similarity of age labels, the underlying correlations between the feature representations of face samples are not taken into account, leaving a perform space can be filled. Fortunately, the distribution space of face instances typically comply with some Gaussian distributions [Bocklet, Maier, Bauer et al. (2008)]. Motivated by this fact and the Gaussian modeling scheme [Yu and Yeung (2012)], we here propose to adopt the correlation matrix (denoted as $\Omega$) to characterize the mutual relationships between the age feature representations. Inspired by the fact that the feature presentations of an instance is operated inter-product with the regression coefficient matrix (see Eq. (3)), we can resort to model the mutual relationships between the feature representations of the instances nobly by regularizing the correlations between the regression coefficient matrix $W\in R^{d\times K}$ using correlation matrix $\Omega$.

**Objective Function.** Through the above analyzed modeling scheme, the objective function of feature relationships leaning incorporated AE with the assistance of CA coding can be formulated as

$$\begin{aligned}&\min_{W,b,\Omega}\ \ \frac{1}{2}\sum_{i=1}^{N}\left\|y_i-\left(W^Tx_i+b\right)\right\|_F^2+\frac{\lambda_1}{2}\left\|W\right\|_F^2+\frac{\lambda_2}{2}tr\left(W^T\Omega^{-1}W\right),\\&s.t.\ \ \ \Omega\succeq 0,\\&\qquad tr\left(\Omega\right)=1,\end{aligned}\tag{4}$$

where $W=\left[w_1,\ \ w_2,\ \ \cdots,\ \ w_d\right]^T \in R^{d\times K}$ is the projection coefficient matrix, $b\in R^K$ is the bias vector, $\Omega\in R^{d\times d}$ is the correlation matrix used to model the mutual relationships between the feature rows of $W$. The constraints are dedicated to control the complexity of the model.

Because of the semi-definiteness of $\Omega$, it thus can be characterized the underlying relationships between the rows (corresponding to each component of the feature representations) of projection coefficient matrix. More specifically, because the element $\Omega_{i,j}$ accordingly characterizes the correlations between the *i*-th and *j*-th components of age feature representations, so its magnitude and signal (positive, negative or zero) happen to recognize the degree and type (*positive-related, negative-related or unrelated*) of mutual relationships between the two feature components. In turn, these recognized correlations can be used to regularize the learning of the projection matrix. By this way, the underlying correlations between the age feature representations can be incorporated to benefit the AE. More importantly, such a modeling strategy can automatically recognize

the correlation degrees and types of the feature representations from the training samples, without requiring any prior knowledge. The proposed modeling strategy helps AE escapes from the drawbacks invoked by human defined unnatural correlation priors.

**Optimization.** It can be found that objective function in Eq. (4) is biconvex w.r.t. the variables $\{W,b,\Omega\}$. Concisely, the objective function is convex w.r.t. $\{W,b\}$ when $\Omega$ is fixed, and convex w.r.t. $\Omega$ if $\{W,b\}$ are fixed. To obtain the variables in Eq. (4), we can thus take an alternating optimization algorithm to solve it.

- With fixed $\Omega$, Eq. (4) essentially becomes

$$\min_{W,b} \quad \frac{1}{2}\sum_{i=1}^{N}\left\|y_i-\left(W^T x_i+b\right)\right\|_F^2+\frac{\lambda_1}{2}\|W\|_F^2+\frac{\lambda_2}{2}tr\left(W^T\Omega^{-1}W\right). \tag{5}$$

Calculating the gradients of Eq. (5) w.r.t. $\{W,b\}$ yields

$$\begin{aligned}
&\frac{\partial G}{\partial W}=\left(\sum_{i=1}^{N}x_i x_i^T+\lambda_1 I_d+\lambda_2\Omega^{-1}\right)W+\sum_{i=1}^{N}x_i b^T-\sum_{i=1}^{N}x_i y_i^T=0\\
&\frac{\partial G}{\partial b}=\sum_{i=1}^{N}x_i^T W+Nb^T-\sum_{i=1}^{N}y_i^T=0\\
&\Rightarrow\begin{bmatrix}W\\ b^T\end{bmatrix}=\left(Q^TQ\right)^{-1}Q^TP,
\end{aligned} \tag{6}$$

where $I_d$ is an identity matrix and

$$Q=\begin{pmatrix}\sum_{i=1}^{N}x_i x_i^T+\lambda_1 I_d+\lambda_2\Omega^{-1} & \sum_{i=1}^{N}x_i\\ \sum_{i=1}^{N}x_i^T & N\end{pmatrix},P=\begin{pmatrix}\sum_{i=1}^{N}x_i y_i^T\\ \sum_{i=1}^{N}y_i^T\end{pmatrix}.$$

The solutions of $\{W,b\}$ can be obtained analytically from Eq. (6).

