Joint Gender Classification and Age Estimation by Nearly Orthogonalizing Their Semantic Spaces

Qing Tian, Songcan Chen

PII: S0262-8856(17)30169-5
DOI: doi: 10.1016/j.imavis.2017.10.003
Reference: IMAVIS 3655

To appear in: *Image and Vision Computing*

Received date: 21 October 2016
Revised date: 26 September 2017
Accepted date: 31 October 2017

Please cite this article as: Qing Tian, Songcan Chen, Joint Gender Classification and Age Estimation by Nearly Orthogonalizing Their Semantic Spaces, *Image and Vision Computing* (2017), doi: 10.1016/j.imavis.2017.10.003

Joint Gender Classification and Age Estimation by Nearly Orthogonalizing Their Semantic Spaces

Qing Tian[1,2,3], Songcan Chen[3*]

1 *School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, P.R.China*
2 *Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, P.R.China*
3 *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P.R.China*

Abstract

In human face-based biometrics, gender classification and age estimation are two important research topics. Although a variety of approaches have been proposed to handle them, just a few of them are solved jointly, even so, these joint methods do not specifically concern the semantic difference between human gender and age, which is intuitively helpful for joint learning, consequently leaving us a room of further improving their performance. To this end, in this work we firstly propose a general learning framework for jointly estimating human gender and age by attempting to formulate such semantic relationships as a form of near-orthogonality regularization and then to incorporate it into the objective of the joint learning framework. In order to evaluate the effectiveness of the proposed framework, we exemplify it by respectively taking the widely used binary-class SVM for gender classification, and two threshold-based ordinal regression methods (i.e., the discriminant learning for ordinal regression and support vector ordinal regression) for age estimation, and crucially coupling both through the proposed semantic formulation. Moreover, we construct its nonlinear counterpart by deriving a representer theorem for the joint learning strategy. Finally, extensive experiments on four aging datasets, i.e., FG-NET, Morph Album I, Album II and Images of Groups demonstrate the effectiveness and superiority of the proposed strategy.

Keywords: Gender classification; Age estimation; Nearly orthogonal semantic spaces; Support vector ordinal regression; Discriminant learning for ordinal regression

1. Introduction

Human face conveys rich biometric characteristics, such as gender, age, ethnicity and expression, in which the estimation of face-based gender and/or age has attracted extensive attentions due to its wide applications in recommendation systems [1], [2], security access control [3], [4], [5], biometrics [6], [7], entertainment [8], [9], [10], [11] and cosmetology [12], [13], etc. Consequently, in this work we concentrate on the research of estimating human gender and age simultaneously. Before introducing our approach, we first review research references on human gender and age estimations below.

1.1. Separate naive estimation of human facial gender or age

The concept of separate *naive* estimate of human facial gender or age means estimating the facial gender or age separately without considering the interrelationship between them. Along this line, large amount of works have been proposed to estimate them below.

Gender classification (GC) is typically addressed as a binary classification problem due to its binariness of male or female. Along this line, the Bayesian classifier [14], [15], [16], SVM classifier [17], [18], [19], [20], [21], [22], [23], [24], [25], Random Forest classifier [26], Boosting classifier [27], [28], [29], [30], and ELM classifier [31] have successfully been adopted to handle the problem of GC. More recently, due to the wide success of deep convolutional

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

neural networks (CNNs), the CNNs-based methods [32], [33], [34], [35], [36], [37] have been employed to learn powerful feature representations for GC and achieved state-of-the-art estimation performance.

Age estimation (AE) is more challenging, compared with the GC problem, due to its multi-class characteristic as well as the ordinality of ages. So far, researchers have paid extensive attention on it and proposed a variety of approaches. According to the modelling strategies of these methods, they can mainly been grouped into three categories: classification-based, regression-based and their hybrids. To be specific, when the AE problem is treated as ordinary classification problem, it can be addressed by the SVM [38], [39], [40], [41], OHRank [42], [43], [44], kNN [45] and CNNs [46], [47], [47], [48], [49], [50] classifiers. Although most of the above methods, especially the CNNs based deep models can classify the ages with good results, the AE problem should be viewed as regression more than classification because the facial appearance is aging continuously. To this end, the support Vector Regression (SVR) [51], [52], [53], [54], Ridge Regression (RR) [54], [55] as well as Sparse Regression (SR) [56], [57], [58] methods have been adopted to regress human ages. Apart from the commonly-used regressors above, [59] adopted the Lie algebra Gaussian method to extract age-related feature representations from the face images and took the refined regressing for AE. [60] proposed the AGES method and applied it to the reconstruction of missing age patterns. [61] extracted age-related features from the face image through spectral decomposition and performed age regression in nonlinear kernel space. [62] first learned an ordered metric space, and then used the kNN to realize the regression of facial age patterns in the learned metric space. [63] employed transfer learning to perform AE across different aging databases. To learn more powerful feature representations, [53], [64], [65], [66] adopted CNNs for AE. Besides the above separate classification or regression based methods, the hybrid strategy of combining classification and regression has also been adopted to AE and received more competitive results.

1.2. Separate estimation of human facial gender or age with caring their interrelationships

The aforementioned separate estimation methods of gender or age handle the GC or AE problem separately without definitely considering the mutual correlations between them. Actually, there exists mutual influence between their estimations [67], [38], i.e., the GC result is affected by the variation of facial aging, and vice versa. Therefore, in the process of GC or AE estimation we should take into account the influence of the other. Along this line, to perform GC eliminating the influence by facial aging, [68] first performed age groups classification (*elderly*, *middle-aged* and *young*) and then conducted AE for each of the age groups. [69] separated the faces into two age groups of *mature* and *immature*, and then performed AE for each of the groups, finding that the GC accuracy of the mature is higher than the immature. On the other hand, to reduce the influence of gender on AE, [70], [38], [40] first performed GC to get male and female aging groups and then in each group made AE.

1.3. Joint estimation of human facial gender and age

Compared with the naive separate methods, although the methods reviewed in Section 1.2 have achieved competitive estimation results, which perform GC or AE with discriminating gender-difference in aging or aging-difference between the male and female by first performing GC and then AE or first AE followed by GC, the interrelationships between gender and age are not used in them. To preserve the relationship between human gender and age, a more natural way is to estimate them together rather than separately. Along this line, [67] and [71] employed a multi-output regressor PLS to jointly regress the gender and age. However, PLS is a universal regressor and has not definitely considered the semantic relationship between the gender and age. To capture the commonality of male and female in their aging, [72] adopted the l_{21}-norm to perform group feature selection and then performed AE for male and female on the selected feature representations. Although this multi-task based method has achieved promising results, it is essentially a pseudo-joint method because it does not actually perform GC and AE simultaneously, leading to the ignorance of their interrelationships. More recently, motivated by the success of deep models, [73], [74], [75] constructed CNNs-based deep models for jointly estimating human gender and age. It is worth noting that these deep models, especially the method of [75], achieved state-of-the-art performance. Although the CNNs-based deep models mostly yield quite promising GC/AE results, their training is terribly time-consuming, and even worse, it is practically difficult to tune for the large number of model parameters and incorporate the interrelationships between the gender and age in the architecture of the complicated deep models.

1.4. Challenges and contributions of this work

From the reference review above, It can be found that there mainly exist several drawbacks among existing methods: 1) the ordinality characteristic of human aging sequence is not definitely preserved; 2) the underlying correlation between the gender and age is not strategically considered; 3) the semantic discrepancy between the gender and age is not desirably preserved. Therefore, it is desirable to jointly estimating human gender and age while overcoming these drawbacks of previous methods. However, *to achieve this goal is quite challenging*, because GC and AE are heterogeneous tasks and the *binariness* of gender and *ordinality* of age are heterogeneous characteristics. As a result, we have to develop novel modelling strategy for this purpose. To this end, in this paper we propose a general joint learning framework for human gender and age, in which gender estimation is took as a binary classification and age prediction as an ordinal regression problem. More crucially, as a key ingredient of reflecting the semantic discrepancy between human gender and age, the underlying relationship between their semantic spaces is formulated as a near-orthogonality regularizer and further incorporated in the objective function. In order to evaluate the effectiveness of the proposed framework, we exemplify it by respectively taking the widely used binary-class SVM for gender classification, and the discriminant learning for ordinal regression/support vector ordinal regression for age estimation, and particularly coupling them via the semantic regularizer. Then we kernelize the joint learning framework by deriving a representer theorem. Finally, through experiments on four real-world aging datasets, we demonstrate the effectiveness and superiority of the proposed joint learning strategy. Specifically, *our contributions are five-fold as follows*:

1. Formulate the semantic relationship between human gender and age as a near-orthogonality regularizer.
2. Propose a joint estimation framework for human gender and age based on the proposed semantic regularizer.
3. Exemplify the proposed framework with two specific examples.
4. Kernelize the proposed framework by deriving corresponding representing theorem.
5. Experimentally demonstrate the effectiveness and superiority of the proposed methods.

The rest of this paper is organized as follows. In Section 2, we specifically review representative works on joint estimation of human gender and age. Then, in Section 3 and Section 4 we introduce our method. In Section 5, we conduct experiments to evaluate the effectiveness and superiority of the proposed methods. Finally, Section 6 concludes this paper.

