Age Estimation via Selecting Discriminated Features and Preserving Geometry

Qing Tian[1,2,3*], Heyang Sun[1§], Chuang Ma[1§], Meng Cao[1], Yi Chu[1]
[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[e-mail: tianqing@nuist.edu.cn]
[2] Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] National Laboratory of Pattern Recognition, Institue of Automation, Chinese Academy of Sciences, 100190, Beijing, China
*Corresponding author: Qing Tian

Received January 21, 2020; revised February 27, 2020; accepted March 30, 2020; published April 30, 2020

Abstract

Human apparent age estimation has become a popular research topic and attracted great attention in recent years due to its wide applications, such as personal security and law enforcement. To achieve the goal of age estimation, a large number of methods have been pro-posed, where the models derived through the cumulative attribute coding achieve promised performance by preserving the neighbor-similarity of ages. However, these methods afore-mentioned ignore the geometric structure of extracted facial features. Indeed, the geometric structure of data greatly affects the accuracy of prediction. To this end, we propose an age estimation algorithm through joint feature selection and manifold learning paradigms, so-called Feature-selected and Geometry-preserved Least Square Regression (FGLSR). Based on this, our proposed method, compared with the others, not only preserves the geometry structures within facial representations, but also selects the discriminative features. Moreover, a deep learning extension based FGLSR is proposed later, namely Feature selected and Geometry preserved Neural Network (FGNN). Finally, related experiments are conducted on Morph2 and FG-Net datasets for FGLSR and on Morph2 datasets for FGNN. Experimental results testify our method achieve the best performances.

Keywords: Age Estimation; Least Square Regression; Cumulative Attribute Coding; Feature Selection; Manifold Learning

This work was partially supported by the National Natural Science Foundation of China under grant 61702273, 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, the Fundamental Research Funds for the Central Universities No. NJ2019010, the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund and the Open Projects Program of National Laboratory of Pattern Recognition (202000007).

http://doi.org/10.3837/tiis.2020.04.017 **ISSN : 1976-7277**

1. Introduction

Age estimation (AE) has become a popular research topic in computer vision which typically predicts the age value for a given human facial image. It can be used in many scenes such as security surveillance [1], advertisement recommendation [2] and ancillary identity authentication [3]. The traditional AE process includes several steps, i.e. image preprocessing, feature extraction [4], feature reduction and decision making. The whole procedure is shown in **Fig. 1**. First, preprocessing is a necessary process which includes facial cropping, histogram equalization and fiducial points marking over eyes, nose and lips. Second, the handcrafted feature is extracted from the facial image, including Anthropometric Models [5], Active Appearance Models (AAM) [6], Local Binary Pattern (LBP) [7], Aging pattern Subspace (AGES) [8]. Third, the dimension of the extracted feature is reduced by PCA or KPCA [9] to remove the noise. At last, the reduced feature would be used for AE by a proper regressor or classifier to predict the age scalar. In this process, there are still some common problems around AE. For instances, two nearly aged human facial images seem to be similar, whose characteristic is called neighbor-similarity in AE. Thus, it is difficult to distinguish these neighbor instances in feature spaces. What's more, the fact [10] that the images at the older ages are especially rare, leading to the imbalanced categories, also influences the accuracy of predicted result. To solve those problems, a large number of methods have been proposed.

