Human Age Estimation by Considering both the Ordinality and Similarity of Ages

Qing Tian[a], Hui Xue[b], Lishan Qiao[c,*]

[a] *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*
[b] *School of Computer Science and Engineering, Southeast University, Nanjing 210096, China*
[c] *School of Mathematical Sciences, Liaocheng University, Liaocheng 252059, China*
tianqing@nuaa.edu.cn; hxue@seu.edu.cn; qiaolishan@nuaa.edu.cn

Abstract

The age sequence of human beings exhibits two striking characteristics: *ordinal* in age values and *similar* in facial appearance of neighboring ages. Although it has been demonstrated that such ordinality especially the neighboring similarity has positive influence on age estimation, existing approaches have yet not simultaneously taken the two types of information into the estimation. In this paper to conduct age estimation with considering both the ordinality and the neighbor similarity which we call *soft-age-contribution* (SAC), we take the widely used discriminant method LDA and the least squares regression (LS) as the research baseline, respectively. Firstly, we construct *inequality-based* large margin ordinal constraints and *equality-based* ordinal regression constraints and, respectively, incorporate them into LDA and LS to develop their respective ordinal counterpart, coined as OrLDA and OrLS. Next, in order to utilize the SAC information, we formulate two types of membership function to depict such neighboring similarity and embed them into OrLDA and OrLS, yielding soft and ordinal variants of LDA and LS, called SAC-OrLDA and SAC-OrLS in which both the ordinality and the neighboring similarity of ages are considered. Finally, through experiments on benchmark aging datasets, we demonstrate the effectiveness of our strategies in utilizing the two types of information to improving age estimation. In addition, we also quantitatively explore the similarity of neighboring ages, finding that generally about neighboring four years are similar in facial appearance to each other.

Keywords: age estimation; neighboring similarity; ordinal relationship; discriminant analysis; least squares regression

1. Introduction

Human face carries a large amount of biological information, such as age, gender, race, facial expression, and the condition of physical health, etc. As an important facial trait, the age information has been used in a wide range of real world applications, such as web security control (Guo et al., 2008; Lanitis et al., 2004), ancillary identity authentication (Jain et al., 2004), and advertisement recommendation (ROMANO JR and FJERMESTAD, 2006), etc.

In order to conduct age estimation based on human face, a variety of approaches have been proposed to date. When we consider each age as a separate class, the age estimation can be performed 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 success to some extent. 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 aging 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, 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.

Although the methods afore-mentioned can yield age estimation with accuracy to different extents, they ignore the fact that there exists natural ordinality among the ages. As illustrated in Figure 1, the facial appearance of 12 years old is older than that of 7, but younger than that of 16. To capture such a characteristic into estimation, Li et al. (Li et al., 2012a) proposed a ordinal metric learning method for age estimation, in which the ordinal relationships of ages are incorporated into the metric process and on FG-NET they obtained compet-

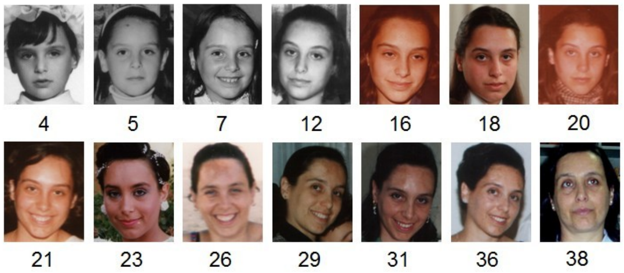


Figure 1: Facial age range of one instance from FG-NET dataset, where the label below every image represents its facial age.

itive accuracy. Moreover, they (Li et al., 2012b) took the ordinality-preserving capability of the features as a criterion to reduce dimensionality and conducted age regression with much better results. More recently, to represent the ordinal characteristic of an aging range, Chen et al. (Chen et al., 2013) proposed the strategy of "cumulative attribute" coding to depict the ordinal characteristic of ages, and by this way on the benchmark aging datasets they obtained more competitive age estimation results.

Actually, by comparing the facial appearance of facial images at close ages in Figure 1, we can find that the closer the ages, the more similar the facial appearance is. For example, the facial appearances of 5 and 12 years old both are quite similar to that of 7 years, compared with the similarity between 16 years and 7 years. In other words, neighboring ages have high similarity in facial appearance, which we call *soft-age-contribution* (SAC). According to such a characteristic, Liu et al. (Liu et al., 2014) collected large quantities of web human face images, and manually labeled them with age ranges rather than exact ages to expand the aging dataset they had, and based on the expanded aging dataset they obtained improved estimation results using the ε-SVR. To take the similarity information of neighboring ages into the learning, Geng et al. (Geng et al., 2013) presented the label distribution concept, which essentially is similar to the SAC in concept. Using the theorem of Bayesian probability and the maximum entropy, they transformed the problem of automatic age predicting into age label distribution learning, named IIS-LLD. However, their model holds under the assumption that the aging distribution is consistent with the assumed entropy condition, which may not hold in reality. To get free from such an assumption, they (Geng et al., 2013) further employed a three-layered neural network (NN) to generate an optimization objective similar to that in IIS-LLD, called conditional probability neural network (CPNN). Although the conditional aging contribution assumption in IIS-LLD is no longer required for CPNN, the solution trained by NN is usually just locally optimal and the learning process of NN to converge is seriously time-consuming. What is worse is that their way of using the label distribution information (or say the SAC) is not flexible to be extended in other models outside of their framework. Practically speaking, it will be preferable if the SAC information could be depicted by an analytical formulation. Along this line, Gao et al. (Gao and Ai, 2009) manually designed three kinds of age-group membership functions (as shown in Figure 2) to take the SAC information into the discriminant analysis (i.e., LDA) for age-group estimation. Although on some datasets such as the FG-NET, they obtained competitive results than the traditional LDA, several concerns are posed: On the one hand, the age-group membership functions manually designed may be not consistent with biological distributions of the ages, thus misleading the estimation; On the other hand, besides the SAC characteristic afore-mentioned, there also exist ordinal relationships among the ages.