2. Related work

Before introducing our methods, in this section we review three types of representative works on joint estimation of human gender and age.

2.1. Joint age estimation for male and female with multi-task (group-lasso) learning

In order to conduct joint age estimation for males and females with feature selection, [72] adopted the multi-task framework regularized with the l_{21}-norm (*a.w.a*, Group-Lasso [76]) by taking male-oriented and female-oriented age estimations as two tasks. More specifically, for a given training set $\{x_i^t, y_i^t\} \in \mathbb{R}^D \times \mathbb{R}, i = 1, ..., N_t, t = 1, 2$, where x_i^t and y_i^t denote the i-th D-dimensional instance and its label from the t-th task, respectively, they performed joint age estimation for males and females through using the off-the-shelf multi-task feature selection learning which is formulated as

$$\min_{W=[w^1,w^2]} \frac{1}{\sum_{t=1}^{2} N_t} \sum_{t=1}^{2} \sum_{i=1}^{N_t} \|y_i^t - (w^t)^T x_i^t\|_2^2 + \lambda \sum_{d=1}^{D} \|w_d\|_2^2, \tag{1}$$

where $W = [w^1, w^2] \in \mathbb{R}^{D\times 2}$ consists of the w^1 and w^2 which respectively represent the projection weights for the two tasks (corresponding to the male-oriented and female-oriented age estimations), and the second term in (1) is the well-known group-lasso regularization and is usually used for joint feature selection among the tasks, λ is the nonnegative trade-off parameter.

Taking the selected features learned through optimizing (1) as new feature representations, the authors of [72] then adopted RR to regress ages for male and female, respectively. From the modeling strategy of [72], it can be found that this method is essentially a pseudo-joint method, because the gender information is just used to cater for the feature selection of AE.

2.2. Joint GC and AE using partial least squares (PLS)

PLS is a typical multi-output regressor that depicts the mapping from the input features to the responses. More concretely, for given input matrix X and output response matrix Y (whose each response vector is concatenated from the multi-output variables, *e.g.*, the gender and age), both of which are centered with zero-means, the PLS aims to compute two weight vectors, w and c, to maximize the following covariance

$$cov(t, u) = max_{|w|=|c|=1} cov(X^T w, Y^T c), \tag{2}$$

where $cov(t, u)$ represents the covariance between the score vectors $t = X^T w$ and $u = Y^T c$. Regressions then can be performed for both X and Y based on the score vectors t and u computed by optimizing the (2), such that

$$X = pt^T + X_1, \quad Y = qu^T + Y_1, \tag{3}$$

where $p = \frac{Xt}{t^T t}$ and $q = \frac{Yu}{u^T u}$ are called loading vectors, and X_1 and Y_1 are the regression residuals of X and Y, respectively. Then, based on (3), we can compute a sequence of score and loading vectors to make the regression residuals small enough as

$$X = p_1 t_1^T + \cdots + p_k t_k^T + X_k, \quad Y = q_1 u_1^T + \cdots + q_k u_k^T + Y_k. \tag{4}$$

According to the regression relationships between the loading vectors $T = \{t_1, \cdots, t_k\}$ and $U = \{u_1, \cdots, u_k\}$ ([77]), we have

$$Y = B^T X + R_Y, \tag{5}$$

where $B = XU(T^T X^T XU)^{-1} T^T Y^T$, and R_Y stands for the regression residual. Actually, if the number k of iterations is properly assigned, the residual R_Y can be omitted for practical applicability. For the sake of following involved notations of PLS, we summarize them in Table 3 (see Appendix 2).

When PLS is employed to perform joint estimation for human gender and age as in [67] and [71], its two-dimensional output indicates the regressed gender and age results, respectively. However, by this way the heterogenicity between the discrete binaryness of gender and the continuous ordinality of age is seriously destroyed. Even worse, it does not definitely take into account the ordinal characteristic of human age, nor the semantic discrepancy between gender and age.

2.3. Joint GC and AE using hybrid deep networks

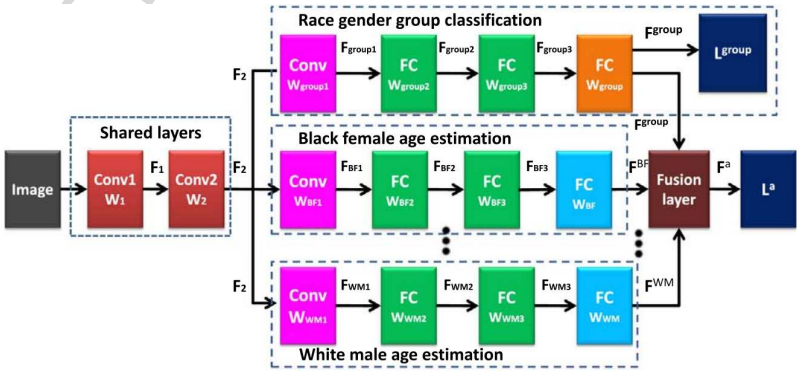

Figure 1: The architecture of Net_{hybrid}^{VGG} (more details can be referred to [75]).

Motivated by the success of deep CNNs in image analysis, [75] [1] designed a hybrid deep convolutional networks (as shown in Figure 1, we call it Net_{hybrid}^{VGG}) to simultaneously estimate the gender and age. In Net_{hybrid}^{VGG}, a hybrid multi-task learning architecture is designed to jointly estimate human gender and age by assigning the Softmax probabilities

[1] Although apart from [75], other deep convolutional models have been proposed in [73] and [74] to handle GC and AE, literature [75] is newest, representative and its performance is state-of-the-art among them. Therefore, here we just review the work of [75] and interested readers can refer to these references.

of gender classification results as fusion weights of male estimation result and female age estimation result. The model parameters of Net_{hybrid}^{VGG} can be obtained by optimizing the following objective function:

$$F^a = \sum_{k \in BF,BM,WF,WM} F_k^{group} F^k, \tag{6}$$

where the k-th element F_k^{group} of F^{group} stands for corresponding race gender group, i.e., Black Female (BF), Black Male (BM), White Female (WF) and White Male (WM). The Net_{hybrid}^{VGG} model mainly comprises of two functional types of sub-networks: Race Gender group classification and race-gender-oriented age estimation sub-models (i.e., BF, BM, WF and WM age estimations). To be specific, the Race Gender group classification is in charge to classify input faces into BF, BM, WF or WM, the BF, BM, WF and WM age estimation models respectively perform AE for corresponding race and gender group. Then, the softmax output of Race Gender group classification is assigned as weights to fuse the AE results of the race-gender-oriented age estimation. Finally, the final predicted age of input face can be obtained.

From the architecture of Net_{hybrid}^{VGG} shown in Figure 1, we can find that it is through sharing some representation layers to achieve the goal of utilizing the mutual correlations between human gender and age. Although such a deep modelling strategy can achieve quite promising AE and GC, its tuning of large number of parameters is quite time-consuming, and moreover, there is no reliable knowledge that *how many*, *how deep* and *which* convolutional layers should be shared between the gender and age. Therefore, the rationality of the architecture of Net_{hybrid}^{VGG} might be an open problem.

3. Proposed Methodology

In this section, we first propose a general learning framework to estimate human gender and age simultaneously, in which the binaryness of human gender and the ordinality of human age are both explicitly considered, and in particular, the semantic relationship between human gender and human age is also explored and exploited to improve their estimations. Then, we exemplify the proposed framework for the sake of the following empirical evaluation.

3.1. A Novel Joint Framework for Human Gender Classification and Age Estimation: Learning in Nearly Orthogonal Semantic Spaces

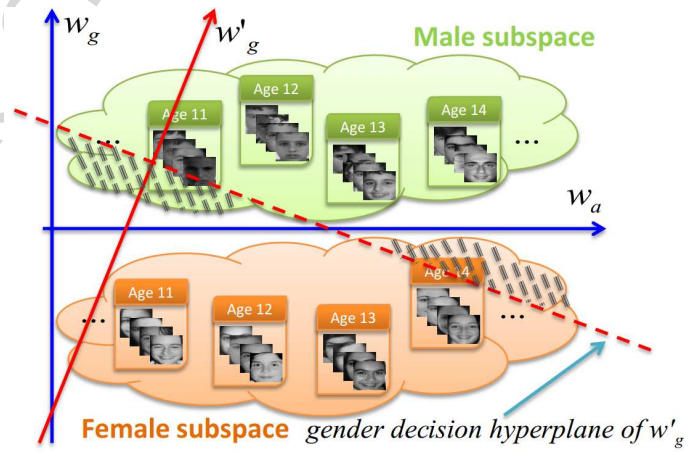

Figure 2: Illustration of the proposed methodology. w_g (w_g') and w_a indicate human gender discriminant direction and aging direction, respectively.