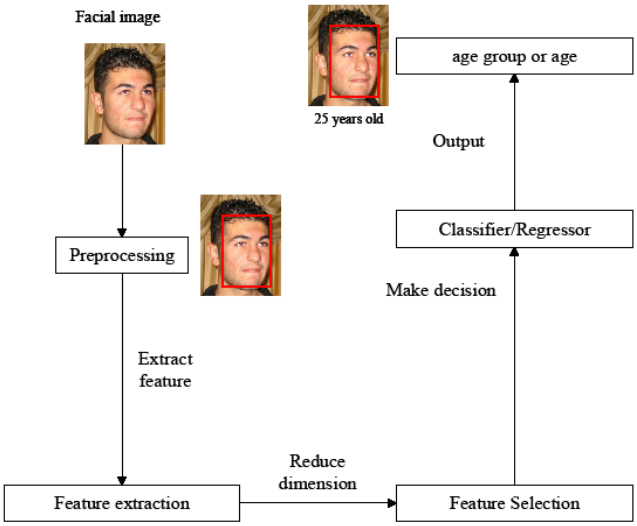


Fig. 1. The procedure of AE

If ages are regarded as separate classes of age or age group, AE can be performed as a classification problem [11,12,13]. Along this line, [14] proposed a method to extract facial features and employed the k-NN [15] as the classifier to estimate the facial age. [16] used Gaussian mixture model to classify the age group. In [17], the conditional probability net-works model (CPNN) was proposed for AE by considering the distributions of all age sequences. [18] proposed a robust framework for multi-view AE based on certain distance measure. [19] proposed a method to select the most discriminative CLBP patterns to represent faces for age classification.

Moreover, based on regression, some researchers thought AE as a regression-based problem due to the continuity of human aging, and corresponding solutions were put forward [20,21,22,23]. [24] proposed to regress facial age values by fitting a quadratic function. [25] proposed an AE method based on canonical correlation analysis (CCA) to fuse multiple feature in order to enhance facial AE. [26] introduced an ordinary preserving manifold analysis for AE, which approached to seek a low-dimensional subspace.

In recent years, deep learning has been widely used for AE [27,28,29,30]. [27] proposed a cumulative hidden layer which is supervised by a point-wise cumulative signal based on Convolutional Neural Network (CNN) architecture network [31] to combat the instance imbalance problem. [32] proposed an End-to-End learning approach based on CNN [33], trans-forming AE into a series of binary classification sub-problems. [28] introduced a method called Deep EXpectation (DEX) of apparent age, which fine-tuned on the facial images and achieved a good performance. [34] proposed Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs) to estimate age.

To solve the problem of *neighbor-similarity*, so many works were proposed. [35] proposed correlation component manifold space learning (CCMSL). In this method, a common feature space was learned by capturing the correlation between heterogeneous databases, so as to achieve better performance. Moreover, Cumulative Attribute (CA) coding was proposed for AE [36]. Differing from one-hot coding, who was just trained regraded as independent class, CA coding can despite the neighbor-similarity characteristic in AE. Least Square Regression (LSR) with CA coding is as shown in **Fig. 2**. Based on this, [37] proposed multi-variate ridge regression (mRR) to predict human facial ages and [38] derive relations between the CA coding in certain order to construct two relation matrices and then incorporate them on human facial images. However, all methods aforementioned could not embody superficial relations and embed the extracted feature in facial images. To this end, we propose a method to solve AE problem which combines the joint feature selection and manifold learning, so called Feature-selected and Geometry-preserved Least Square Regression (FGLSR). Moreover, a deep learning method based on FGLSR is also proposed, namely Feature-selected and Geometry-preserved Neural Network (FGNN). The main contributions of this paper are as follows:

1. A novel AE method called Feature-selected and Geometry-preserved Least Square Regression (FGLSR) is constructed through utilizing the neighbor-similarity of facial inherent characteristic in manifold space and selecting more discriminative feature representation.

2. To enhance AE, a deep learning method based on FGLSR is proposed, called Feature-selected and Geometry-preserved Neural Network (FGNN) which works by reconstructing the network objective function.

3. Effectiveness of the proposed method are verified through extensive experiments on both proposed methods.

The rest of this paper is organized as follows: Section 2 briefly introduces the status of research on AE. Section 3 describes the proposed method. Section 4 shows the experiment results and Section 5 concludes the paper and gives a prospect for further research.

2. Related Work

As a classic regression model, LSR has been applied on AE. In traditional LSR, the researchers usually use one-hot coding to estimate the facial age. Assuming that the *i*th sample x_i corresponds to the age A, the one-hot coding is shown in Eq. (1).