To take both the ordinality and the SAC information into age estimation, in this work without loss of generality we adopt the classical LDA (Li et al., 2006) and the least squares regression (LS) (Geladi and Kowalski, 1986) as the base method, respectively. Firstly we model their specific ordinal constraints to

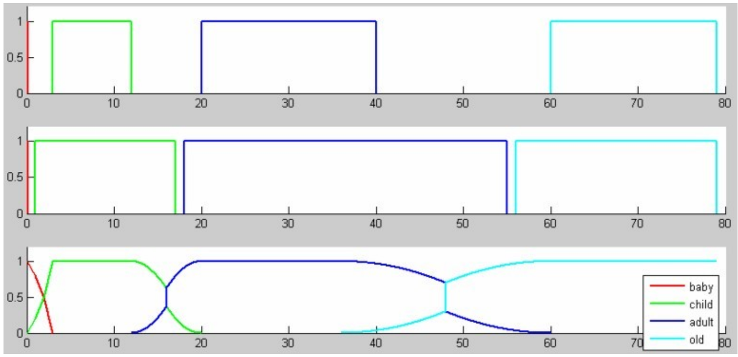


Figure 2: Three kinds of age-group membership functions originally shown in (Gao and Ai, 2009).

capture the ordinality of the ages to develop their respective ordinal variants, called OrLDA and OrLS, and then formulate two kinds of membership functions to depict the SAC relationships and thus develop the final models, SAC-OrLDA and SAC-OrLS, by embedding both the ordinal constraints and the SAC terms into the base. Finally, through experiments we demonstrate the effectiveness of the proposed strategies.

The main contributions of the work are as follows:

- According to the model structure of LDA and LS, construct *inequality-based* large margin ordinal constraints and *equality-based* ordinal regression constraints and, respectively incorporate them into the methods to develop their ordinal counterparts, coined as OrLDA and OrLS.
- Formulate two kinds of membership function to depict the similarity of neighboring ages (or say SAC), and embed them into the OrLDA and OrL-S, yielding the models considering both the ordinality and the similarity of ages, called SAC-OrLDA and SAC-OrLS.
- Conduct experiments to validate the effectiveness of the proposed strategies in improving age estimation, and more importantly, quantitatively explore the similarity degree of neighboring ages.

The rest of the paper is organized as follows. In Section 2, we briefly review the ordinary LDA and LS, respectively. Then, according to the structure of the model, we design specific ordinal constraints and thus propose ordinal variants of the LDA and LS, i.e., OrLDA and OrLS, by imposing the ordinal constraints in Section 3. In Section 4, we further incorporate the SAC information to OrLDA and OrLS to develop their soft counterparts SAC-OrLDA and SAC-OrLS. Experimental results and analyses are reported in Section 5. Finally, we conclude the paper in Section 6.

2. Review of ordinary LDA and LS

Before introducing our research, let us briefly review the ordinary linear discriminant analysis (LDA) and the least squares regression (LS), respectively.

2.1. LDA

Given a set of N samples $\{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \Re^d$ being the instance belonging to K classes and $y_i \in \Re$ the label, then the mean vector m_k of the k-th class can be obtained by $m_k = \frac{1}{N_k} \sum x_k^i$, where N_k is the number of samples in the k-th class, and the total mean of samples can be calculated by $m = \frac{1}{N} \sum x^i$. The LDA aims to find a projection direction along which the total intra-class scatters are compressed while the inter-class scatters are as separated as possible, i.e., the LDA is to find a projection direction $w \in \Re^{d \times 1}$ along which the following objective can be minimized (Li et al., 2006)

$$\textit{minimize} \quad J(w) = \frac{w^T S_w w}{w^T S_b w}. \tag{1}$$

where $S_w = \sum_{k=1}^{K} \sum_{x \in X_k} (x - m_k)(x - m_k)^T$ stands for the intra-class scatter matrix, and $S_b = \sum_{k=1}^{K} N_k \cdot (m_k - m)(m_k - m)^T$ the inter-class scatter matrix. By solving objective (1) through *generalized Rayleigh Quotient* (Bathe and Wilson, 1976), we can obtain the optimal projection direction w.

2.2. LS

The least squares regression (LS) (Geladi and Kowalski, 1986) aims to seek for a projection vector $w \in \Re^d$ by minimizing the following mean-squared errors

$$\textit{minimize} \quad J(w) = \sum_{i=1}^{N} \parallel w^T x_i - y_i \parallel^2, \tag{2}$$

or an equivalent form

$$\textit{minimize} \quad J(w) = w^T X X^T w - 2 w^T X Y^T + Y Y^T, \tag{3}$$

where $X \in \Re^{d \times N}$ and $Y \in \Re^{1 \times N}$ denote the instances and the ground-truth regression values, respectively. Actually, the last term in (3) could be omitted due to its irrelevance to the w.

3. Ordinal variants of LDA and LS: OrLDA and OrLS

From equations (1) and (2), we can find that neither the LDA nor the LS can well suit for such ordinal estimation problems as human age estimation, because of their ignorance of the ordinal characteristic of the problems. To this end, according to their model structure, in this section we first design their specific ordinal constraints and then incorporate them into the original LDA and LS to develop their ordinal counterparts, coined as OrLDA and OrLS, to cater for ordinal estimations.

3.1. Ordinal LDA: OrLDA

From (1), it can be found that for a multi-class discriminant projection, LDA just concerns to find a projection direction along which the ratio of the projected total intra-class scatters over the inter-class scatters is minimized, but has not considered the ordinal relationships between such data as human ages when applied in such ordinal problems. As a result, it may finally meet a projection direction satisfying the maximal discriminant principle but violating the ordinal relationships of the data, as illustrated in Figure 3.