- With fixed $\{W,b\}$, Eq. (4) can essentially be rewritten as

$$\begin{aligned}
&\min_{\Omega} \quad tr\left(W^T\Omega^{-1}W\right)\\
&s.t. \quad \Omega\succeq 0\\
&\quad tr\left(\Omega\right)=1,
\end{aligned} \tag{7}$$

which also has analytical solution as

$$\Omega=\frac{\left(WW^T\right)^{\frac{1}{2}}}{tr\left(\left(WW^T\right)^{\frac{1}{2}}\right)}. \tag{8}$$

In order to cater for the constraints of Eq. (4), $\Omega$ should be properly initialized. For this, we assume all the feature components between training samples are unrelated, so $\Omega$ can

be initially set to $\Omega=\frac{1}{d}I_d$. Then, we repeat the above two operation steps alternatingly until maximal iterations meet or convergence. With the obtained $\{W,b\}$, we can perform facial image AE.

### *3.2 Gender-aware AE with incorporating within- and between-gender feature relationships*

Although the feature correlations between age feature representations are incorporated to benefit AE, the gender variance between human male and female is not taken into account yet. Related research [Tian and Chen (2018)] shows that the male and female are aging significantly differently, which affect the performance of AE. Motivated by this fact, we here take into account such underlying gender variance to AE by incorporating the underlying correlations between human male and female. To this end, we need to separate the loss term in Eq. (4) for male and female samples, respectively. Besides that, in order to model the underlying relationships *within-gender* and *between-gender* simultaneously, we here also construct correlation matrices to characterize them.

**Objective Function.** Based on the analysis above, we can accordingly formulate the objective function of gender-aware AE as follows

$$\begin{aligned}\min_{W_m,W_f,b_m,b_f,\Omega}\quad & \frac{1}{2}\sum_{i=1}^{N_m}\left\|y_{mi}-\left(W_m{}^T x_{mi}+b_m\right)\right\|_F^2+\frac{1}{2}\sum_{i=1}^{N_f}\left\|y_{fi}-\left(W_f{}^T x_{fi}+b_f\right)\right\|_F^2 \\ & +\frac{\lambda_1}{2}\left\|\begin{bmatrix}W_m\\W_f\end{bmatrix}\right\|_F^2+\frac{\lambda_2}{2}tr\left(\begin{bmatrix}W_m\\W_f\end{bmatrix}^T\Omega_{mf}{}^{-1}\begin{bmatrix}W_m\\W_f\end{bmatrix}\right)\end{aligned} \tag{9}$$

$$\begin{aligned}s.t.\quad & \Omega_{mf}\succeq 0,\\ & tr\left(\Omega_{mf}\right)=1,\end{aligned}$$

where $W_m=\left[w_{m1},\ \cdots,\ w_{md}\right]^T\in R^{d\times K}$, $W_f=\left[w_{f1},\ \cdots,\ w_{fd}\right]^T\in R^{d\times K}$, $b_m\in R^K$ and $b_f\in R^K$ are the projection coefficient matrices and biases for male and female, respectively; $N_m$ and $N_f$ denote the numbers of male and female training samples, while $\{W_m,b_m\}$ and $\{W_f,b_f\}$ denote the projection coefficient matrices and biases for male and female, respectively. The first and second terms are introduced to denote the empirical losses for male and female, respectively. The third term is dedicated to penalize the model complexity. More importantly, as the fourth term showing, the age feature correlations **within** male and female, together with those **between** male and female are characterized by $\Omega_{mf}\in R^{2d\times 2d}$ in a unified regularization term.