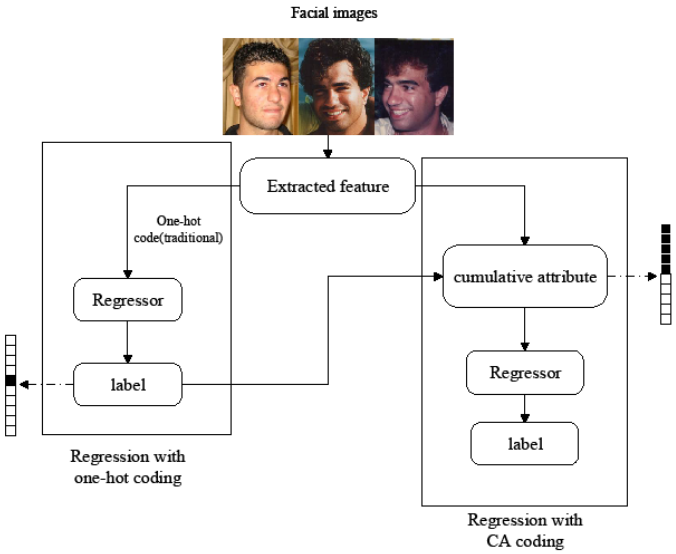


Fig. 2. Illustrations on one-hot coding and CA coding, in which the black Cube represents 1, and the other represents 0

$$y_i^j = \begin{cases} 1 & if\ j = A \\ 0 & if\ j \neq A \end{cases} \tag{1}$$

where y_i^j indicates the jth element y_i, the coding of sample x_i. Although this coding is often used on deep learning method, it treats all age as discrete classes, and do not take the relation-ship between ages into consider. For example, age 11 is close to age 12 but far away from age 60. In other words, one-hot coding is not suitable for AE problem. To solve this problem, [32] developed LSR combining CA coding and achieved a better result. CA coding is shown as follows:

$$y_i^j = \begin{cases} 1 & if\ j \leq A \\ 0 & if\ j > A \end{cases} \tag{2}$$

In Eq. 2, we can observe that when two ages are similar, CA coding will exhibit the similarity of the ages due to the difference between eleven and thirteen is two years while the difference between eleven and sixty is forty-nine years. The process of aging will be enlarged in CA coding. So, in [35], a method combined LSR and CA coding was proposed, which can be shown as following:

$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} \left\| y_i^k - \left(\mathbf{w}_k^T \mathbf{x}_i + b_k \right) \right\|_2^2 + \frac{\lambda}{2} \|\mathbf{W}\|_F^2 \tag{3}$$

where $\|\cdot\|_2^2$ means the squared L_2 norm and $\|\cdot\|_F^2$ means the squared Frobenius norm of a matrix. $\mathbf{W} = [\boldsymbol{w_1}, \ldots \boldsymbol{w_K}] \in \mathbb{R}^{d \times K}$ means the projection matrix. $\mathbf{b}$ is the regression bias and y_i^j indicates the jth regression output of $\boldsymbol{x_i}$. In fact, the value of y_i^j is decided by the vector $\boldsymbol{w_j}$ and the bias b_j. To simplify the model, we set $\widetilde{\mathbf{W}} = [\mathbf{W}; \mathbf{b}] \subseteq \mathbb{R}^{(d+1) \times K}$ and $\widetilde{\mathbf{X}} = [\mathbf{X}; \mathbf{1}] \subseteq \mathbb{R}^{(d+1) \times N}$, then Eq. (3) can be expressed.

$$\min_{\widetilde{\mathbf{W}}} \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{y_i} - \widetilde{\mathbf{W}}^T \tilde{\mathbf{x}}_i \right\|_2^2 + \frac{\lambda}{2} \left\| \widetilde{\mathbf{W}} \right\|_F^2 \tag{4}$$

To extract more useful feature, $L_{2,1}$ norm was used on AE which is used for jointly feature selection. $L_{2,1}$ norm can be seen as the sum of L_2 norm of each row in matrixes. $L_{2,1}$ norm combines the weight of the same feature of different modes, so that some common features can be selected jointly. The multi-row elements of weight matrix $\mathbf{W}$ obtained by multi-task feature selection model are all 0, which is called row-sparse matrix. The value of $L_{2,1}$ norm is calculated.

$$\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{N} x_{i,j}^2} = \sum_{i=1}^{d} \|\mathbf{x}_i\|_2 \tag{5}$$

where $x_{i,j}$ is the element of the ith row and the jth column and $\mathbf{x}_i$ means the ith row in the matrix $\mathbf{X}$.