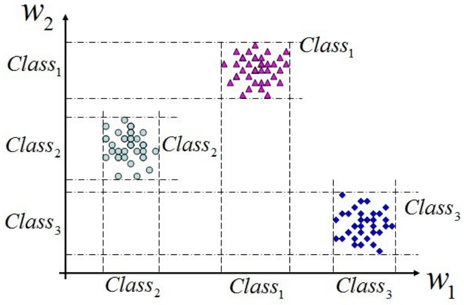


Figure 3: An illustration of two projection directions for ordinal classes. w_1 is the direction obtained by LDA, along which the ordinal relationships among the classes are disordered, while along w_2 they can be preserved well.

According to the analyses above, we propose to impose ordinal constraints on the class-mean of the data to guide for seeking a projection direction along which to preserve the ordinal orders of the classes while as much abiding the discriminant principle as possible, and the developed ordinal variant of LDA, coined as OrLDA, can be formulated as

$$\begin{aligned} &minimize \quad J(w,\xi) = \frac{1}{2} w^T (S_w - \frac{\lambda_1}{2} S_b) w + \lambda_2 \sum_{k=1}^{K-1} \xi_k \\ &s.t. \quad w^T(m_{k+1} - m_k) \geq 1 - \xi_k,\ k = 1,2,...,K-1, \\ &\qquad \xi_k \geq 0,\ k = 1,2,...,K-1, \end{aligned} \tag{4}$$

where S_w and S_b are the intra-class and the inter-class scatter matrices, respectively as defined in (1), $m_k, k = 1,2,...,K$, denotes the mean vector of the k-th class, and λ_1 and λ_2 are two nonnegative trade-off parameters controlling the balance between the structure risk and the empirical risk. In consideration of the degree of margin between two neighboring classes might be unequal, we introduce slack variables to regularize the solutions.[1]

[1] It is worthwhile to note that the proposed OrLDA in (4) is essentially different from the KDLOR (Sun et al., 2010), in both the objective and the form of constraints. In KDLOR, it is assumed that the degree of margins between two neighboring classes is equal, which may be inconsistent with actual classes distributions.

The objective of OrLDA in (4) is a convex quadratic programming (QP) problem, provided that the sub-term of $S_w - \frac{\lambda_1}{2} S_b$ is positive semi-definite (PS-D). Actually, both the S_w and S_b are PSD by definition, thus by tuning the trade-off parameter λ_1, this assumption could hold. So, it could be optimized using off-the-shelf QP algorithms such as SMO (Boser et al., 1992). Definitely, we can also solve it efficiently through deriving its dual problem via the Lagrangian theorem as

$$\begin{aligned} & minimize \quad J(\alpha) = \frac{1}{2}\alpha^T M^T (S_w - \lambda_1 S_b)^{-1} M\alpha - \sum_{k=1}^{K-1} \alpha_k \\ & s.t. \quad 0 \leq \alpha_k \leq \lambda_2, \quad k = 1, 2, ..., K-1, \end{aligned} \tag{5}$$

where $M = [m_2 - m_1, ..., m_K - m_{K-1}]$. Obviously, the problem (5) with dual variable vector α of length $K-1$ is relatively easier to optimize with less computational complexity than the primal problem (4) with variables (w, ξ). After obtaining the α by solving (5), we can obtain the primal variable w in (4) by

$$w = (S_w - \frac{\lambda_1}{2} S_b)^{-1} M\alpha.$$

3.2. Ordinal LS: OrLS

Analyzing the ordinary LS formulated in (2), it can be found that the LS aims to seek a projection direction along which the squared absolute residual between the regressed and the ground-truth values is minimized. In this sense, when LS is employed to perform ordinal estimation, the ordinal relationships among the ordinal data can not be preserved, due to that we can make regression to approach the true values along two directions, i.e., along the ordinal direction or the reverse direction, as illustrated in Figure 4. In Figure 4, the actual order relationship is $y_i = k$ smaller than $y_j = k + 1$ by 1. However, the regressed

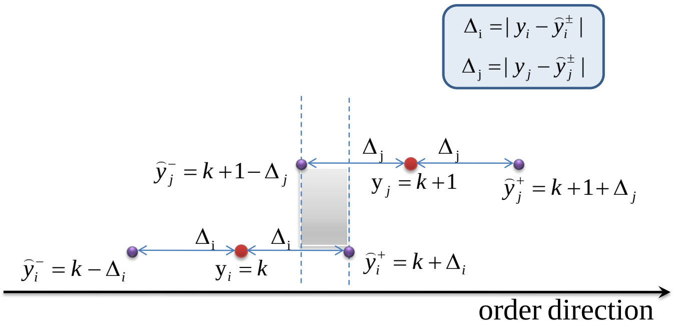

$\Delta_i = | y_i - \hat{y}_i^{\pm} |$
$\Delta_j = | y_j - \hat{y}_j^{\pm} |$
$\hat{y}_j^- = k+1-\Delta_j$
Δ_j
Δ_j
$y_j = k+1$
$\hat{y}_j^+ = k+1+\Delta_j$
Δ_i
Δ_i
$\hat{y}_i^- = k-\Delta_i$
$y_i = k$
$\hat{y}_i^+ = k+\Delta_i$
order direction

Figure 4: An illustration of LS for ordinal data regression, where "$\triangle_i$" ("$\triangle_j$") denotes the regression residual between the ground-truth y_i (y_j) and the regressed value $\widehat{y_i^+}$ ($\widehat{y_j^+}$) along the order direction or $\widehat{y_i^-}$ ($\widehat{y_j^-}$) along the reverse order direction.