Eq. (9) can be simplified by introducing projection matrix $W$ to unify $W_m$ and $W_f$ as

$$\begin{aligned}\min_{W,b_m,b_f,\Omega_{mf}} \quad & \frac{1}{2}\sum_{i=1}^{N_m}\left\|y_{mi}-\left(\left(A_mW\right)^T x_{mi}+b_m\right)\right\|_F^2+\frac{1}{2}\sum_{i=1}^{N_f}\left\|y_{fi}-\left(\left(A_fW\right)^T x_{fi}+b_f\right)\right\|_F^2 \\ & +\frac{\lambda_1}{2}\|W\|_F^2+\frac{\lambda_2}{2}tr\left(W^T\Omega_{mf}{}^{-1}W\right) \end{aligned} \tag{10}$$

$$s.t. \quad \Omega_{mf}\succeq 0,$$

$$tr\left(\Omega_{mf}\right)=1,$$

where $W=\begin{bmatrix}W_m\\W_f\end{bmatrix}\in R^{2d\times K}$, $A_m=\begin{bmatrix}I_{d\times d} & 0_{d\times d}\end{bmatrix}\in R^{d\times 2d}$,

and $A_f=\begin{bmatrix}0_{d\times d} & I_{d\times d}\end{bmatrix}\in R^{d\times 2d}$.

**Optimization.** Obviously, Eq. (10) is also biconvex w.r.t. to all variables, thus we can also take an alternating optimization algorithm to solve it.

- With fixed $\Omega_{mf}$, Eq. (10) can be written as

$$\begin{aligned}\min_{W,b_m,b_f} \quad & \frac{1}{2}\sum_{i=1}^{N_m}\left\|y_{mi}-\left(\left(A_mW\right)^T x_{mi}+b_m\right)\right\|_F^2 \\ & +\frac{1}{2}\sum_{i=1}^{N_f}\left\|y_{fi}-\left(\left(A_fW\right)^T x_{fi}+b_f\right)\right\|_F^2 \\ & +\frac{\lambda_1}{2}\|W\|_F^2+\frac{\lambda_2}{2}tr\left(W^T\Omega_{mf}{}^{-1}W\right)\end{aligned} \tag{11}$$

We calculate the derivatives of Eq. (11) w.r.t. $\{W,b_m,b_f\}$ and let them to zeros, yielding that

$$\begin{aligned}&\frac{\partial G_{mf}}{\partial W}=\left(\sum_{i=1}^{N_m}A_m{}^T x_{mi}x_{mi}{}^T A_m+\sum_{i=1}^{N_f}A_f{}^T x_{fi}x_{fi}{}^T A_f+\lambda_1 I_{2d}+\lambda_2\Omega_{mf}{}^{-1}\right)W \\ &+\sum_{i=1}^{N_m}A_m{}^T x_{mi}b_m{}^T+\sum_{i=1}^{N_f}A_f{}^T x_{fi}b_f{}^T-\sum_{i=1}^{N_m}A_m{}^T x_{mi}y_{mi}{}^T-\sum_{i=1}^{N_f}A_f{}^T x_{fi}y_{fi}{}^T=0 \\ &\frac{\partial G_{mf}}{\partial b_m}=\sum_{i=1}^{N_m}x_{mi}{}^T A_mW+N_mb_m{}^T-\sum_{i=1}^{N_m}y_{mi}{}^T=0 \\ &\frac{\partial G_{mf}}{\partial b_f}=\sum_{i=1}^{N_f}x_{fi}{}^T A_fW+N_fb_f{}^T-\sum_{i=1}^{N_f}y_{fi}{}^T=0 \\ &\Rightarrow\begin{bmatrix}W\\b_m^T\\b_f^T\end{bmatrix}=\left(Q^TQ\right)^{-1}Q^TP,\end{aligned} \tag{12}$$

where

$$Q=\begin{pmatrix} \sum_{i=1}^{N_m} A_m{}^T x_{mi} x_{mi}{}^T A_m + \sum_{i=1}^{N_f} A_f{}^T x_{fi} x_{fi}{}^T A_f & \sum_{i=1}^{N_m} A_m{}^T x_{mi} & \sum_{i=1}^{N_f} A_f{}^T x_{fi} \\ +\lambda_1 I_{2d} + \lambda_2 \Omega_{mf}{}^{-1} & & \\ \sum_{i=1}^{N_m} x_{mi}{}^T A_m & N_m & 0 \\ \sum_{i=1}^{N_f} x_{fi}{}^T A_f & 0 & N_f \end{pmatrix}, P=\begin{pmatrix} \sum_{i=1}^{N_m} A_m{}^T x_{mi} y_{mi}{}^T \\ +\sum_{i=1}^{N_f} A_f{}^T x_{fi} y_{fi}{}^T \\ \sum_{i=1}^{N_m} y_{mi}{}^T \\ \sum_{i=1}^{N_f} y_{fi}{}^T \end{pmatrix}$$

- With fixed $\{W, b_m, b_f\}$, Eq. (10) then can be reformulated as

$$\begin{aligned} &\min_{\Omega_{mf}} \quad tr\left(W^T \Omega_{mf}{}^{-1} W\right), \\ &s.t. \quad \Omega_{mf} \succeq 0, \\ &\qquad tr\left(\Omega_{mf}\right) = 1, \end{aligned} \tag{13}$$

which in form has the same closed-form solution as Eq. (7).