Now consider the situation that there is a set of human face samples associated with gender labels and age annotations, what we need do is to train a joint estimator in terms of the gender and age using the training set. For the sake of presenting our strategy, we illustrate it in Figure 2, in which w_a denote the aging direction, and w_g or w_g' indicates the candidate gender discriminant direction. As claimed in, the face samples distribute monotonously and

orderly according to their age along w_a. So, if $w_g^{'}$ is chosen as the gender discriminant direction, the male and female gender subspaces will cross the gender decision hyperplane of $w_g^{'}$ (see the area shaded by oblique lines in Figure 2), implying that a severe gender misclassification. By contrast, if we choose w_g, which is nearly orthogonal to w_a [2], as the gender discriminant direction, the male and female gender subspaces will more clearly distribute on two sides of its gender decision hyperplane. Consequently, we should choose w_a and w_g as human aging direction and gender discriminant direction, respectively. For the sake of constructing a joint learning framework for human gender and age, we mathematically formulate learning in such *nearly orthogonal semantic spaces* (NOSSpaces) as a regularization term

$$\begin{aligned}\mathcal{R}_{NOSSpaces} &:= (w_a^T w_g)^2 \\ &= w_g^T w_a w_a^T w_g \\ &= w_a^T w_g w_g^T w_a\end{aligned} \tag{7}$$

Clearly, according to the above analysis the $\mathcal{R}_{NOSSpaces}$ should be minimized as small as possible to achieve a desirable joint estimation on human gender and age, as illustrated in Figure 2.

With the proposed joint learning regularizer $\mathcal{R}_{NOSSpaces}$, we are in position to propose a joint learning framework to simultaneously perform GC and AE as follows:

$$\begin{aligned}&min_{\{w_g, w_a\}} \\ &\mathcal{L}_g(w_g; X, Y_g) + \frac{\lambda_1}{2}\|w_g\|^2 + \mathcal{L}_a(w_a; X, Y_a) + \frac{\lambda_2}{2}\|w_a\|^2 + \frac{\lambda_3}{2}\mathcal{R}_{NOSSpaces}\end{aligned} \tag{8}$$

where $\mathcal{L}_g(w_g; X, Y_g)$ refers to the binary classification loss function regard to GC, *e.g.*, the widely-used *hinge loss function* [78], $\mathcal{L}_a(w_a; X, Y_a)$ refers to the estimation loss regard to AE, such as *squares loss* [79], λ_1, λ_2 and λ_3 are predefined nonnegative trade-off parameters to balance the loss functions and model regularization terms. Through such a modelling strategy, both GC and AE submodels can be trained jointly, and more importantly, the semantic relationships between them can be explicitly incorporated.

It is worth noting that

1. Although the regularizer in (7) in form is seemingly trivial, to the best of our knowledge, it is the first work to explore the semantic relationship between human gender and age.
2. The NOSSpaces is similar *in spirit* to but different *in essence* from generic multi-task learning, because that in multi-task learning, different tasks are exclusively trained on their respective training data without data sharing, while in our framework the gender classification and age estimation are performed based on the same training data. More importantly, the formulation of NOSSpaces is relatively more concise than most of the multi-task learning methods.
3. In theory, while there may be other strategies that can capture the semantic relationships between human gender and age, our joint learning methodology formulated in (7) is concise and easy-to-implement, and more importantly, we later will experimentally demonstrate its effectiveness and superiority over related methods.

3.2. Joint GC and AE in the NOSSpaces

In order to evaluate the proposed framework and without loss of generality, we exemplify (8) by taking the widely used binary-class SVM for gender classification, while the discriminant learning for ordinal regression (KDLOR) and support vector ordinal regression (SVOR) for age estimation, respectively.

3.2.1. SVM for GC and Discriminant Learning for Ordinal Regression for AE in the NOSSpaces

With the proposed joint learning framework (8), we construct the first exemplified joint estimation model by respectively substituting the $\mathcal{L}_g(w_g; X_g, Y_g)$ and the $\mathcal{L}_a(w_a; X_a, Y_a)$ with the widely used binary SVM and the discrim-

[2]Note that the w_g and w_a are not required to be strictly orthogonal to each other, which will be verified in Section 5.4.

inant learning for ordinal regression (KDLOR)[3] [80] as

$$
\begin{aligned}
&min_{\{w_g,b_g,w_a,\rho\}} \\
&\quad \frac{1}{2}\|w_g\|^2 + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\} \\
&\quad + w_a^T S_w w_a - \lambda_2 \rho + \lambda_3 (w_g^T w_a)^2 \\
&s.t. \\
&\quad w_a^T (m_{k+1} - m_k) \geq \rho,\ k = 1, 2, ..., K-1,
\end{aligned}
\tag{9}
$$

where N_g denotes the number of samples used for training in gender classification, w_g and b_g, respectively, denote the weight vector and intercept for gender classification, $\{x_i^g, y_i^g\}_{i=1}^{N_g}$ denote the i-th instance and corresponding gender label of N_g samples, w_a denotes the weight vector for age estimation, $S_w = \frac{1}{N_a}\sum_{k=1}^{K}\sum_{x\in X_k}(x - m_k)(x - m_k)^T$ stands for the within-class scatter with N_a being the total number of training samples of K classes, X_k the training samples set of the k-th class, and m_k the mean vector of the k-th class, ρ is interval margin between the classes, and λ_1, λ_2 and λ_3 are non-negative hyper-parameters.

Due to the bi-convexity of (9) with respect to $\{w_g, b_g\}$ and $\{w_a\}$, i.e., it is convex with respect to $\{w_g, b_g\}$ with fixed $\{w_a\}$, and vice versa. Therefore, we can take an alternative strategy to solve $\{w_g, b_g, w_a\}$. More specifically, for fixed w_a, then (9) becomes

$$
\begin{aligned}
&min_{\{w_g,b_g\}} \\
&\quad \frac{1}{2}\|w_g\|^2 + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\} + \lambda_3 (w_g^T w_a)^2
\end{aligned}
\tag{10}
$$

which is equivalent to

$$
\begin{aligned}
&min_{\{w_g,b_g\}} \\
&\quad \frac{1}{2} w_g^T (\mathcal{I} + 2\lambda_3 w_a w_a^T) w_g + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\},
\end{aligned}
\tag{11}
$$

where $\mathcal{I}$ is the identity matrix of proper size. The sub-problem (11) is convex and can be similarly solved by the same way for SVM [78].

When w_g and b_g are solved by (10) or (11), then they are constant and we come to optimize (9) with respect to $\{w_a\}$ as

$$
\begin{aligned}
&min_{\{w_a\}} \\
&\quad w_a^T S_w w_a - \lambda_2 \rho + \lambda_3 (w_g^T w_a)^2 \\
&s.t. \\
&\quad w_a^T (m_{k+1} - m_k) \geq \rho,\ k = 1, 2, ..., K-1,
\end{aligned}
\tag{12}
$$

which in form is equivalent to

$$
\begin{aligned}
&min_{\{w_a\}} \\
&\quad w_a^T (S_w + \lambda_3 w_g w_g^T) w_a - \lambda_2 \rho \\
&s.t. \\
&\quad w_a^T (m_{k+1} - m_k) \geq \rho,\ k = 1, 2, ..., K-1,
\end{aligned}
\tag{13}
$$

(12) in form is quite similar to the primal problem of KDLOR, and thus can be solved similarly as in [80].

By alternating between (11) and (13) until convergence, we can obtain the final solution for w_g, b_g, w_a, and b_a. And we summarize the complete procedure for joint SVM and KDLOR in Table 1.

[3]Note that in order to follow the abbreviation, we still call its linear counterpart KDLOR, unless specified with linear or nonlinear description.

Table 1: Algorithm of joint learning of SVM and KDLOR in the NOSSpaces.

Input:	Training instances X, and labels Y_{gender} and Y_{age}; Parameters λ_1, λ_2, and λ_3.
Output:	w_g, b_g, w_a.
1. Initialize w_a;	
2. **for** $t = 1, 2, ..., T_{max}$ **do**	
3. Compute w_g and b_g based on (11);	
4. Compute w_a and b_a based on (13);	
5. **end for**	
6. Return w_g, b_g, and w_a.	

Using w_g, b_g, and w_a solved by the Algorithm in Table 1, we can fulfil the goal of making joint estimation for human gender and age in the NOSSpaces.

3.2.2. SVM for GC and Support Vector Ordinal Regression for AE in the NOSSpaces

In order to evaluate the general feasibility of the proposed joint learning framework. Besides the KDLOR, we also adopt the support vector ordinal regression (SVOR) [81] method, a well-known ordinal regression method, for age estimation. And we can construct the corresponding joint model by substituting the formulations of SVM and SVOR into (8) as

$$
\begin{aligned}
&min_{\{w_g, b_g, w_a, b_a := \{b_j\}_{j=1}^K, \xi^{(*)}\}} \\
&\quad \frac{1}{2}\|w_g\|^2 + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\} \\
&\quad + \frac{1}{2}\|w_a\|^2 + \lambda_2 \sum_{j=1}^{K} \sum_{i=1}^{N_j} (\xi_i^j + \xi_i^{j*}) + \lambda_3 (w_g^T w_a)^2 \\
&s.t. \\
&\quad w_a^T x_i^j - b_j \le -1 + \xi_i^j, \quad \xi_i^j \ge 0, \\
&\quad w_a^T x_i^{j*} - b_{j-1} \ge 1 - \xi_i^{j*}, \quad \xi_i^{j*} \ge 0, \\
&\quad b_{j-1} \le b_j,
\end{aligned}
\tag{14}
$$

where w_g and b_g, respectively, denote the weight vector and intercept for gender classification, $\{x_i^g, y_i^g\}_{i=1}^{Ng}$ denote the i-th instance and corresponding gender label of N_g samples, w_a and b_a, respectively, denote the weight vector and intercept for age estimation, x_i^j stands for the i-th instance from the j-th age of totally K ages, $\xi^{(*)}$ represent the slack variables, $b_a := \{b_j\}_{j=1}^K$ are the thresholds of SVOR to be optimized, and λ_1, λ_2 and λ_3 are also non-negative trade-off parameters.