order by LS might be $\widehat{y}_j^-$ smaller than $\widehat{y}_i^+$. Consequently, the original order relationships of the data are disordered. To handle this problem, according to the structure of LS we introduce the equality ordinal constraints into (3) (*it is equivalent to* (2)) and thus develop its ordinal counterpart, coined as OrLS, written as

$$\begin{aligned} &\textit{minimize} \quad J(w,\xi) = w^T XX^T w - 2w^T XY^T + YY^T + \lambda \sum_{k=2}^{K} \sum_{i=1}^{N_k} (\xi_k^i)^2 \\ &\textit{s.t.} \quad w^T(x_{k+1}^i - m_k) = 1 - \xi_{k+1}^i, \quad k = 1, 2, ..., K-1;\ i = 1, 2, ..., N_{k+1}, \end{aligned} \tag{6}$$

where x_{k+1}^i denotes the i-th training sample from class $k+1$, m_k the mean vector of the k-th class, and the meanings of X, Y and M are the same as that in (3) or (5). Note that slack variables ξ here are not required to be nonnegative and they are squared minimized in the objective function. Fortunately, the regression projection vector w of (6) can be obtained analytically as

$$\begin{aligned} w = \big(XX^T + \lambda \sum_{k=1}^{K-1} \sum_{i=1}^{N_k} (x_{k+1}^i - m_k)(x_{k+1}^i - m_k)^T\big)^{-1} \\ \cdot \big(XY^T + \lambda \sum_{k=1}^{K-1} \sum_{i=1}^{N_k} (x_{k+1}^i - m_k)\big) \end{aligned} \tag{7}$$

It is worthwhile to point out that although ordinal constraints are incorporated into the ordinary LS to develop its ordinal counterpart, a mathematically analytical solution still can be obtained. This attractive property is crucial in practical applications, especially those sensitive to computational complexity.

Remarks

Comparing the ordinal constraints of OrLDA in (4) and OrLS in (6), we can find that the main difference lies in the form of constraints: *inequality* constraints for the OrLDA and *equality* constraints for the OrLS. Taking inequality ordinal constraints into LDA and equality ordinal constraints into LS for OrLDA and OrLS is due to the following considerations, respectively. As for LDA, according to the claim that large between-class margins can boost the generalization ability of classifiers (Agresti, 2010; Zhang and Zhou, 2013), so to make the compressed sets of data of different classes to be separated apart as far as possible while arranged in an ordinal order, it is desirable to impose ordinal inequality constraints on the classes to perform ordinal discriminant learning. As for LS, it is a regression approach and aims to make the regression residual minimal between the ground truth and regressed values. So the large margin ordinal constraints for LDA are no longer suitable for LS, instead, we are expected to impose ordinal constraints on the regression values without increasing the regression residuals, as illustrated in Figure 4. In consideration of the analyses above, we propose to impose equality ordinal constraints between data points of one class and their neighboring class center (or say class mean) with fixed margin 1, and in order to regularize the solution space, we also introduce slack variables in the constraints and consequently obtain the OrLS formulated in (6).

4. Soft variants of OrLDA and OrLS: SAC-OrLDA and SAC-OrLS

Besides the ordinal relationships among the ages, next we are in position to introduce the SAC information into age estimation by incorporating the similarity of neighboring ages (i.e., the SAC information) into OrLDA and OrLS, respectively. Before that, let us firstly define two types of SAC membership function that will be embedded into our ordinal LDA/LS to achieve the goal of performing age estimation with further considering the similarity of the ages, besides the ordinal relationships.

4.1. SAC membership function

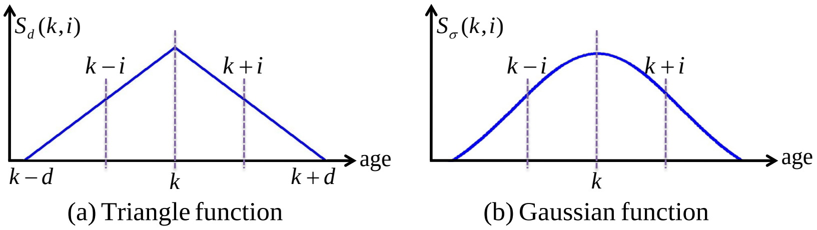


Figure 5: Two types of SAC membership function.

As illustrated in Figure 1, neighboring ages exhibit similarity in facial appearance, and the closer the ages, the more similar the facial appearance is. According to such an observation (i.e., the SAC characteristic of ages), we design two types of membership function to depict the SAC and call them Triangle membership function (TF, $S_d(k,i)$) and Gaussian membership function (GF, $S_\sigma(k,i)$), shown in Figure 5 (a) and (b), respectively, and formulate them as

$$S_d(k,i) = \frac{max\{(d-|i-k|)/d, 0\}}{\sum_{i=1}^{K} max\{(d-|i-k|)/d, 0\}} \tag{8}$$

and

$$S_\sigma(k,i) = \frac{e^{-\frac{(k-i)^2}{\sigma}}}{\sum_{i=1}^{K} e^{-\frac{(k-i)^2}{\sigma}}}, \tag{9}$$

where parameters d and σ play a role in controlling the age span of SAC, in TF and GF, respectively. The $S_d(k,i)$ or $S_\sigma(k,i)$ depicts the normalized similarity between age k and i.

It is worth pointing out that although so-called label distributions introduced in (Geng et al., 2013) are in concept similar to our TF and GF, their formulations are not presented at all and are learned by assumption-based entropy learning or highly time-consuming neural networks training, more importantly, their way of considering the label distributions might not be flexible to be extended in other methods outside their framework. As a contrast, the membership functions explicitly formulated in (8) and (9) can be easily embedded into existing methods including the OrLDA and OrLS.

4.2. Soft OrLDA: SAC-OrLDA

To take the SAC information into discriminant age estimation besides the ordinal constraints, we come to achieve this goal by embedding the SAC membership function, either TF or GF, into (4) and thus generate its soft counterpart, coined as SAC-OrLDA, as

$$\begin{aligned} &minimize \quad J(w,\xi) = \frac{1}{2} w^T (S_w^{SAC} - \frac{\lambda_1}{2} S_b^{SAC}) w + \lambda_2 \sum_{k=1}^{K-1} \xi_k \\ &s.t. \quad w^T (m_{k+1}^{SAC} - m_k^{SAC}) \geq 1 - \xi_k, \; k = 1,2,...,K-1, \\ &\qquad \xi_k, \; k = 1,2,...,K-1, \end{aligned} \tag{10}$$

where

$$S_w^{SAC} = \sum_{k=1}^{K} \sum_{j=1}^{K} \sum_{x \in X_j} S(k,j) \cdot (x - m_j)(x - m_j)^T,$$

$$S_b^{SAC} = \sum_{k=1}^{K} \sum_{j=1}^{K} S(k,j) \cdot N_j \cdot (m_j - m)(m_j - m)^T,$$

and

$$m_k^{SAC} = \sum_{j=1}^{K} S(k,j) \cdot m_j.$$

The SAC membership function $S(k,j)$ in (10) can be either the TF or GF, or even other valid membership functions. The problem of (10) in form is the same as (4), thus we can directly borrow the implementation for OrLDA to solve it.