Manifold learning is also used on regularization. Manifold in machine learning borrows the concept of manifold in mathematics. More often, the manifold in machine learning means that data is distributed on a low-dimensional manifold in a high-dimensional space, such as **Fig. 3**.

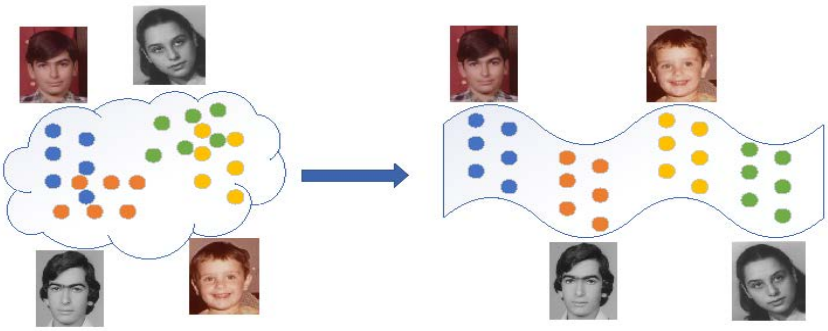


Fig. 3. The original samples and manifold embedding samples

In **Fig. 3**, the facial feature information shows the characteristics of high dimension, but in fact, they also can be expanded in low dimension space. Moreover, many real data have similar properties, such as the same face in different lighting environments. And the manifold tries to recover low dimensional data from high dimensional space to embody the sample features more intuitively. In this paper, we combine the traditional LSR with jointly feature selection and manifold, and achieve a good performance.

3. Proposed Methodology

3.1 Overall of the Model

As described in Eq. (4), although CA coding regressors are trained on the same data, they are trained separately and fail to take advantage of the potential correlation between and within them. In order to overcome these shortcomings, in this section, we propose a method called Feature-selected and Geometry-preserved Least Square Regression (FGLSR) and its deep learning extension Feature-selected and Geometry-preserved Neural Network (FGNN) to construct a proper regularizer and leverage the joint feature selection of the model. These correlations are embedded into the CA-coding regression process and merged to improve the accuracy of the subsequent CA coding based on AE multi-output regression model.

3.2 Joint Feature Selection

$L_{2,1}$ norm is usually used on feature selection. It can be seen as the L_2 norm in each row of the origin matrix. The reason we choose $L_{2,1}$ norm is that the facial feature has the characteristic of sparsity. Multi task feature selection makes use of the relevant information between tasks, and there are similar sparse patterns among the variables of different tasks. $L_{2,1}$ norm combines the weights of the same features of different modes, so that some common features can be selected

jointly. The elements of weight matrix $\mathbf{W}$ obtained by the multi task feature selection model are all 0, which is "row-sparsity".

3.3 The Regularization of Manifold

In this paper we take another regularization to constrain our model through manifold learning. Assuming that the samples $\mathbf{z}_i$, $\mathbf{z}_j$ where $\mathbf{z}_a = \widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_a$, $a = i$ or j and their relationship parameter $\boldsymbol{W}_{\mathrm{i,j}}$. We choose minimize the following objective function Eq. (6) to close the two connected points.

$$\sum_{i,j=1}^{N}\left(\boldsymbol{z}_i-\boldsymbol{z}_j\right)^2\mathbf{S}_{ij} \tag{6}$$

The objective function with our choice of $\mathbf{S}_{ij}$ incurs a heavy penalty if neighboring points $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are mapped far apart. So, minimizing Eq. (6) can ensure $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ is close enough. Combining $\mathbf{z}_i = \widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i$, Eq. (6) can be transferred as follows:

$$\begin{aligned}\frac{1}{2}\sum_{i,j=1}^{N}\left(\boldsymbol{z}_i-\boldsymbol{z}_j\right)^2\mathbf{S}_{ij} &= \frac{1}{2}\sum_{i,j=1}^{N}\left(\widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i-\widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_j\right)^2\mathbf{S}_{ij}\\ &= \sum_{i=1}^{N}\tilde{\mathbf{x}}_i{}^{\boldsymbol{T}}\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i\mathbf{D}_{\boldsymbol{ii}}-\sum_{i,j=1}^{N}\tilde{\mathbf{x}}_j{}^{\boldsymbol{T}}\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_{\boldsymbol{j}}\mathbf{S}_{ij}\\ &= tr\left(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}(\mathbf{D}-\mathbf{S})\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}\right)=tr\left(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}\right)\end{aligned} \tag{7}$$

where $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ means Laplacian matrix, $\boldsymbol{D}$ is the diagonal matrix where $\boldsymbol{D}=\sum_{\mathrm{j}=1}^{\mathrm{N}}\mathbf{S}_{ij}$. $\mathbf{S}$ is an auxiliary matrix which is defined as follows:

$$\mathbf{S}_{ij}=\begin{cases}1 & if\ \mathbf{x}_i\ and\ \mathbf{x}_i\ in\ the\ same\ class\\ 0 & otherwise\end{cases} \tag{8}$$

Therefore, minimizing the last term can hold the original structure information, which is referred as follows:

$$\min_{\widetilde{\mathbf{W}}} tr\left(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}\right) \tag{9}$$

3.4 The Summary of our Model

Combining Eq. (4), $L_{2,1}$ norm and Eq. (9), our model can be described as following:

$$\min_{\widetilde{\mathbf{W}}}\frac{1}{2}\left\|\widetilde{\mathbf{W}}\right\|_F^2+\frac{\lambda_1}{2}\left\|\widetilde{\mathbf{W}}\right\|_{2,1}+\frac{\lambda_2}{2}\sum_{i=1}^{N}\left\|\mathbf{y}_i-\widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i\right\|_2^2+\frac{\lambda_3}{2}tr\left(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}\right) \tag{10}$$

To sum up, we choose LSR as our baseline for the popularity in age estimation. Then, considering the row-sparsity in the projection $\widetilde{\mathbf{W}}$, we add an additional regularization $\|\widetilde{\mathbf{W}}\|_{2,1}$. Moreover, we find the complex facial feature can be embedded into a manifold space which can describe the feature in a low demission space. At last, our model is completed successfully.

3.5 Model Optimization

Let

$$\boldsymbol{J} = \frac{1}{2}\|\widetilde{\mathbf{W}}\|_F^2 + \frac{\lambda_1}{2}\|\widetilde{\mathbf{W}}\|_{2,1} + \frac{\lambda_2}{2}\sum_{i=1}^{N}\|\mathbf{y}_i - \widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i\|_2^2 + \frac{\lambda_3}{2}tr(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}) \tag{11}$$

We propose to use the accelerated approximate gradient (APG) algorithm [39] to optimize the objective function (11). Specifically, Eq. (11) is first divided into two parts, namely smooth part

$$f(\widetilde{\mathbf{W}}) = \frac{1}{2}\|\widetilde{\mathbf{W}}\|_F^2 + \frac{\lambda_2}{2}\sum_{i=1}^{N}\|\mathbf{y}_i - \widetilde{\mathbf{W}}^T\tilde{\mathbf{x}}_i\|_2^2 + \frac{\lambda_3}{2}tr(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{W}}) \tag{12}$$

and non-smooth part

$$g(\widetilde{\mathbf{W}}) = \frac{\lambda_1}{2}\|\widetilde{\mathbf{W}}\|_{2,1} \tag{13}$$

Then, a function is built which is shown as follows:

$$\Omega_l(\widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}_i) = f(\widetilde{\mathbf{W}}_i) + \langle \widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}_i, \nabla f(\widetilde{\mathbf{W}}_i)\rangle + \frac{l}{2}\|\widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}_i\|_F + g(\widetilde{\mathbf{W}}) \tag{14}$$

where $\nabla f(\widetilde{\mathbf{W}}_i)$ means the gradient of $f(\widetilde{\mathbf{W}})$ in the ith iteration, l means the length of step whose value can be determined by linear search. Finally, the update step of $\widetilde{\mathbf{W}}_i$ in APG is shown as following:

$$\begin{aligned}\widetilde{\mathbf{W}}_{i+1} &= \arg\min_{\widetilde{\mathbf{W}}} \frac{1}{2}\|\widetilde{\mathbf{W}} - \mathbf{V}\|_F^2 + \frac{1}{l}g(\widetilde{\mathbf{W}}) \\ &= \arg\min_{\widetilde{\mathbf{w}}_1,\ldots,\widetilde{\mathbf{w}}_d} \frac{1}{2}\sum_{i=1}^{d}\|\widetilde{\mathbf{w}}_j - \mathbf{v}_j\|_2^2 + \frac{\beta}{l}\|\widetilde{\mathbf{w}}_j\|_2^2\end{aligned} \tag{15}$$

where $\widetilde{\boldsymbol{w}}_i$ and $\boldsymbol{v}_i$ is the ith row of $\widetilde{\mathbf{W}}$ and $\mathbf{V}$, where $\mathbf{V}$ is defined as

$$\mathbf{V} = \widetilde{\mathbf{W}}_i - \frac{1}{l}\nabla f(\widetilde{\mathbf{W}}_i) \tag{16}$$

So, according to Eq. (15), the problem to be optimized is transformed into a k-subproblem. Moreover, the key of APG is how to solve this update step. Additionally, the analytical solution in Eq. (15) is shown as following:

$$\widetilde{\mathbf{w}}_j^* = \begin{cases} \left(1 - \frac{\beta}{l\|\mathbf{v}_j\|_2}\right)\mathbf{v}_j & \|\mathbf{v}_j\|_2 > \frac{\beta}{l} \\ 0 & otherwise \end{cases} \tag{17}$$

Additionally, [38] proposed an alternative to solve the gradient descent, which uses another search point to optimize the problem, as shown in Eq. (18):

$$\mathbf{Q}_i = \widetilde{\mathbf{W}}_i + \alpha_i\left(\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_{i-1}\right) \tag{18}$$

where $a_i = \frac{(1-\rho_{i-1})\rho_i}{\rho_{i-1}}$ and $\rho_i = \frac{2}{i+3}$.

We repeat the above steps alternately until convergence and generate the optimal solutions of the value of Eq. (11). The optimization procedure is summarized in Algorithm 1.

Algorithm 1: The Optimization Algorithm for our method

Input:

The training data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ from training set, along with their corresponding response vector $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$

Output:

The weighted matrix $\widetilde{\mathbf{W}}$

Initialize $\beta \geq 0, \gamma \geq 0, \sigma > 1, l_0 \geq 0, \widetilde{\mathbf{W}}_0 = \widetilde{\mathbf{W}}_1 = \mathbf{0}$;

repeat

  Compute the search point $\mathbf{Q}_i$ according to Eq. (18);

  Compute $\mathbf{W}_i$ according to Eq. (15);

  repeat

    $l = \sigma l$;

  until $\left(f\left(\widetilde{\mathbf{W}}_{i+1}\right) + g\left(\widetilde{\mathbf{W}}_{i+1}\right)\right) > \Omega_l\left(\widetilde{\mathbf{W}}_{i+1}, \boldsymbol{Q}_i\right)$;

  $i = i + 1$;

until Convergence;

3.6 The Deep Learning Method

By far, we have introduced our FGLSR for AE. However, the traditional model cannot meet the existing accuracy requirements. To this end, a method called Feature-selected and Geometry-preserved Neural Network (FGNN) based on FGLSR is proposed by adding the FGLSR term and incorporating with CNN. Based on [39], the features gradually become specific when the layers go toward the top one, which results in increasing dataset bias for the higher layers of features. Therefore, additional regularizer based on $L_{2,1}$ norm and manifold embedding will perform a good role compared with common regularizer. Above all, the objective function of FGNN is constructed as follows:

$$\min_{\mathbf{W}} \sum_{i=1}^{N} l\left(x_i, y_i; \mathbf{W}\right) + \lambda_1 \| \mathbf{W} \|_{2,1} + \lambda_2 \operatorname{trace}\left(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}\right) \tag{19}$$

where $\mathbf{W}$ denotes the model parameters to be learned, λ_1 and λ_2 are two tradeoff parameters. The first term of Eq. (19) is the Least Square Regression loss. The second term is the $L_{2,1}$ norm loss and the last term is the manifold regularization.