Clearly, the formulation (14) is also bi-convex with respect to $\{w_a, b_a, \xi^{(*)}\}$ and $\{w_g, b_g\}$. Therefore, we also take the alternative optimization to solve it. More precisely, for fixed $\{w_a, b_a, \xi^{(*)}\}$, (14) then becomes

$$
\begin{aligned}
&min_{\{w_g, b_g\}} \\
&\quad \frac{1}{2}\|w_g\|^2 + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\} + \lambda_3 (w_g^T w_a)^2
\end{aligned}
\tag{15}
$$

which is equivalent to

$$
\begin{aligned}
&min_{\{w_g, b_g\}} \\
&\quad \frac{1}{2} w_g^T (\mathcal{I} + 2\lambda_3 w_a w_a^T) w_g + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(w_g^T x_i^g + b_g)\},
\end{aligned}
\tag{16}
$$

where $\mathcal{I}$ is the identity matrix of proper size. The sub-problem (16) is convex and can be similarly solved by the same way for SVM [78]. Then, when $\{w_g, b_g\}$ are obtained and fixed, the problem (14) becomes

$$
\begin{aligned}
&min_{\{w_a, b_a, \xi^{(*)}\}} \\
&\frac{1}{2}\|w_a\|^2 + \lambda_2 \sum_{j=1}^{K} \sum_{i=1}^{N_j} (\xi_i^j + \xi_i^{j*}) + \lambda_3 (w_g^T w_a)^2 \\
&s.t. \\
&\quad w_a^T x_i^j - b_j \le -1 + \xi_i^j, \quad \xi_i^j \ge 0, \\
&\quad w_a^T x_i^{j*} - b_{j-1} \ge 1 - \xi_i^{j*}, \quad \xi_i^{j*} \ge 0, \\
&\quad b_{j-1} \le b_j,
\end{aligned}
\tag{17}
$$

or equivalently,

$$
\begin{aligned}
&min_{\{w_a, b_a, \xi^{(*)}\}} \\
&\frac{1}{2} w_a^T (\mathcal{I} + 2\lambda_3 w_g w_g^T) w_a + \lambda_2 \sum_{j=1}^{K} \sum_{i=1}^{N_j} (\xi_i^j + \xi_i^{j*}) \\
&s.t. \\
&\quad w_a^T x_i^j - b_j \le -1 + \xi_i^j, \quad \xi_i^j \ge 0, \\
&\quad w_a^T x_i^{j*} - b_{j-1} \ge 1 - \xi_i^{j*}, \quad \xi_i^{j*} \ge 0, \\
&\quad b_{j-1} \le b_j,
\end{aligned}
\tag{18}
$$

which is the same in form as the problem (5) of [81], and can be similarly solved as in the literature.

To obtain the final $\{w_g, b_g, w_a, b_a, \xi^{(*)}\}$, we repeat the alternative optimization process between (16) and (18) until convergence, and summarize the algorithm in Table 2.

Table 2: Algorithm of joint learning of SVM and SVOR in the NOSSpaces.

Input:	Training instances X, and labels Y_{gender} and Y_{age}; Parameters λ_1, λ_2, and λ_3.
Output:	w_g, b_g, w_a, b_a.
1. Initialize w_a, b_a, $\xi^{(*)}$;	
2. **for** $t = 1, 2, ..., T_{max}$ **do**	
3.	Compute w_g and b_g based on (16);
4.	Compute w_a and b_a based on (18);
5. **end for**	
6. Return w_g, b_g, w_a, and b_a.	

When w_g, b_g, w_a, and b_a are obtained by the Algorithm listed in Table 2, we can make predictions for human gender and age.

4. Joint Gender Classification and Age Estimation in the Nonlinear NOSSpaces

4.1. Nonlinear NOSSpaces

In real applications, usually the associated data are not linearly separable. In order to handle such applications, we need to extend the joint learning framework in (8) to higher-dimensional nonlinear spaces. Before that, we firstly give the specific representer theorem for the w_g and w_a involved in the joint learning framework as below.

Lemma 4.1. *The w_g and w_a in* (8) *can be respectively expressed as a linear combination of training samples as $w_g = \sum_i \alpha_i \phi(x^i)$ and $w_a = \sum_i \beta_i \phi(x^i)$, with α and β being the combination coefficients and $\phi(\cdot)$ the feature mapping function defined on training samples $\{x_i\}_{i=1}^N$. (the proof can be found in Appendix 1.)*

According to Lemma 4.1, the w_g and w_a can be respectively expressed as a combination of the training samples: $w_g = \sum_i \alpha_i \phi(x^i)$ and $w_g = \sum_i \beta_i \phi(x^i)$, and then NOSSpaces in (7) can be mapped into the nonlinear feature space with kernel trick and expressed as

$$\begin{aligned} \mathcal{R}_{Nonlinear-NOSSpaces} &:= (\beta^T \mathbf{K}\alpha)^2 \\ &= \alpha^T \mathbf{K}\beta\beta^T \mathbf{K}\alpha \\ &= \beta^T \mathbf{K}\alpha\alpha^T \mathbf{K}\beta, \end{aligned} \tag{19}$$

where $\mathbf{K}$ stands for the kernel matrix with $\mathbf{K}_{(i,j)} = \phi(x^i)^T \phi(x^j)$.

4.2. Joint GC and AE in the nonlinear NOSSpaces

With Lemma 4.1, the joint learning models in (9) can be reformulated in the nonlinear NOSSpaces as

$$\begin{aligned} &min_{\{\alpha, b_g, \beta, \rho\}} \\ &\frac{1}{2}\alpha^T \mathbf{K}\alpha + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(\alpha^T \mathbf{K}_{(:,i)} + b_g)\} \\ &+ \frac{1}{N_a}\beta^T \mathbf{K}\mathbf{K}\beta - \lambda_2\rho + \lambda_3 \alpha^T \mathbf{K}\beta\beta^T \mathbf{K}\alpha \\ &s.t. \\ &\beta^T\left(\frac{1}{N_{k+1}}\mathbf{K}_{(:,X_{k+1})}\mathbf{1}_{k+1} - \frac{1}{N_k}\mathbf{K}_{(:,X_k)}\mathbf{1}_k\right) \geq \rho, \ k = 1, 2, ..., K-1, \end{aligned} \tag{20}$$

with the assumption that the training samples have been centralized within each class beforehand, and $\mathbf{K}_{(:,i)}$ represents the i-th column of the total kernel matrix $\mathbf{K}$, and $\mathbf{K}_{(:,X_k)}$ stands for the sub-block kernel matrix between the k-th class samples and the entire training set, i.e., it is a sub-matrix of the $\mathbf{K}$.

Similarly, the joint SVM and SVOR in (14) can be transformed into the form as

$$\begin{aligned} &min_{\alpha, b_g, \beta, b_a := \{b_j\}_{j=1}^K, \xi^{(*)}} \\ &\frac{1}{2}\alpha^T \mathbf{K}\alpha + \lambda_1 \sum_{i=1}^{N_g} max\{0, 1 - y_i^g(\alpha^T \mathbf{K}_{(:,i)} + b_g)\} \\ &+ \frac{1}{2}\beta^T \mathbf{K}\beta + \lambda_2 \sum_{j=1}^{K}\sum_{i=1}^{N_j}(\xi_i^j + \xi_i^{j*}) + \lambda_3 \beta^T \mathbf{K}\alpha\alpha^T \mathbf{K}\beta \\ &s.t. \\ &\beta^T \mathbf{K}_{(:,X_i^j)} - b_j \leq -1 + \xi_i^j, \quad \xi_i^j \geq 0, \\ &\beta^T \mathbf{K}_{(:,X_i^{j*})} - b_{j-1} \geq 1 - \xi_i^{j*}, \quad \xi_i^{j*} \geq 0, \\ &b_{j-1} \leq b_j, \end{aligned} \tag{21}$$

where $\mathbf{K}_{(:,X_i^{j(*)})}$ denotes the i-th column of the $\mathbf{K}_{(:,X_j)}$, where the meaning of $\mathbf{K}_{(:,X_j)}$ is the same as in (20).

By comparing the forms of Eqs (20) with (9), (21) with (14), respectively, it can be found that their forms are correspondingly similar, and thus the models in Eqs. (20) and (21) can be solved similarly according to the Algorithms summarized in Tables 1 and 2, respectively.

5. Experiment

In this section, we conduct experiments to evaluate the proposed strategy on four real world aging datasets, respectively.