4.3. Soft OrLS: SAC-OrLS

Similarly, we also introduce the SAC information into OrLS by embedding the membership function into (6) and consequently obtain the remodeled soft counterpart, coined as SAC-OrLS, formulated as

$$\begin{aligned} &minimize \; J(w,\xi) = w^T X S_1^{SAC} X^T w - 2Y S_2^{SAC} X^T w + Y S_3^{SAC} Y^T + \lambda \sum_{k=2}^{K} \sum_{i=1}^{N_k} (\xi_k^i)^2 \\ &s.t. \quad w^T (x_{k+1}^i - m_k^{SAC}) = 1 - \xi_k^i, \; k = 1,2,...,K-1, \; i = 1,2,...,N_{k+1}, \end{aligned} \tag{11}$$

where

$$S_1^{SAC} = \begin{pmatrix} S_{11}^1 & & & \\ & S_{22}^1 & & \\ & & \ddots & \\ & & & S_{KK}^1 \end{pmatrix}_{N\times N},$$

$$S_{ii}^1 = \begin{pmatrix} \sum_{k=1}^{K} S(i,k) & & \\ & \ddots & \\ & & \sum_{k=1}^{K} S(i,k) \end{pmatrix}_{N_i\times N_i};$$

$$S_2^{SAC} = \begin{pmatrix} S_{11}^2 & S_{12}^2 & \cdots & S_{1K}^2 \\ S_{21}^2 & S_{22}^2 & \cdots & S_{2K}^2 \\ \vdots & \vdots & \ddots & \vdots \\ S_{K1}^2 & S_{K2}^2 & \cdots & S_{KK}^2 \end{pmatrix}_{N\times N},$$

$$S_{ij}^2 = \begin{pmatrix} \frac{S(i,j)}{N_i} & \cdots & \frac{S(i,j)}{N_i} \\ \vdots & \ddots & \vdots \\ \frac{S(i,j)}{N_i} & \cdots & \frac{S(i,j)}{N_i} \end{pmatrix}_{N_i\times N_i};$$

$$S_3^{SAC} = \begin{pmatrix} S_{11}^3 & & & \\ & S_{22}^3 & & \\ & & \ddots & \\ & & & S_{KK}^3 \end{pmatrix}_{N\times N},$$

$$S_{ii}^3 = \begin{pmatrix} \sum_{k=1}^{K} S(k,i) & & \\ & \ddots & \\ & & \sum_{k=1}^{K} S(k,i) \end{pmatrix}_{N_i\times N_i}.$$

Note that the third term $YS_3^{SAC}Y^T$ in the objective of (11) actually can be omitted, due to that it does not affect the objective optimization. Similar to the OrLS, there also exists analytical solution for the w in (11) as

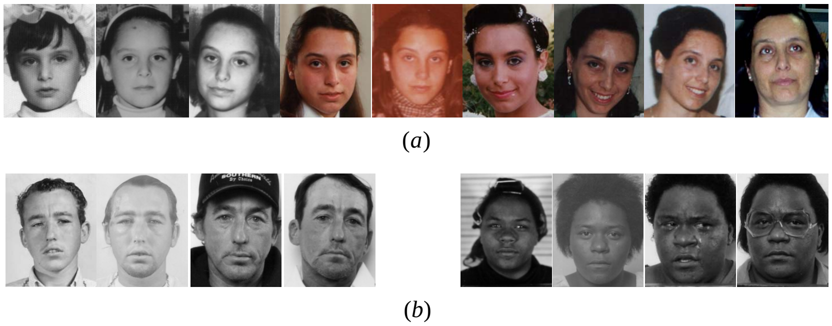
$$\begin{aligned} w = &\big(XS_1^{SAC}X^T + \lambda \sum_{k=1}^{K-1}\sum_{i=1}^{N_k}(x_{k+1}^i - m_k^{SAC})(x_{k+1}^i - m_k^{SAC})^T\big)^{-1} \\ &\cdot \big(X(S_2^{SAC})^T Y^T + \lambda \sum_{k=1}^{K-1}\sum_{i=1}^{N_k}(x_{k+1}^i - m_k^{SAC})\big). \end{aligned} \tag{12}$$

5. Experiments

In this section we conduct experiments to make evaluations on the proposed methods. To be specific, we first make an introduction on the human aging datasets that will be used in the experiments and the experimental settings, then on the datasets we perform experiments to detailedly evaluate the proposed methods and explore the similarity relationship of the ages.

5.1. Datasets and Settings

Datasets: In this work we conduct human age estimations on two commonly used benchmark datasets, i.e., the FG-NET and Morph. The FG-NET consists of 1,002 facial images from 82 individuals of the European, and the age ranges from 0 to 69 years. As for the Morph, it contains 1,690 images of about 631 persons mainly from the Africa and Europe, with the age ranging from 15 to 68 years. Image examples from the two datasets are exhibited in Figure 6, respectively.



(*a*)

(*b*)

Figure 6: Image examples from the FG-NET (the top row) and the Morph (the bottom row).