4. Experimental Classification Results and Analysis

To evaluate the proposed method, we performed experiments on facial datasets from the viewpoint of performance comparison and parameter analysis.

4.1 Datasets

In this experiment, Morph Album 2 and FG-Net were used respectively. Morph Album 2 (referred to as Morph2) is one of the most popular age estimation datasets at present. Morph2 is also a cross-time dataset, which contains images of the same person in different age. Specifically, the dataset includes 55134 pictures of about 13000 people, in which the time span of image collection is from 2003 to 2007 and the age of the volunteers is from 16 to 77 years old, 33 years old average. In addition, the dataset of morph2 records other information about the volunteers, such as gender, race, whether to wear glasses, etc. FG-Net contains 82 images of different ages, and provides 68 facial key point information in each image. In view of the cross-age characteristics of this dataset, FG-Net can be used in age estimation, cross age face recognition, age progression and other research directions. **Fig. 4** shows the samples of the two datasets.

(a) Morph2

(b) FG-Net

Fig. 4. Facial image examples from (a) Morph2 and (b) FG-Net

4.2 Experimental Setup

In this experiment, our proposed method was compared with the related methods such as LSR and LSR with $L_{2,1}$ norm. In this experiment, we extracted BIF coefficients as feature representation on Morph2 and reduce their dimension to 146 (90%) through PCA and extracted AAM coefficients as feature representation on FG-Net, reducing dimension to 200. For hyperparameters, all were set to five-fold cross validation. For AE, 50%, 70% and 90% overall dataset was selected as training sets and the other as testing sets, respectively. All experiments were repeated ten times with random disruption of samples. In addition, a common performance measure, Mean Absolute Errors (MAE) was used in all experiments, which is shown as following:

$$\mathrm{MAE}_{y_i,\hat{y}_i} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{20}$$

where $\hat{y}_i$ means the predicted value and y_i means the true value.

4.3 Results and Analysis

4.3.1 Results on Datasets

We performed evaluations on Facial datasets. The results are shown in **Table 1** and **Table 2**. From **Table 1** and **Table 2** we observe the following findings. Firstly, our method achieves the best result compared with the other methods on different datasets. Secondly, with the training data increasing, the value of MAE decreases rapidly. This phenomenon is reasonable, because when the number of training samples increase, the characteristics that can be learned from the samples become more obvious, leading to more reasonable prediction. Thirdly, our method achieved the minimum variance, which shows that our method has the stability of prediction to some extent. Additionally, the Firedman test with $\alpha = 0.05$ followed by Nemenyi Test, conducted on Morph2 dataset in different methods, as shown in **Fig. 5**, also proves the effectiveness of our method.

Table 1. The MAE on Morph2, where lower is better

	LSR	LSR with $L_{2,1}$ norm	FGLSR
70%	5.37 ± 0.41	4.88 ± 0.32	4.47 ± 0.23
50%	5.37 ± 0.40	5.20 ± 0.39	4.49 ± 0.21
30%	5.73 ± 0.40	5.14 ± 0.50	4.79 ± 0.32

Table 2. The MAE on Morph2, where lower is better

	LSR	LSR with $L_{2,1}$ norm	FGLSR
70%	7.02 ± 1.19	5.34 ± 0.69	5.24 ± 0.95
50%	6.91 ± 0.89	5.80 ± 0.39	5.20 ± 0.58
30%	6.99 ± 0.45	6.62 ± 0.50	5.71 ± 0.40