5.1. Dataset

In the experiments, we make evaluations of the proposed methods on four aging datasets, *i.e.*, FG-NET, Morph Album I, Morph Album II and the Images of Groups (or *Groups* for short). The FG-NET dataset consists of 1,002 facial images captured from 82 persons. In order to evaluate the proposed joint learning methods with respect to male and female age prediction, we select a subset of FG-NET with age ranging from 0 to 36 years old, since that there are very few female samples older than 36 years. For the Morph Album I dataset, there are about 1,690 facial images from about 631 persons aging from 16 to about 77 years old. And we select the subset from Morph Album I with age ranging from 16 to 44 years. Morph Album II is a relatively large aging dataset with over 55,000 images. We select the Caucasians from Album II with age ranging from 16 to 60 years for experiment. For the Groups database, it is the largest and most challenging real-world aging database captured from unrestricted Flickr. it exhibits large variations in image revolution, pose and illumination, and it contains more than 28000 persons from 5080 real life groups with ages from 0 to 80 years old. Image examples from the four datasets are respectively shown in Figure 3.

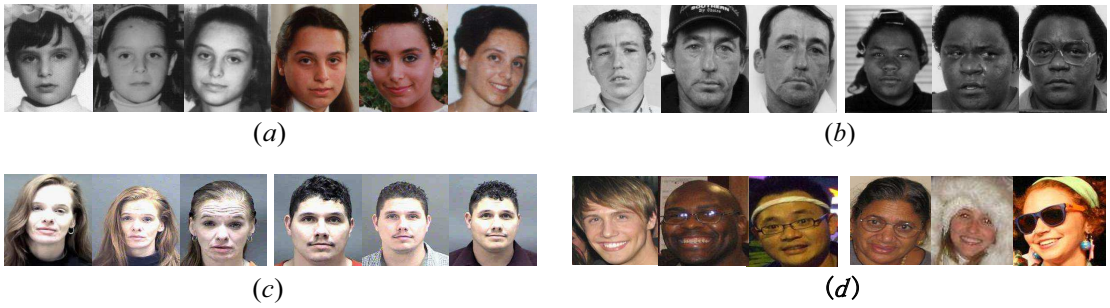

Figure 3: Image examples from the FG-NET (a), Morph Album I (b), Album II (c) and Groups (d) databases.

5.2. Experimental Setup

With the four aging datasets, we extract 200-dimensional AAM features from the FG-NET and Morph Album I databases, 152-dimensional BIF features from the Album II database, and 260-dimensional Gabor features from the Groups database, respectively. In the experiments, the optimal values of all hyper-parameters involved are searched through *cross-validation* in the range of {1e-5, 1e-2, 1e0, 1e2, 1e5}. And it is worth noting that in order to make the semantic space of gender as orthogonal to that of age as possible, we assign the λ_3, in Eqs. (9) and (14), with a relatively large value, and in the experiments we tune it in {1e0, 1e3, 1e6, 1e9, 1e12, 1e15}. Besides, for the sake of facilitating fair comparison, we accordingly converted the age labels {1, 5, 10, 16, 28, 51, 75} of the Groups database to {1, 2, 3, 4, 5, 6, 7}.

For performance measure, we uniformly adopt the commonly used classification *Accuracy Rate* (Acc., $Acc. := \frac{N_{correct}}{N_{total}}$ with $N_{correct}$ denoting the number of test samples correctly classified and N_{total} the total number of test samples) for gender classification; Concerning age estimation, we take the *Mean Absolute Error* (MAE, $MAE := \frac{1}{N}\sum_{i=1}^{N}|\widehat{l_i} - l_i|$ with l_i and $\widehat{l_i}$ denoting the ground-true and predicted age values, respectively) as the measure. To evaluate the effectiveness and superiority of the proposed methods, we compare them with the related methods, especially these state-of-the-art, i.e., PLS-based method [67], [71], multi-task based method [72], *Ensemble* method [37] and Net^{VGG}_{hybrid} [75] (*for nonlinear case comparison*).

5.3. Experimental Results and Analysis

5.3.1. Linear Case

With the extracted feature representations for FG-NET, Morph Album I and II, we conducted comparative experiment by randomly selecting a certain number of male and female samples from each age class for training, with the remaining samples for test. We averaged the experimental results over ten trials and show them in Figures 4 and 5.

From the results shown in Figure 4, it can be found that from the evaluation perspective of gender classification, joint learning based SVOR and KDLOR (in the NOSSpaces) can generate higher accuracy than their respectively

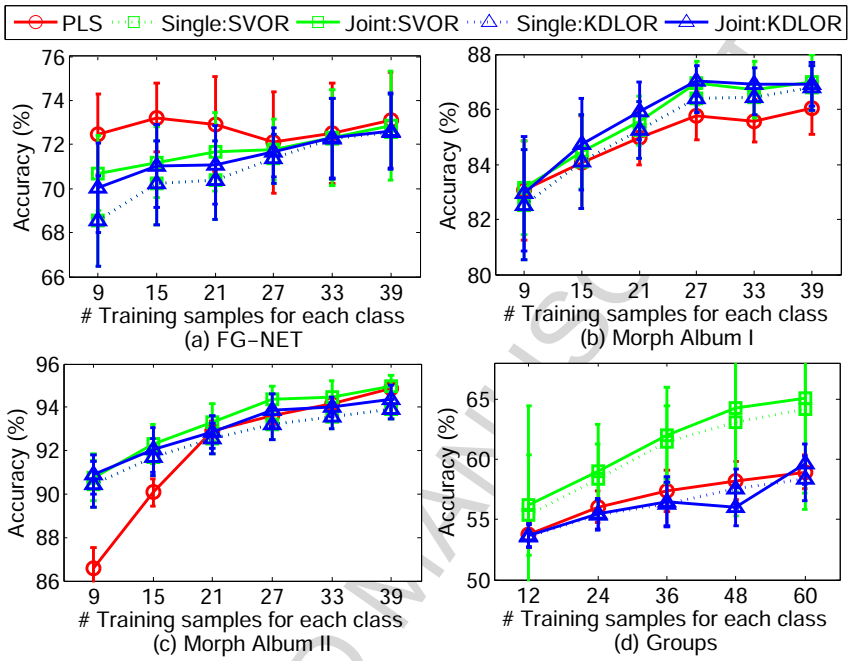

Figure 4: Comparison between the methods in terms of gender classification in linear case.

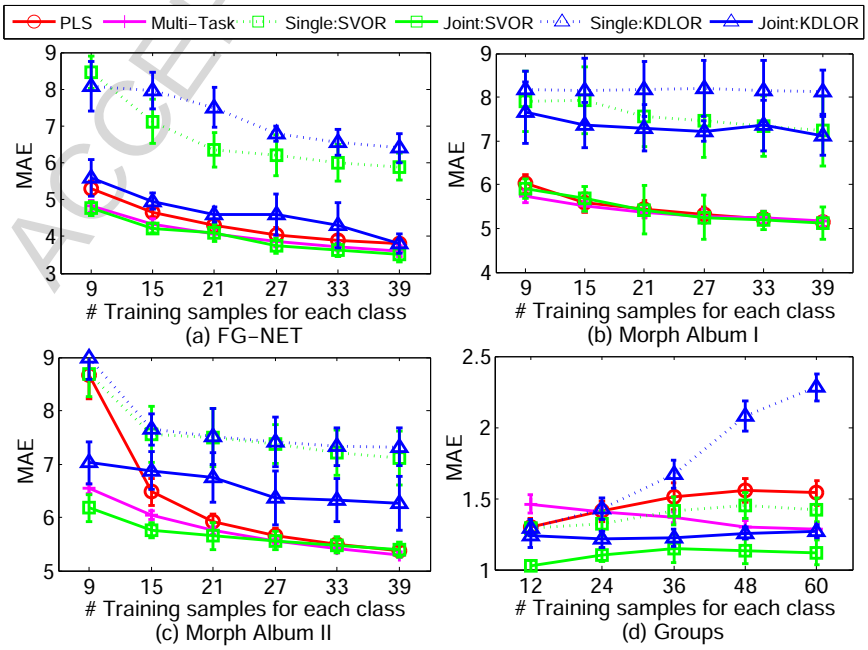

Figure 5: Comparison between the methods in terms of age estimation in linear case.

single learning based counterparts. It demonstrates the effectiveness and superiority of learning in the NOSSpaces to gender recognition.

For age estimation, the learned regressors SVOR and KDLOR in the NOSSpaces can produce significantly lower MAEs than their jingle learning based methods, as shown in Figure 5. More specifically, by the setting of joint learning in the NOSSpaces, the SVOR (KDLOR) can reduce the age estimation MAE by about 40% (35%), 30% (10%), and 25% (over 10%) on the FG-NET, Morph Album I and II, respectively. And in most cases, the joint learning based SVOR yields the lowest age estimation errors. These results demonstrate that learning in the NOSSpaces can dramatically improve the performance of human age estimation.