Settings: Analyzing the age distributions of the two datasets, i.e., FG-NET and Morph, it can be easily found that they are imbalanced, even some ages are associated with no samples, as shown in Figures 7 and 8, respectively. As a result, to make fair comparisons with such methods involved with sample class-mean such as LDA and KDLOR (Sun et al., 2010), we take an age range of 0 to 55 years and 15 to 56 years, generating 56 and 42 age classes from the FG-NET and Morph, respectively for experiments. Then, we adopt the AAM method (Cootes et al., 1998) to extract 95 components from the selected aging datasets as the feature representation. And, we also make experimental comparisons for the proposed methods with several representative ordinal estimation approaches including the SVOR (Chu and Keerthi, 2005), SVR (Mu et al., 2009), and KDLOR (Sun et al., 2010). All the hyper-parameters involved are set by performing cross-validation. Finally, we report the averaged results over 20 runs by random training data split and measure them by the *mean absolute error* (MAE), $MAE := \frac{1}{N}\sum_{i=1}^{N} |l_i - \widehat{l_i}|$ with l_i and $\widehat{l_i}$ denoting the ground-true and predicted values, respectively.

5.2. Evaluation on the Effectiveness of the Proposed Strategies

We first explore the effectiveness of introducing the ordinal information of age classes, incorporating the SAC to age estimation with increasing quantity of training samples. Specifically, we randomly select certain number of samples from each age class for training and take the rest for test, and report the averaged

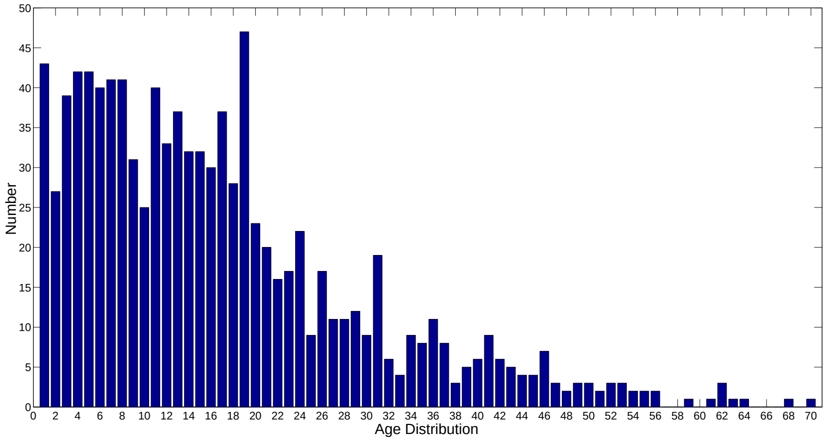


Figure 7: Sample distribution of different ages on FG-NET.

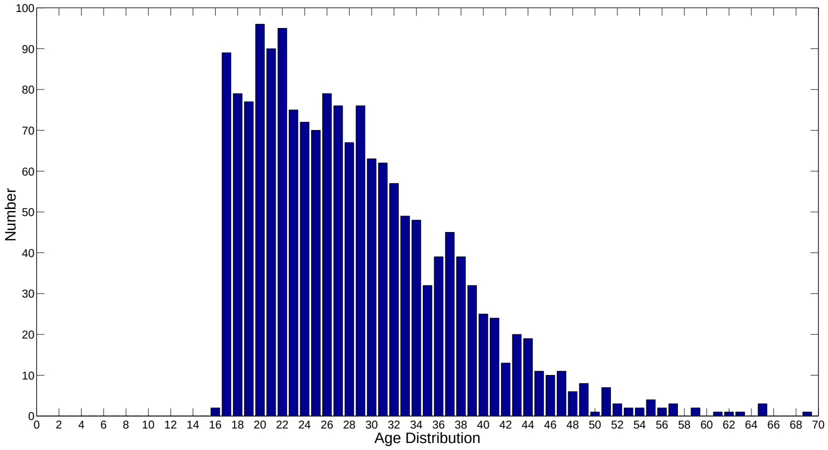


Figure 8: Sample distribution of different ages on Morph.

results over 20 runs in Tables 1-4. From them, we can obtain the following findings.

• Comparing the MAEs (corresponding to the results of OrLDA/OrLS in Tables 1-4) of the methods imposed with ordinal constraints against those without such constraints (corresponding to the results of LDA/LS in the four tables), it can be found that the former ones are much lower than the latter ones, especially when the number of training data is relatively not large. For example, on FG-NET, OrLDA performs age estimation with MAE reduced by about 19% (from 10.41 to 8.46 in case of #samples = 3) from that of LDA, and similarly, OrLS performs age estimation with MAE 7.80 reduced by about 9% from 8.57 of the LS. It indicates that to both the discriminant analysis and the regression

Table 1: Comparison of age estimation (mean ± std, MAE in years) between LDA, its variants and other methods on FG-NET.

# samples	LDA	OrLDA	SAC-OrLDA (TF)	SAC-OrLDA (GF)	SVOR	SVR	KDLOR
3	10.41±1.06	8.46±0.76	**7.06±0.31**	7.21±0.36	10.64±0.59	8.37±0.74	9.24±0.58
6	8.88±0.98	7.80±0.40	6.60±0.32	**6.54±0.30**	7.32±0.43	8.24±0.40	8.78±0.62
9	6.45±0.49	5.98±0.29	**5.84±0.31**	**5.80±0.27**	6.14±0.29	7.01±0.41	7.06±0.80
12	5.35±0.50	5.09±0.30	**4.97±0.23**	**5.00±0.26**	5.32±0.21	6.31±0.25	6.22±0.42
15	4.93±0.42	4.68±0.21	**4.51±0.19**	**4.52±0.18**	4.64±0.20	5.93±0.29	5.65±0.38
18	4.51±0.41	4.30±0.21	**4.10±0.21**	**4.11±0.25**	4.25±0.20	5.61±0.38	5.24±0.40
21	4.41±0.63	4.16±0.29	4.08±0.20	4.04±0.19	**3.95±0.19**	5.33±0.32	4.88±0.58

Table 2: Comparison of age estimation (mean ± std, MAE in years) between LDA, its variants and other methods on Morph.