We repeat the above two steps until maximal iterations meet or convergence. Then, all the model variables of Eq. (10) can be obtained. It should be noted that the initialization of $\Omega_{mf}$ is set to $\Omega_{mf} = \frac{1}{2d} I_{2d}$ different from that in Section 3.1.

## 4 Experiment

### *4.1 Datasets and setting*

To validate effectiveness of the proposed method, we conduct age estimation experiments on three widely-used benchmark datasets, i.e. FG-NET [Guo, Fu, Dyer et al. (2008)], Morph Album I and Album II [Geng, Yin and Zhou (2013)]. Specially, FG-NET consists of 1,002 images taken from 82 individuals aging from 0 to 69; Morph Album I contains 1,690 facial pictures from 631 African and European people with age ranging from 15 to 68; the Morph Album II database is relatively large and has over 55,000 face pictures aged between 16 and 77. Image examples of the three datasets are shown in Fig. 3.

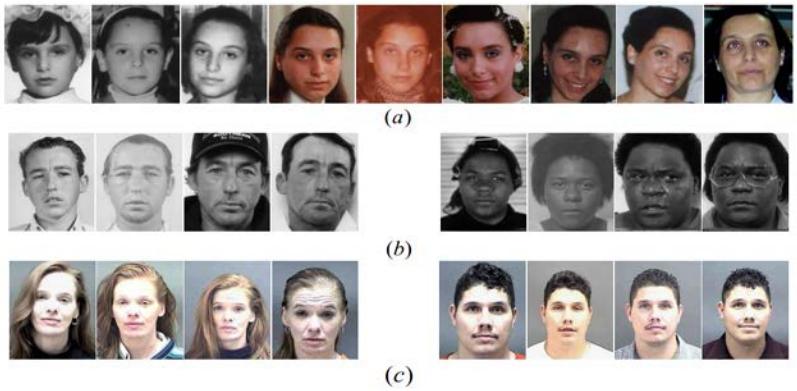


**Figure 3:** Image examples from FG-NET (a), Morph Album I (b) and Album II (c)

For experimental setting, we extract AAM from FG-NET and Morph Album I and BIF from Morph Album II as their feature representations, respectively. All the regularization parameters are tuned through five-fold cross-validation in the searching range $\{1e-5, \ \ldots, \ 1e5\}$. Following related literature [Geng, Yin and Zhou (2013); Fu, Xu and Huang (2007)], we also take the MAE criterion as performance measurement. For the second stage regression from CA codings to age labels, we uniformly adopt the nearest class center regressor for final AE decision making.

### *4.2 AE Experiment disregarding the gender-variance*

To validate the effectiveness of the proposed method, we first perform AE experiments without particularly considering the aging difference between human male and female. More concisely, we randomly select 10%-90% of feature extracted datasets for training and the remaining for test. We repeat the runs ten times with random data splitting. The experimental results are shown in Tab. 1.

We can see from the above results that the proposed method consistently reduces the MAEs by up to 2% (from 5.81 to 5.67 on FG-NET), especially when the percent of training set is less than 40%. This finding shows that when the percent of training samples is low, so limited scale of training samples are not abundant to meet the whole distributions of sample feature space, leading to non-optimal solutions or even biased solutions without proper regularizations. By contrast, when the correlations between feature representations of limited training set are embedded (as a regularizer) in the objective function of AE learning for joint optimization, the underlying distribution relationships between the features can be taken into account to regularize the solutions of AE to better performance. On the other hand, with increasing training samples, the natural spatial distributions and correlations between feature representations can be increasingly covered by the training samples themselves. Besides, we also research the magnitudes and types of the correlations between sample representing features in a visual manner in Fig. 4.