5.3.2. Nonlinear Case

In order to evaluate the effectiveness of the proposed learning strategy in nonlinear feature space (*i.e.*, the nonlinear NOSSpaces), we uniformly adopted the RBF kernel function to map the original feature representations of the four aging datasets (*i.e.*, FG-NET, Morph Album I and Morph Album II) to high-dimensional feature space, and conducted comparative experiment by randomly selecting a certain number of male and female samples from each age class for training, with the remaining samples for test. We averaged the experimental results over ten trials and demonstrated them in Figures 6 and 7. By making a comparison between Figures 6 versus 4, and Figures 7 versus 5, respectively

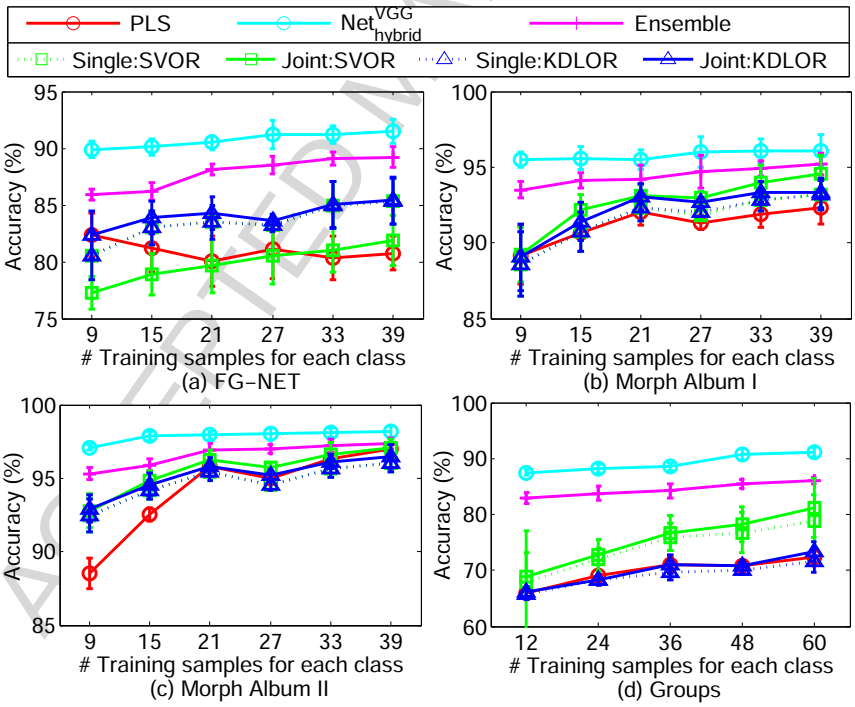

Figure 6: Comparison between the methods in terms of gender classification in nonlinear case.

in terms of GC and AE, we can find that the estimation accuracies obtained in the kernel-induced feature space are correspondingly higher than those in the original feature space. More specifically, for gender classification, the general average accuracy is increased by about 1%, while the general age prediction error is reduced by over 6% on average. And by making joint estimation in the kernel NOSSpaces, both the accuracies of gender classification and age estimation are further improved. It demonstrates the effectiveness of nonlinear NOSSpaces in improving the joint estimation of gender and age. It should be noticed that although in terms of both accuracy of GC and MAE value of AE, Net_{hybrid}^{VGG} achieves the best results, it is a highly-nonlinear method with several layers of feature nonlinear

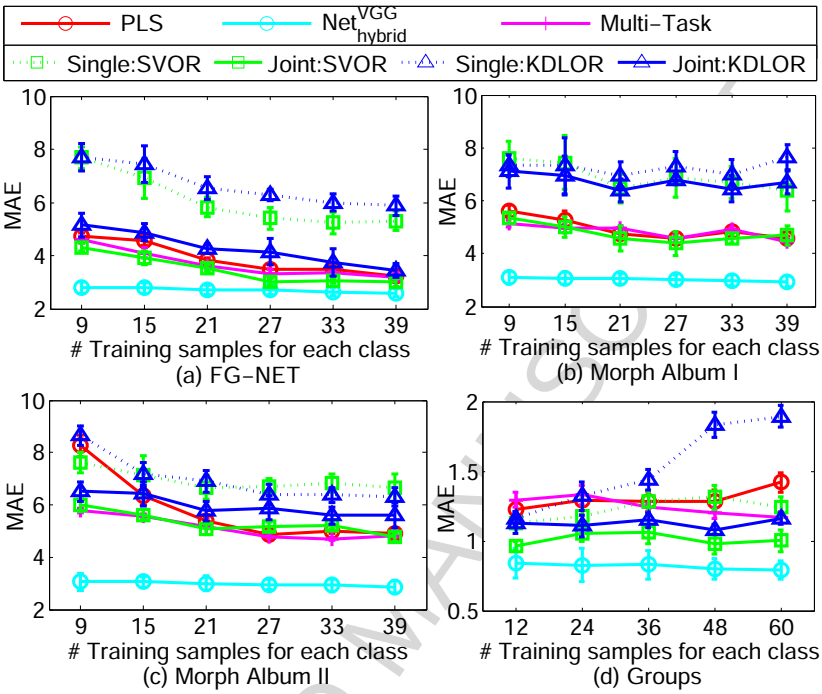

Figure 7: Comparison between the methods in terms of age estimation in nonlinear case.

transformations, and was pretrained on a large number of data samples (e.g., LFW) [4] and then fine-tuned on the experimented databases. Therefore, from the viewpoint of fair comparison with model complexity, the effectiveness and promising superiority of the proposed methods can be verified.

5.4. Near-Orthogonality Between Human Gender and Age Semantic Spaces

Now in order to investigate the near-orthogonality between human gender and age semantic spaces, we denote the intersection angle between human gender semantic direction w_g and age semantic direction w_a as θ

$$cos(\theta) = \frac{< w_g, w_a >}{\|w_g\| \cdot \|w_a\|}. \tag{22}$$

In the experiments, we find that generally, the θ lies near to $\frac{\pi}{2}$[5]. That is, the gender direction w_g is nearly but not strictly orthogonal to the age direction w_a. It witnesses the reasonableness of performing joint estimation for human gender and age in the nearly orthogonal semantic spaces (*i.e.*, the NOSSpaces).

5.5. Convergence Analysis

Through analyzing the algorithms for joint learning of SVM with KDLOR and SVM with SVOR, summarized in Tables 1 and 2, and performing the experiments, we find that it just requires 3 rounds of iterations for convergence of the joint human gender classification and age estimation in the NOSSpaces. For the sake of clarification, we provide an intuitive convergence analysis in Figure 8.

As shown in Figure 8, we make joint estimation for age estimation and gender classification in the second and third rounds of iterations of the joint learning strategy (in the NOSSpaces), respectively. And in the nonlinear NOSSpaces, it also performs age estimation and gender classification in the second and third round iterations, respectively. In other words, our nonlinear NOSSpaces just requires three rounds of iterations for the joint estimation task.

[4] Actually, the *Ensemble* method [37] is also a highly nonlinear algorithm by combining hand-crafted features with many layers of CNNs features pretrained on large image data set.

[5] In the nonlinear NOSSpaces, the intersection angle between α and β also lies near to $\frac{\pi}{2}$.

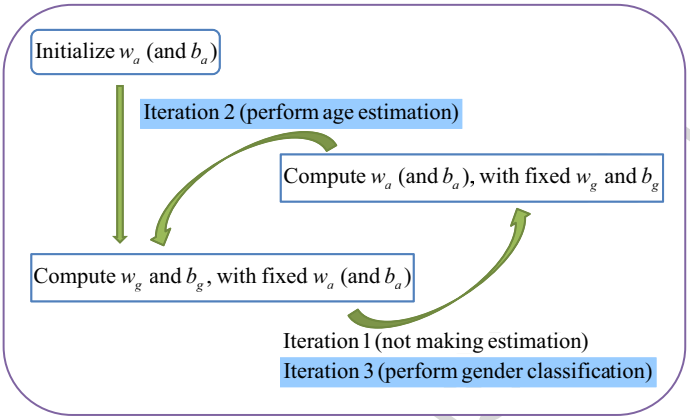

Figure 8: Convergence demonstration of the joint learning strategy for gender classification and age estimation.

6. Conclusions

In this work, we proposed a general framework for jointly estimating human gender and age, in which the binaryness of gender was considered by taking it as a binary classification problem, the ordinality of age was respected by treating it as an ordinal regression, and in particular, the semantic discrepancy between human gender and age was captured and expressed by nearly orthogonalizing their semantic spaces. In order to evaluate the proposed learning framework, we exemplified it by taking the widely used binary-class SVM for gender classification, while the discriminant learning for ordinal regression and support vector ordinal regression for age estimation, and then kernelized the joint learing framework by deriving a specific representer theorem. Finally, through experimental evaluations on four aging datasets, we demonstrated the effectiveness and superiority of the proposed methods. In the future, we consider to extend our methods to handle other joint estimation problems, such as joint estimation for human face-based age and expression, etc.

Acknowledgment

The authors first want to thank the *Pattern Recognition and Neural Computing (ParNec)* laboratory of Nanjing University of Aeronautics and Astronautics because the initial work of this paper was done there. Besides, we would like to thank Junliang Xing, et al. (Institute of Automation, Chinese Academy of Sciences) who provided source codes for our experimental comparison. This work was partially supported by the National Natural Science Foundation of China under grants 61702273 and 61472186, the Natural Science Foundation of Jiangsu Province under grant BK20170956, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant 17KJB520022, a Project Funded by the Priority Academic Program Development of Jiangsu Higer Education Institutions, and the Startup Foundation for Talents of Nanjing University of Information Science and Technology.