# samples	LDA	OrLDA	SAC-OrLDA (TF)	SAC-OrLDA (GF)	SVOR	SVR	KDLOR
3	14.43±0.85	9.15±0.52	**8.65±0.55**	**8.67±0.51**	8.97±0.40	9.22±0.52	9.28±0.64
6	13.94±0.71	8.65±0.38	**8.38±0.29**	**8.39±0.28**	8.51±0.30	8.99±0.61	8.48±0.19
9	12.44±0.71	8.36±0.28	8.30±0.26	8.28±0.26	7.80±0.26	8.38±0.28	**7.48±0.00**
12	9.82±0.65	7.62±0.23	7.42±0.27	7.46±0.39	6.98±0.20	7.62±0.23	**6.46±0.00**
15	8.52±0.62	7.22±0.22	7.09±0.26	7.12±0.19	6.49±0.13	7.22±0.24	**6.42±0.00**
18	7.74±0.56	7.03±0.18	6.85±0.18	6.87±0.18	**6.16±0.14**	7.03±0.18	6.40±0.00
21	7.20±0.48	6.83±0.18	6.63±0.19	6.65±0.18	**5.84±0.14**	6.83±0.18	5.90±0.00

Table 3: Comparison of age estimation (mean ± std, MAE in years) between LS, its variants and other methods on FG-NET.

# samples	LS	OrLS	SAC-OrLS (TF)	SAC-OrLS (GF)	SVOR	SVR	KDLOR
3	8.57±0.36	7.80±0.40	**7.56±0.37**	**7.58±0.31**	10.64±0.59	8.37±0.74	9.24±0.58
6	7.33±0.45	6.70±0.40	**6.45±0.28**	**6.43±0.28**	7.32±0.43	8.24±0.40	8.78±0.62
9	6.64±0.31	6.18±0.38	**5.83±0.23**	**5.83±0.19**	6.14±0.29	7.01±0.41	7.06±0.80
12	6.22±0.31	5.89±0.28	**5.20±0.21**	**5.21±0.23**	5.32±0.21	6.31±0.25	6.22±0.42
15	5.99±0.29	5.60±0.21	5.14±0.22	5.18±0.23	**4.64±0.20**	5.93±0.29	5.65±0.38
18	5.58±0.36	5.29±0.27	4.80±0.25	4.91±0.27	**4.25±0.20**	5.61±0.38	5.24±0.40
21	5.39±0.26	5.03±0.31	4.67±0.30	4.70±0.30	**3.95±0.19**	5.33±0.32	4.88±0.58

Table 4: Comparison of age estimation (mean ± std, MAE in years) between LS, its variants and other methods on Morph.

# samples	LS	OrLS	SAC-OrLS (TF)	SAC-OrLS (GF)	SVOR	SVR	KDLOR
3	8.91±0.52	8.80±0.50	**8.72±0.40**	**8.75±0.39**	8.97±0.40	9.22±0.52	9.28±0.64
6	7.86±0.28	7.69±0.26	**7.58±0.32**	**7.59±0.26**	8.51±0.30	8.99±0.61	8.48±0.19
9	7.48±0.25	7.23±0.24	**7.10±0.21**	**7.08±0.25**	7.80±0.26	8.38±0.28	7.48±0.00
12	7.38±0.31	7.19±0.23	6.91±0.28	6.90±0.23	6.98±0.20	7.62±0.23	**6.46±0.00**
15	7.34±0.28	7.17±0.20	6.76±0.31	6.80±0.21	6.49±0.13	7.22±0.24	**6.42±0.00**
18	7.08±0.20	6.87±0.19	6.56±0.14	6.57±0.14	**6.16±0.14**	7.03±0.18	6.40±0.00
21	6.94±0.19	6.80±0.18	6.42±0.11	6.46±0.16	**5.84±0.14**	6.83±0.18	5.90±0.00

learning, introducing proper ordinal constraints can significantly improve their ordinal estimation performance.

- The MAEs of SAC-OrLDA (with either TF or GF) are correspondingly lower than those of OrLDA, as shown in Tables 1 and 2), and it is similar to that SAC-OrLS over OrLS (as shown in Tables 3 and 4). For example, on the FG-NET, SAC-OrLDA and SAC-OrLS improve the age estimation by reducing MAE by about 16% from 8.46 (in the case of #samples = 3), by about 11% from 5.89 (in the case of #samples = 12), respectively. And on the Morph dataset, incorporating the SAC information into OrLDA/OrLS also significantly improves their ability in age estimation. As shown in Tables 2 and 4, both the SAC-OrLDA and SAC-OrLS reduce the MAEs of age estimation by up to about 6% from those of OrLDA or OrLS. More importantly, when the number of training data is relatively insufficient (for example, not larger than 15 on FG-NET dataset, or not larger than 12 on Morph), both the SAC-OrLDA and the SAC-OrLS yield the best age estimation accuracy, among all the compared methods. It shows that in human age estimation, especially in case of sparse samples, taking the SAC relationships (i.e., the similarity of neighboring ages) into account for learning can effectively improve the estimation since that the characteristic of neighbor-similarity of ages allows to represent an age with samples from its neighboring ages in addition to its own limited samples and consequently alleviates the performance limit caused by inadequate samples. This characteristic is practically useful in real human age estimation, since collecting samples is usually costly or even unpractical, especially for some age-specific people.

5.3. Similarity Relationship of Neighboring Ages

From the experimental results in Section 5.2, we have learnt that incorporating the SAC information, i.e., the similarity of neighboring ages, into age estimation can significantly improve the its accuracy. Next, we come to quantitatively explore the similarity relationship of neighboring ages, as well as its influence on age estimation. To this end, we conduct experiments and show the results obtained in Figure 9. From the four sub-figures (A)-(D) in Figure 9, the following interesting common rule can be found.