**Table 1:** Age estimation (AE) results (MAE±STD) comparison on the three datasets

| | FG-NET | | Morph Album Ⅰ | | Morph Album Ⅱ | |
|---|---|---|---|---|---|---|
| Training Percent | CA coding based AE | **ours** | CA coding based AE | **ours** | CA coding based AE | **ours** |
| 10% | 5.81±0.36 | **5.67±0.34** | 5.11±0.12 | **5.02±0.12** | 5.13±0.09 | **5.06±0.11** |
| 20% | 5.34±0.18 | **5.29±0.17** | 4.83±0.09 | **4.75±0.09** | 4.89±0.04 | **4.80±0.03** |
| 30% | 5.02±0.17 | **4.96±0.12** | 4.74±0.07 | **4.70±0.06** | 4.79±0.07 | **4.73±0.06** |
| 40% | 4.91±0.25 | **4.86±0.17** | 4.68±0.05 | **4.67±0.05** | 4.74±0.07 | **4.72±0.07** |
| 50% | 4.84±0.25 | **4.80±0.19** | 4.64±0.07 | **4.63±0.07** | 4.71±0.08 | **4.70±0.08** |
| 60% | 4.74±0.24 | **4.72±0.06** | 4.60±0.12 | **4.59±0.12** | 4.68±0.07 | **4.67±0.07** |
| 70% | 4.69±0.31 | **4.68±0.32** | 4.61±0.16 | **4.60±0.16** | 4.66±0.09 | **4.65±0.08** |
| 80% | 4.80±0.50 | **4.79±0.51** | 4.57±0.25 | **4.56±0.25** | 4.69±0.13 | **4.68±0.13** |
| 90% | 4.71±0.66 | **4.71±0.65** | 4.55±0.34 | **4.53±0.33** | 4.72±0.15 | **4.71±0.15** |

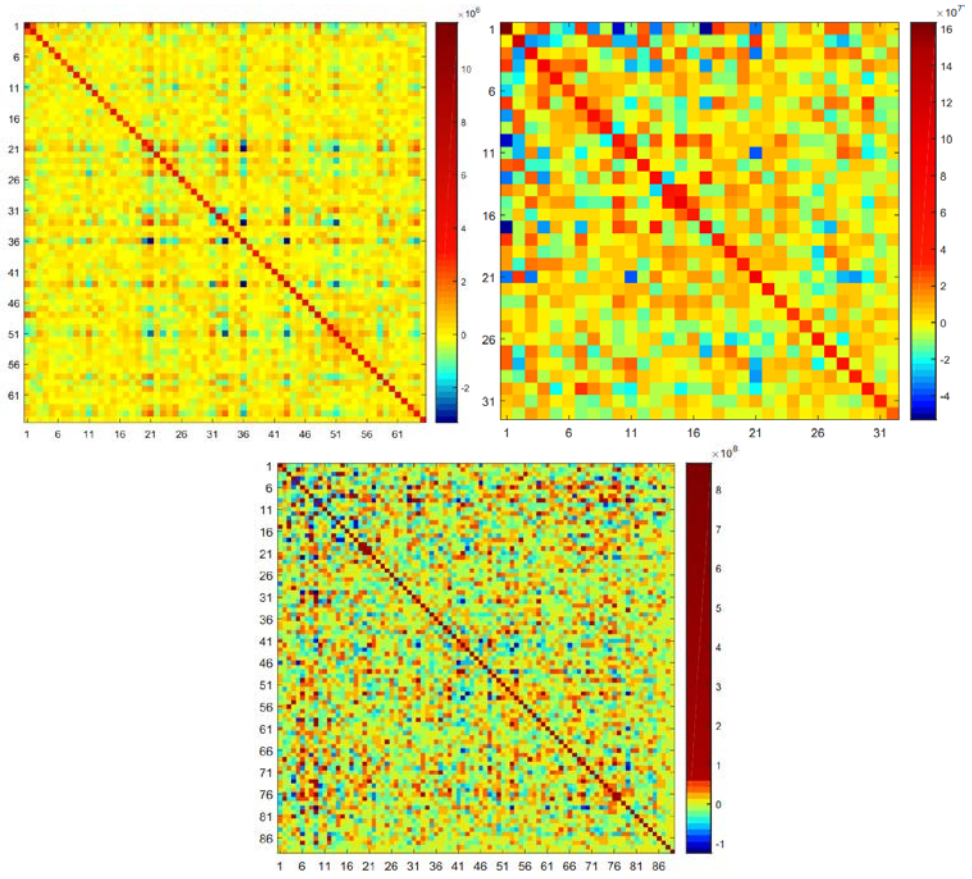

**Figure 4:** Visualization of the learned average correlation matrix between the age feature representations on FG-NET (*left*), Morph Album I (*middle*) and Album II (*right*)

From the correlation visualizations, we can discover that the correlation magnitudes between different elements of age features are not identical but varying. Moreover, these correlations can be grouped into three types: positive-related (*corresponding elements of correlation matrix are positive*), negative-related (*corresponding elements of correlation matrix are negative*), or unrelated (*corresponding elements of correlation matrix are zeros*). More interestingly, they are learned automatically from training data with human interference.