Appendix 1: The Proof of Lemma 4.1

Proof. (8) can be detailed as

$$\begin{aligned}\mathcal{J} &= \mathcal{L}_g(w_g; X, Y_g) + \frac{\lambda_1}{2}\|w_g\|^2 + \mathcal{L}_a(w_a; X, Y_a) + \frac{\lambda_2}{2}\|w_a\|^2 + \frac{\lambda_3}{2}(w_g^T w_a)^2. \\ &= \sum_i \mathcal{L}_g(w_g; x^i, y_g^i) + \frac{\lambda_1}{2}\|w_g\|^2 + \sum_i \mathcal{L}_a(w_a; x^i, y_a^i) + \frac{\lambda_2}{2}\|w_a\|^2 + \frac{\lambda_3}{2}(w_g^T w_a)^2.\end{aligned} \tag{23}$$

Computing the derivatives of (23) w.r.t. w_g and w_a and making them equal to zero leads to

$$\frac{\partial \mathcal{J}}{\partial w_g} = \sum_i \mathcal{L}_g^{'}(w_g; x^i, y_g^i)x^i + \lambda_1 w_g + \lambda_3(w_g^T w_a)w_a = 0, \tag{24a}$$

$$\frac{\partial \mathcal{J}}{\partial w_a} = \sum_i \mathcal{L}_g^{'}(w_a; x^i, y_a^i)x^i + \lambda_2 w_a + \lambda_3(w_g^T w_a)w_g = 0. \tag{24b}$$

Substituting (24a) into (24b) with $\lambda_2^{'} = \lambda_3(w_g^T w_a)$ yields

$$\sum_i \mathcal{L}_g^{'}(w_a; x^i, y_a^i)x^i + \lambda_2^{'}(-\sum_i \mathcal{L}_g^{'}(w_g; x^i, y_g^i)x^i - \lambda_1 w_g) + \lambda_2^{'} w_g = 0, \tag{25}$$

which further yields

$$w_g = \frac{\lambda_2^{'}\sum_i \mathcal{L}_g^{'}(w_g; x^i, y_g^i)x^i - \sum_i \mathcal{L}_g^{'}(w_a; x^i, y_a^i)x^i}{(\lambda_2^{'} - \lambda_2^{'}\lambda_1)}. \tag{26}$$

From (28), it can be found that w_g can be expressed by a linear combination of the training samples as

$$w_g = \sum_i \alpha_i x^i, \tag{27}$$

, with

$$\alpha_i = \frac{\lambda_2^{'}\sum_i \mathcal{L}_g^{'}(w_g; x^i, y_g^i) - \sum_i \mathcal{L}_g^{'}(w_a; x^i, y_a^i)}{(\lambda_2^{'} - \lambda_2^{'}\lambda_1)}. \tag{28}$$

Similarly, by substituting the $\lambda_1 w_g$ in (24a) with (24b), we can similarly obtain a combination expression with the training samples.

Finally, introducing a feature mapping function $\phi(\cdot)$ on the samples x^i along with the above proof can prove Lemma 4.1.

□

Appendix 2: Notations of PLS in Section 2.2

Table 3: Meaning of notations involved in PLS.

Notation	Meaning
X	the input matrix
Y	the output matrix
w, c	the weight vectors
t_i, u_i	the score vectors
p, q	the loading vectors

References

[1] K. K. Shashidhar, D. H. Manjaiah, Electronic customer relationship management (e-crm): Data integration for technical institutions, Advances in Intelligent Systems and Computing 216 (2014) 169–178.

[2] G. S. Linoff, M. J. Berry, Data mining techniques: for marketing, sales, and customer relationship management, John Wiley and Sons.

[3] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34 (1) (2004) 621–628.

[4] N. Ramanathan, R. Chellappa, Face verification across age progression, IEEE Transactions on Image Processing 15 (11) (2006) 3349–3361.

[5] Y. Wu, Z. Wei, H. DENG, et al., Attribute-based access to scalable media in cloud-assisted content sharing, IEEE Transactions on Multimedia 15 (4) (2013) 778–788.
[6] P. Van Leeuwen, S. Lange, A. Klein, D. Geue, D. H. Grönemeyer, Dependency of magnetocardiographically determined fetal cardiac time intervals on gestational age, gender and postnatal biometrics in healthy pregnancies, BMC pregnancy and childbirth 4 (1) (2004) 6.
[7] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, M. King, Aspects of age variation in facial morphology affecting biometrics, in: IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–6.
[8] M. Das, A. C. Loui, Automatic face-based image grouping for albuming, in: IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, 2003, pp. 3726–3731.
[9] A. C. Gallagher, T. Chen, Understanding images of groups of people, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 256–263.
[10] Y.-Y. Chen, A.-J. Cheng, W. H. Hsu, Travel recommendation by mining people attributes and travel group types from community-contributed photos, IEEE Transactions on Multimedia 15 (6) (2013) 1283–1295.
[11] H. Dibeklioglu, A. A. Salah, T. Gevers, Recognition of genuine smiles, IEEE Transactions on Multimedia 17 (3) (2015) 279–294.
[12] F. Yun, Z. Nanning, L. Jianyi, Z. Ting, Facetransfer: A system model of facial image rendering, in: IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, 2004, pp. 2180–2185.
[13] Y. Fu, N. Zheng, M-face: An appearance-based photorealistic model for multiple facial attributes rendering, IEEE Transactions on Circuits and Systems for Video Technology 16 (7) (2006) 830–842.
[14] S. Wang, Z. Gao, S. He, M. He, Q. Ji, Gender recognition from visible and thermal infrared facial images, Multimedia Tools and Applications 75 (14) (2016) 8419–8442.
[15] J. G. Wang, J. Li, Y. L. Chong, W. Y. Yau, Dense sift and gabor descriptors-based face representation with applications to gender recognition, in: International Conference on Control Automation Robotics and Vision, 2011, pp. 1860–1864.
[16] J. Aghajanian, J. Warrell, S. J. D. Prince, P. Li, Patch-based within-object classification, in: IEEE International Conference on Computer Vision, 2009, pp. 1125–1132.
[17] C. Wang, D. Huang, Y. Wang, G. Zhang, Facial image-based gender classification using local circular patterns, in: International Conference on Pattern Recognition, 2012, pp. 2432–2435.
[18] A. Dantcheva, F. Brmond, Gender estimation based on smile-dynamics, IEEE Transactions on Information Forensics and Security 12 (3) (2017) 719–729.
[19] B. Patel, R. P. Maheshwari, B. Raman, Compass local binary patterns for gender recognition of facial photographs and sketches, Neurocomputing 218 (2016) 203–215.
[20] W. Zhang, M. L. Smith, L. N. Smith, A. Farooq, Gender and gaze gesture recognition for human-computer interaction, Computer Vision and Image Understanding 149 (2016) 32–50.
[21] Y. Du, X. Lu, L. Chen, W. Zeng, An interval type-2 t-s fuzzy classification system based on pso and svm for gender recognition, Multimedia Tools and Applications 75 (2) (2016) 987–1007.
[22] M. Castrilln-Santana, J. Lorenzo-Navarro, E. Ramn-Balmaseda, Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild, Image and Vision Computing.
[23] W. Yang, C. Sun, W. Zheng, K. Ricanek, Gender classification using 3d statistical models, Multimedia Tools and Applications (2016) 1–13.
[24] M. Castrilln-Santana, J. Lorenzo-Navarro, E. Ramn-Balmaseda, On using periocular biometric for gender classification in the wild , Pattern Recognition Letters 82 (2016) 181–189.
[25] T. Nystrm, P. Mrdn, S. Kjelleberg, Multi-view gender classification using symmetry of facial images, Neural Computing and Applications 21 (4) (2012) 1–9.
[26] B. Xia, B. Ben Amor, H. Drira, M. Daoudi, L. Ballihi, Combining face averageness and symmetry for 3d-based gender classification, Pattern Recognition 48 (3) (2014) 746–758.
[27] D. Huang, H. Ding, C. Wang, Y. Wang, G. Zhang, L. Chen, Local circular patterns for multi-modal facial gender and ethnicity classification , Image and Vision Computing 32 (12) (2014) 1181–1193.
[28] L. Ballihi, B. Ben Amor, M. Daoudi, A. Srivastava, Boosting 3-d-geometric features for efficient face recognition and gender classification, IEEE Transactions on Information Forensics and Security 7 (6) (2012) 1766–1779.
[29] A. Lapedriza, M. J. Marinjimenez, J. Vitria, Gender recognition in non controlled environments, in: International Conference on Pattern Recognition, 2006, pp. 834–837.
[30] Z. Yang, H. Ai, Demographic classification with local binary patterns, in: International Conference on Advances in Biometrics, 2007, pp. 464–473.
[31] S. F. Mahmood, M. H. Marhaban, F. Z. Rokhani, K. Samsudin, O. A. Arigbabu, Fasta-elm: a fast adaptive shrinkage/thresholding algorithm for extreme learning machine and its application to gender recognition, Neurocomputing.
[32] M. D. Coco, P. Carcagni, M. Leo, P. L. Mazzeo, P. Spagnolo, Assessment of deep learning for gender classification on traditional datasets, in: IEEE International Conference on Advanced Video and Signal Based Surveillance, 2016, pp. 271–277.
[33] F. Juefeixu, E. Verma, P. Goel, A. Cherodian, M. Savvides, Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention, in: IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 136–145.
[34] J. Mansanet, A. Albiol, R. Paredes, Local deep neural networks for gender recognition, Pattern Recognition Letters 7 (2016) 80–86.
[35] G. Antipov, S. A. Berrani, J. L. Dugelay, Minimalistic cnn-based ensemble model for gender prediction from face images, Pattern Recognition Letters 70 (2015) 59–65.
[36] S. Jia, T. Lansdall-Welfare, N. Cristianini, Gender classification by deep learning on millions of weakly labelled images, in: IEEE International Conference on Data Mining Workshops, 2016, pp. 462–467.
[37] M. Castrilln-Santana, J. Lorenzo-Navarro, E. Ramn-Balmaseda, Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild, Image and Vision Computing 57 (2017) 15–24.
[38] G. Guo, G. Mu, Human age estimation: What is the influence across race and gender?, in: International Conference on Computer Vision and