In the beginning, the accuracy of age estimation is getting better and better with increasing age span of the SAC membership function, up to the span of 4 years. Then with the age span of the SAC membership function set at 4, the best estimation accuracies are obtained (with the lowest MAEs), and the MAEs have been reduced from about 5.35 to 4.95 years, 7.50 to 7.40 years, 5.75 to 5.20 years, and 7.25 to 6.90 years by about 7%, 2%, 10%, and 5% in Figure 9 (A) to (D), respectively. When the age span of the SAC membership function is larger than 4 years, the MAEs of age estimation are turning to become larger seriously, even worse than those without incorporating the SAC information (i.e., equivalent to such a case where the age span of the SAC membership function is set 1). It might suggest that the facial age appearances of about neighboring 4 years are biologically similar to each other, and in turn such an aging characteristic can be used to help estimate human ages. *To our knowledge, this may be the first quantitative exploration on the issue of similarity of neighboring ages.*

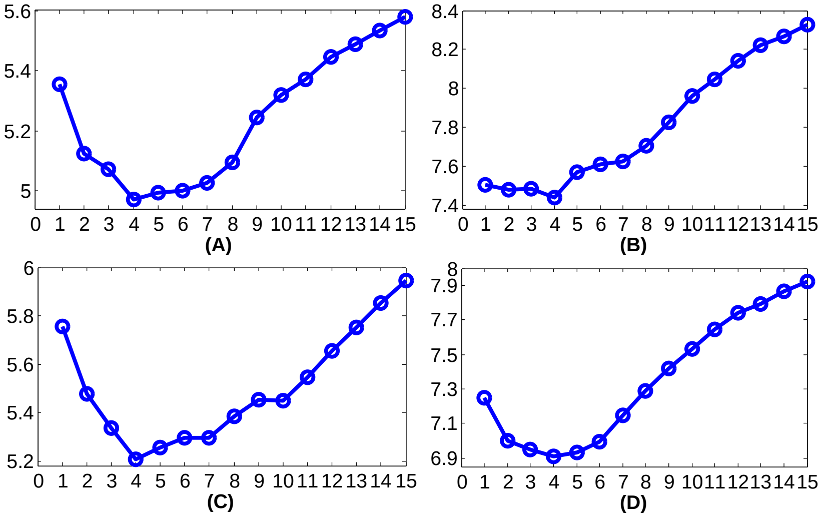


Figure 9: The similarity of neighboring ages and its influence on age estimation. The x-axis indicates the age span of SAC membership function, and the y-axis denotes the age estimation accuracy in MAE. Without loss of generality, here we demonstrate the results of the case #samples = 12 for SAC-OrLDA(TF) on FG-NET (corresponding to (A)) and on Morph (corresponding to (B)), and SAC-OrLS on FG-NET (corresponding to (C)) and on Morph (corresponding to (D)). The rule of other cases is quite similar to this.

6. Conclusion

In this work, in order to take both the ordinal information and the neighboring similarity of the ages into human age estimation, without loss of generality we took the discriminant learning (LDA) and the least squares regression (LS) as research baseline, respectively. Firstly, according to the model structure of LDA and LS we constructed their specific ordinal constraints (i.e., *inequality-based large margin ordinal constraints* for LDA and *equality-based ordinal regression constraints* for LS), incorporated them into the LDA and LS to achieve the goal of taking the ordinality of the ages into estimation, and thus developed their ordinal counterpart, OrLDA and OrLS with solutions, respectively. Then, to put the neighboring similarity (we call *soft-age-contribution*, or SAC for short) of ages into the estimation as well, we formulated two types of membership function to quantitatively depict the SAC information and embedded them into the OrLDA and OrLS to develop their respective soft and ordinal counterpart, coined as SAC-OrLDA and SAC-OrLS, in which both the ordinality and the neighbor similarity of ages is incorporated. Finally, through experiments on two benchmark aging datasets, we demonstrated the effectiveness of the proposed strategies in improving age estimation, especially quantitatively explored the similarity of neighboring ages with the finding that generally about four neighboring ages are in appearance similar to each other.

In order to eliminate the influence of pose variations to age estimation, we will consider to conduct pose-invariant age estimations by taking the pose variations into account for learning as in (Ding et al., 2015), (Ding et al., 2014), and (Ding and Tao, 2015).

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grants 61472186, 61375057 and 61300154, Natural Science Foundation of Jiangsu Province of China under Grant $BK20131298$, Funding of Jiangsu Innovation Program for Graduate Education under Grant $CXLX13\_159$, and the Fundamental Research Funds for the Central Universities. In addition, we would like to express our gratitude to Prof. Songcan Chen (now with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China) for his valuable advice on building the related models in this work.

References

Agresti, A., 2010. Analysis of ordinal categorical data. volume 656. Wiley. com.

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.

Bathe, K.J., Wilson, E.L., 1976. Numerical methods in finite element analysis .

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, ACM. pp. 144–152.

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.

Chu, W., Keerthi, S.S., 2005. New approaches to support vector ordinal regression, in: Proceedings of the 22nd international conference on Machine learning, ACM. pp. 145–152.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 1998. Active appearance models, in: Computer VisionECCV98. Springer, pp. 484–498.

Ding, C., Choi, J., Tao, D., Davis, L.S., 2014. Multi-directional multi-level dual-cross patterns for robust face recognition. arXiv preprint arXiv:1401.5311 .

Ding, C., Tao, D., 2015. A comprehensive survey on pose-invariant face recognition. arXiv preprint arXiv:1502.04383 .

Ding, C., Xu, C., Tao, D., 2015. Multi-task pose-invariant face recognition. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society 24, 980–993.

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.

Gao, F., Ai, H., 2009. Face age classification on consumer images with gabor feature and fuzzy lda method, in: Advances in biometrics. Springer, pp. 132–141.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Analytica chimica acta 185, 1–17.

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

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

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

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, T., Zhu, S., Ogihara, M., 2006. Using discriminant analysis for multi-class classification: an experimental investigation. Knowledge and information systems 10, 453–472.

Liu, J., Ma, Y., Duan, L., Wang, F., Liu, Y., 2014. Hybrid constraint svr for facial age estimation. Signal Processing 94, 576–582.

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.

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.

ROMANO JR, 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.

Sun, B.Y., Li, J., Wu, D.D., Zhang, X.M., Li, W.B., 2010. Kernel discriminant learning for ordinal regression. Knowledge and Data Engineering, IEEE Transactions on 22, 906–910.

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 .

Zhang, T., Zhou, Z.H., 2013. Large margin distribution machine. arXiv preprint arXiv:1311.0989 .