### *4.3 Gender-aware AE experiment*

**Table 2:** Gender-aware age estimation (AE) results (MAE±STD) comparison on Morph Album Ⅰ

| | On Entire test set | | On Male test set | | On Female test set | |
|---|---|---|---|---|---|---|
| Training Percent | CA coding based AE | **ours** | CA coding based AE | **ours** | CA coding based AE | **ours** |
| 10% | 5.04±0.08 | **4.98±0.10** | 4.88±0.06 | **4.85±0.09** | **5.76±0.27** | 5.77±0.25 |
| 20% | 4.83±0.10 | **4.77±0.06** | 4.65±0.10 | **4.60±0.06** | 5.70±0.21 | **5.51±0.26** |
| 30% | 4.70±0.10 | **4.65±0.10** | 4.54±0.13 | **4.52±0.11** | 5.46±0.27 | **5.31±0.20** |
| 40% | 4.60±0.06 | **4.57±0.06** | 4.46±0.09 | **4.45±0.09** | 5.24±0.17 | **5.21±0.18** |
| 50% | 4.54±0.07 | **4.53±0.07** | 4.42±0.08 | **4.41±0.08** | 5.10±0.28 | 5.10±0.27 |
| 60% | 4.51±0.08 | **4.50±0.08** | 4.40±0.09 | 4.40±0.09 | 5.04±0.20 | **5.00±0.27** |
| 70% | 4.48±0.11 | **4.46±0.11** | 4.39±0.12 | **4.38±0.12** | **4.88±0.23** | 4.89±0.24 |
| 80% | 4.43±0.17 | **4.42±0.17** | 4.37±0.19 | 4.37±0.19 | 4.72±0.41 | **4.66±0.40** |
| 90% | 4.37±0.26 | **4.36±0.26** | 4.30±0.33 | 4.30±0.33 | 4.68±0.65 | **4.67±0.63** |

**Table 3:** Gender-aware age estimation (AE) results (MAE±STD) comparison on Morph Album Ⅱ

| | On Entire test set | | On Male test set | | On Female test set | |
|---|---|---|---|---|---|---|
| Training Percent | CA coding based AE | **ours** | CA coding based AE | **ours** | CA coding based AE | **ours** |
| 10% | 4.79±0.06 | **4.77±0.07** | 4.41±0.07 | **4.38±0.07** | 5.96±0.21 | **5.93±0.23** |
| 20% | 4.52±0.05 | **4.51±0.05** | 4.23±0.05 | **4.21±0.06** | 5.41±0.12 | 5.41±0.12 |
| 30% | 4.42±0.05 | **4.41±0.05** | 4.16±0.04 | **4.15±0.06** | 5.20±0.11 | 5.20±0.07 |
| 40% | 4.36±0.04 | **4.35±0.04** | 4.11±0.04 | **4.10±0.04** | 5.13±0.13 | **5.08±0.09** |
| 50% | 4.34±0.03 | **4.31±0.03** | 4.10±0.04 | **4.07±0.04** | 5.06±0.11 | 5.06±0.11 |
| 60% | 4.29±0.03 | **4.28±0.04** | 4.07±0.04 | **4.05±0.04** | 4.97±0.11 | **4.96±0.10** |
| 70% | 4.29±0.05 | **4.26±0.07** | 4.07±0.05 | **4.04±0.07** | 4.97±0.16 | **4.92±0.16** |
| 80% | 4.30±0.06 | **4.26±0.07** | 4.09±0.05 | **4.04±0.05** | 4.94±0.16 | **4.91±0.18** |
| 90% | 4.28±0.08 | **4.22±0.08** | 4.07±0.10 | **4.03±0.10** | 4.92±0.20 | **4.81±0.19** |

To further evaluate the proposed method in handling gender-variance to AE, we conduct gender-aware AE experiments on the Morph Album I and II, both of which are given gender annotations besides the age labels. We also repeat the runs ten times with random data splitting and report the averaged MAE and standard deviations in Tab. 2 and 3.