Pattern Recognition Workshops, 2010, pp. 71–78.
[39] X. Geng, K. Smith-Miles, Facial age estimation by multilinear subspace analysis, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 865–868.
[40] G. Guo, G. Mu, Y. Fu, C. Dyer, T. Huang, A study on automatic age estimation using a large database, in: IEEE International Conference on Computer Vision, 2009, pp. 1986–1991.
[41] X. Wang, A study on human age estimation under facial expression changes, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2012, pp. 2547–2553.
[42] W. Li, Y. Wang, Z. Zhang, A hierarchical framework for image-based human age estimation by weighted and ohranked sparse representation-based classification, in: IAPR International Conference on Biometrics, 2012, pp. 19–25.
[43] J. Lu, V. E. Liong, J. Zhou, Cost-sensitive local binary feature learning for facial age estimation, IEEE Transactions on Image Processing 24 (12) (2015) 5356–5368.
[44] C. Li, Q. Liu, J. Liu, H. Lu, Learning ordinal discriminative features for age estimation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2012, pp. 2570–2577.
[45] C. Li, Q. Liu, J. Liu, H. Lu, Learning distance metric regression for facial age estimation, IEEE International Conference on Pattern Recognition (2012) 2327–2330.
[46] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 4920–4928.
[47] X. Yang, B. B. Gao, C. Xing, Z. W. Huo, X. S. Wei, Y. Zhou, J. Wu, X. Geng, Deep label distribution learning for apparent age estimation, in: IEEE International Conference on Computer Vision Workshop, 2015, pp. 344–350.
[48] R. Rothe, R. Timofte, L. V. Gool, Dex: Deep expectation of apparent age from a single image, in: International Conference on Computer Vision Workshop, 2015, pp. 252–257.
[49] Z. Tan, S. Zhou, J. Wan, Z. Lei, S. Z. Li, Age estimation based on a single network with soft softmax of aging modeling, in: Asian Conference on Computer Vision, 2016, pp. 203–216.
[50] K. Li, J. Xing, W. Hu, S. J. Maybank, D2c: Deep cumulatively and comparatively learning for human age estimation, Pattern Recognition 66 (2017) 95–105.
[51] G. Guo, G. Mu, Y. Fu, T. S. Huang, Human age estimation using bio-inspired features, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 112–119.
[52] K. Chen, S. Gong, T. Xiang, C. L. Chen, Cumulative attribute space for age and crowd density estimation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2013, pp. 2467–2474.
[53] Y. Zhu, Y. Li, G. Mu, G. Guo, A study on apparent age estimation, in: IEEE International Conference on Computer Vision Workshop, 2015, pp. 267–273.
[54] Q. Tian, S. Chen, Cumulative attribute relation regularization learning for human age estimation, Neurocomputing (2015) 456–467.
[55] Y. Liang, L. Liu, Y. Xu, X. Yao, Multi-task gloh feature selection for human age estimation, in: IEEE International Conference on Image Processing, 2011, pp. 565–568.
[56] Y. Fu, T. S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Transactions on Multimedia 10 (4) (2008) 578–584.
[57] J. Lu, Y. P. Tan, Fusing shape and texture information for facial age estimation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, pp. 1477–1480.
[58] J. Lu, Y. P. Tan, Ordinary preserving manifold analysis for human age and head pose estimation, IEEE Transactions on Human-Machine Systems 43 (2) (2013) 249–258.
[59] C. Hu, L. Gong, T. Wang, Q. Feng, Effective human age estimation using a two-stage approach based on lie algebrized gaussians feature, Multimedia Tools and Applications 74 (11) (2015) 1–21.
[60] X. Geng, Z. H. Zhou, Y. Zhang, G. Li, H. Dai, Learning from facial aging patterns for automatic age estimation, in: ACM International Conference on Multimedia, 2006, pp. 307–316.
[61] C. Zhang, G. Guo, Exploiting unlabeled ages for aging pattern analysis on a large database, in: IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 458–464.
[62] C. Li, Q. Liu, J. Liu, H. Lu, Learning distance metric regression for facial age estimation, 2012, pp. 2327–2330.
[63] Y. Su, Y. Fu, Q. Tian, X. Gao, Cross-database age estimation based on transfer learning, in: IEEE International Conference on Acoustics Speech and Signal Processing, 2010, pp. 1270–1273.
[64] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, in: International Conference on Applications of Computer Vision, 2015, pp. 534–541.
[65] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognition 66 (2017) 82–94.
[66] R. Rothe, R. Timofte, L. V. Gool, Deep expectation of real and apparent age from a single image without facial landmarks, International Journal of Computer Vision (2016) 1–14.
[67] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: Cca vs. pls, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2013, pp. 1–6.
[68] G. Guo, C. R. Dyer, Y. Fu, T. S. Huang, Is gender recognition affected by age?, in: IEEE International Conference on Computer Vision Workshops, 2009, pp. 2032–2039.
[69] Y. Wang, K. Ricanek, C. Chen, Y. Chang, Gender classification from infants to seniors, in: IEEE International Conference on Biometrics: Theory Applications and Systems, 2010, pp. 1–6.
[70] G. Guo, C. Zhang, A study on cross-population age estimation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2014, pp. 4257–4263.
[71] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2011, pp. 657–664.
[72] Y. Liang, L. Liu, Y. Xu, X. Yao, Multi-task gloh feature selection for human age estimation, in: IEEE International Conference on Image

Processing, 2011, pp. 565–568.
[73] D. Yi, Z. Lei, S. Z. Li, Age estimation by multi-scale convolutional network, in: Asian Conference on Computer Vision, 2015, pp. 144–158.
[74] G. Levi, T. Hassncer, Age and gender classification using convolutional neural networks, in: International conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 34–42.
[75] J. Xing, K. Li, W. Hu, C. Yuan, H. Ling, Diagnosing deep learning models for high accuracy age estimation from a single image, Pattern Recognition 66 (2017) 106–116.
[76] L. Meier, S. Van De Geer, P. Bühlmann, The group lasso for logistic regression, Journal of the Royal Statistical Society: Series B 70 (1) (2008) 53–71.
[77] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, in: Subspace, latent structure and feature selection, 2006, pp. 34–51.
[78] I. Steinwart, A. Christmann, Support vector machines, Springer Science and Business Media, 2008.
[79] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[80] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, W.-B. Li, Kernel discriminant learning for ordinal regression, IEEE Transactions on Knowledge and Data Engineering 22 (6) (2010) 906–910.
[81] W. Chu, S. S. Keerthi, New approaches to support vector ordinal regression, in: International Conference on Machine learning, 2005, pp. 145–152.

Biography

Qing Tian received the Bachelor degree in computer science from Southwest University for Nationalities, China, the Master degree in computer science from Zhejiang University of Technology, China, and PhD degree in computer science from Nanjing University of Aeronautics and Astronautics (NUAA), awarded \textit{the Sichuan province-level outstanding graduate, the Zhejiang province-level outstanding graduate, and the outstanding graduate of NUAA,} in 2008, 2011 and 2016, respectively. During Feb 2011 to Feb 2012, as a research fellow in the field of machine learning and pattern recognition, he worked at ArcSoft, Inc. USA. Since May 2016, he has been an assistant professor in the School of Computer and Software, Nanjing University of Information Science and Technology. His research interests mainly focus on machine learning and pattern recognition.

Songcan Chen received the B.S. degree from Hangzhou University (now merged into Zhejiang University), the M.S. degree from Shanghai Jiao Tong University and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics (NUAA) in 1983, 1985, and 1997, respectively. He joined in NUAA in 1986, and since 1998, he has been a full-time Professor with the Department of Computer Science and Engineering. He has authored/co-authored over 170 scientific peer-reviewed papers and ever obtained Honorable Mentions of 2006, 2007 and 2010 Best Paper Awards of Pattern Recognition Journal respectively. His current research interests include pattern recognition, machine learning, and neural computing.

Highlights

- Formulate the semantic relationship between human gender and age as near-orthogonality regulation.
- Propose a joint estimation framework for human gender and age based on the semantic regulation.
- Exemplify the proposed framework.
- Kernelize the proposed framework by deriving a representer theorem.
- Experimentally demonstrate the effectiveness and superiority of the proposed methods.