We can see from the comparative results that with increasing training samples, in all cases, our proposed method have reduced the MAEs against the original method just with CA coding, on the entire, male and female test sets. These results again demonstrate the effectiveness of the proposed methodology. Interestingly, the MAEs of AE on female are higher than those on male. This is mainly due to that the percentage of female samples is significantly lower than that of male on the Morph Album I and II datasets. So, if we extend these datasets with more female samples, the MAEs of AE on female will be reduced to comparable with that on male.

We also visually demonstrate the underlying correlations between age features within-gender and between-gender in Fig. 5.

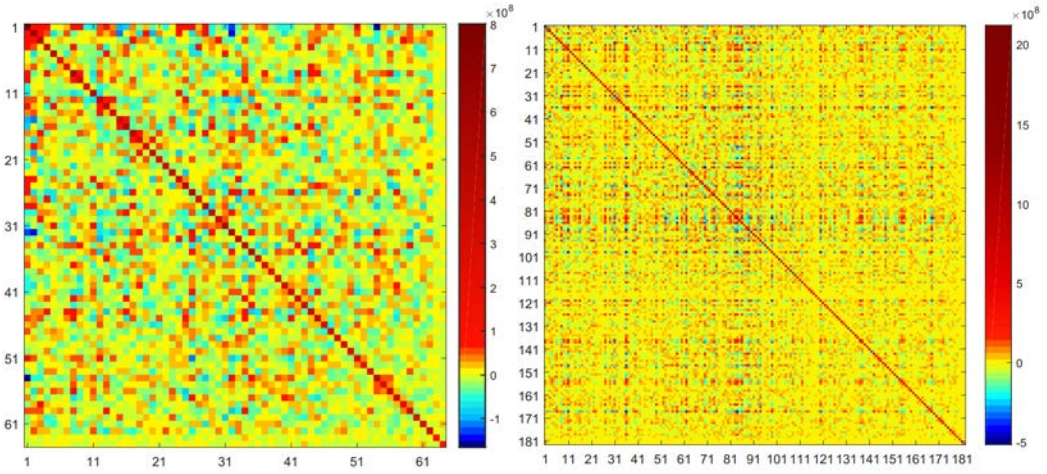

**Figure 5:** Visualization of the learned average correlation matrix between the age feature representations on Morph Album I (*left*) and Album II (*right*)

We can see from Fig. 5 that not only the features **within** male and female are correlated, but those **between** male and female have also relationships. And the types of these relationships are *positive-related, negative-related or unrelated*.

## 5 Conclusion

In this paper, we proposed an AE methodology by encoding the potential age feature relationships with CA coding via an implicit modeling strategy. Although the types of relationships (*positive-related, negative-related, or unrelated*) between age features are not given in advance, they can be recognized nobly through the proposed model. To incorporate the aging difference between human male and female, we also extended the proposed model to gender-aware AE by taking into account the gender variance. Besides, we also presented alternating optimization algorithms to solve the models with analytical solutions. Finally, we experimentally demonstrated the effectiveness of the proposed

methods. In the future, we will consider to extend the proposed models to cross-database age estimation scenarios [Tian and Chen (2017)] with large-margin learning [Gu, Sheng, Tay et al. (2015); Gu, Sun and Sheng (2017)] and deep learning for further performance gain.

**Acknowledgment:** This work was partially supported by the National Natural Science Foundation of China (61702273 and 61472186), the Natural Science Foundation of Jiangsu Province (BK20170956), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (17KJB520022), the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and the Startup Foundation for Talents of Nanjing University of Information Science and Technology.

## 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*, vol. 30, no. 12, pp. 946-953.

**Bocklet, T.; Maier, A.; Bauer, J. G.; Burkhardt, F.; Nöth, E.** (2008): Age and gender recognition for telephone applications based on GMM supervectors and support vector machines. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1605-1608.

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

**Chen, K.; Gong, S.; Xiang, T.; Mary, Q.; Loy, C. C.** (2013): Cumulative attribute space for age and crowd density estimation. *International Conference on Computer Vision and Pattern Recognition*, pp. 2467-2474.

**Fjermestad, J.; Romano, N. C.** (2006): *Electronic Customer Relationship Management*. M. E. Sharpe, USA.

**Fu, Y.; Xu, Y.; Huang, T.** (2007): Estimating human age by manifold analysis of face pictures and regression on aging features. *IEEE International Conference on Multimedia and Expo*, pp. 1383-1386.

**Geng, X.; Zhou, Z.; Smith-Miles, K.** (2007): Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 92, no. 12, pp. 2234-2240.

**Geng, X.; Yin, C.; Zhou, Z.** (2013): Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401-2412.

**Gu, B.; Sheng, V. S.; Tay, K. Y.; Romano, W.; Li, S.** (2015): Incremental support vector learning for ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 7, pp. 1403-1416.

**Gu, B.; Sun, X.; Sheng, V. S.** (2017): Structural minimax probability machine. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1646-1656.

**Guo, G.; Fu, Y.; Dyer, C. R.; Huang, T. S.** (2008): Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, vol. 17, no.7, pp. 1178-1188.

**Guo, G.; Fu, Y.; Huang, T. S.; Dyer, C. R.** (2008): Locally adjusted robust regression for human age estimation. *IEEE Workshop on Applications of Computer Vision*, pp. 1-6.

**Jain, A. K.; Dass, S. C.; Nandakumar, K.** (2004): Soft biometric traits for personal recognition systems. *Biometric Authentication*, vol. 3072, pp. 731-738.

**Kohli, S.; Prakash, S.; Gupta, P.** (2013): Hierarchical age estimation with dissimilarity-based classification. *Neurocomputing*, vol. 120, pp. 164-176.

**Lanitis, A.; Taylor, C. J.; Cootes, T. F.** (2002): Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442-455.

**Lanitis, A.; Draganova, C.; Christodoulou, C.** (2004): Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 621-628.

**Li, C.; Liu, Q.; Liu, J.; Lu, H.** (2012): Learning distance metric regression for facial age estimation. *International Conference on Pattern Recognition*, pp. 2327-2330.

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

**Li, C.; Liu, Q.; Liu, J.; Lu, H.** (2015): Ordinal distance metric learning for image ranking. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 7, pp. 1551-1559.

**Liu, H.; Lu, J.; Feng, J.; Zhou, J.** (2018): Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 292-305.

**Luu, K.; Ricanek, K.; Bui, T. D.; Suen, C. Y.** (2009): Age estimation using active appearance models and support vector machine regression. *International Conference on Biometrics: Theory, Applications, and Systems*, pp. 1-5.

**Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G.** (2016): Ordinal regression with multiple output cnn for age estimation. *International Conference on Computer Vision and Pattern Recognition*, pp. 4920-4928.

**Sai, P.; Wang, J.; Teoh, E.** (2015): Facial age range estimation with extreme learning machines. *Neurocomputing*, vol. 149, pp. 364-372.

**Tian, Q.; Xue, H.; Qiao, L.** (2016): Human age estimation by considering both the ordinality and similarity of ages. *Neural Processing Letters*, vol. 43, no. 2, pp. 505-521.

**Tian, Q.; Chen, S.** (2017): Cross-heterogeneous-database age estimation through correlation representation learning. *Neurocomputing*, vol. 238, pp. 286-295.

**Tian, Q.; Chen, S.** (2018): Joint gender classification and age estimation by nearly orthogonalizing their semantic spaces. *Image and Vision Computing*, vol. 69, pp. 9-21.

**Ueki, K.; Hayashida, T.; Kobayashi, T.** (2006): Subspace-based age-group classify-cation using facial images under various lighting conditions. *International Conference on Face and Gesture Recognition*, pp. 43-48.

**Xing, J.; Li, K.; Hu, W.; Yuan, C.; Ling, H.** (2017): Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, vol. 66, pp. 106-116.

**Yan, S.; Wang, H.; Tang, X.; Huang, T. S.** (2007): Learning auto-structured regressor from uncertain nonnegative labels. *International Conference on Computer Vision*, pp. 1-8.

**Yan, S.; Wang, H.; Huang, T. S.; Yang, Q.; Tang, X.** (2007): Ranking with uncertain labels. *IEEE International Conference on Multimedia and Expo*, pp. 96-99.

**Yang, X.; Gao, B; Xing, C.; Huo, Z.; Wei, X. et al.** (2016): Deep label distribution learning for apparent age estimation. *IEEE International Conference on Computer Vision Workshop*, pp. 344-350.

**Yu, Z.; Yeung, D.** (2012): A convex formulation for learning task relationships in multi-task learning. *IEEE International Conference on Uncertainty in Artificial Intelligence*, pp. 733-742.