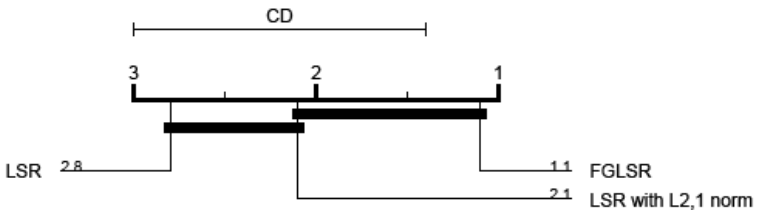
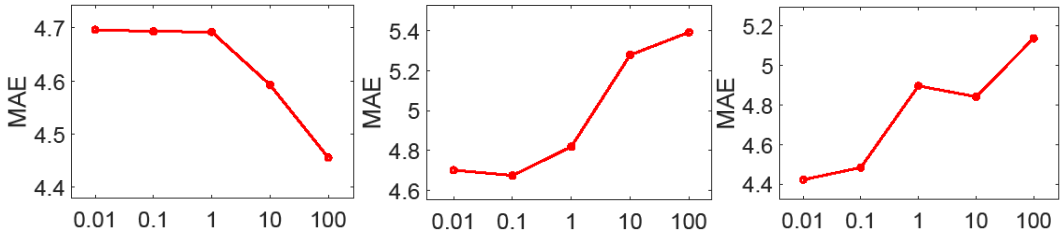


Fig. 5. The Firedman test on Morph2

4.3.2 Parameters Analysis

Then we evaluated the sensitivity of the hyper-parameters involved in our method. To this end, we set the value of λ_1 to λ_3 in the range of [0.01,0.1,1,10,100]. From **Fig. 6** we can observe the following findings: Firstly, with the increase of λ_1, the results get better which may be because the weight of $L_{2,1}$ norm overwhelms the other terms. Secondly, the results get worse with the growing value of λ_2 and λ_3, this proves that too much weight has a negative effect on the result. To sum up, for a better result we decide to decrease the value of λ_2 and λ_3 in the cross validation.



(a) λ_1 (b) λ_2 (c) λ_3

Fig. 6. Sensitivity analysis on hyper-parameters on Morph2 dataset

4.3.3 Results in Deep Learning

Moreover, an additional experiment on FGNN is put into effect. In this experiment, a pre-trained VGG-16 network was used to extract the deep feature [40] on datasets. Then, 80% of images in datasets were used on training and the remaining on testing. To verify the advancement of our method, we chose several methods to compare our method, such as OR-CNN, CasCNN, ARN, SSR-Net. Morph2 was used on experiment while FG-Net was ignored for it' s poor number of images. **Fig. 7** confirms our method still achieves the best result.

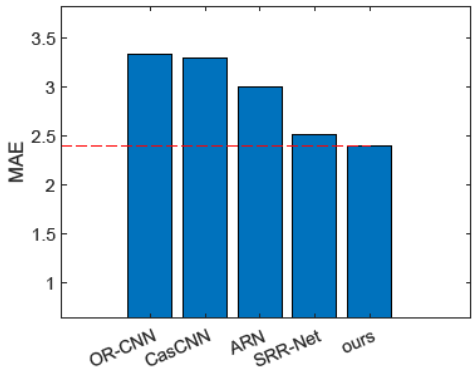


Fig. 7. The results on Morph2, the red line means the value of MAE in our method

5. Conclusion

In this paper, we proposed a novel age estimation model, coined as Feature-selected and Geometry-preserved Least Square Regression (FGLSR). For the sake of describing the facial ageing process, we chose LSR as our empirical loss function with CA coding to depict the facial intrinsic characteristic. Considering there is redundance in feature representations and the original feature space usually lies in low dimensional manifold space, we constructed the L2,1 norm and the manifold embedding as our regularization terms, which ensures the model be able to select more discriminative feature representation components and preserve the

geometry structure simultaneously. For better model performance, a CNN-based deep learning extension was designed, namely Feature-selected and Geometry-preserved Neural Network (FGNN). Finally, extensive experiments in facial datasets have validated that our proposed method can achieve superior performance to the compared ones.

In the future, we will consider to predict age with significant feature selection in adaptive manner, instead of the L2,1 norm regularization. Another future direction is to address age estimation in more challenging heterogeneous domain scenarios.

References

[1] Guo. G, Fu. Y, Dyer. C.R, and Huang. T.S, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol.17, pp.1178–1188, 2008. Article (CrossRef Link)

[2] Fjermestad. J, and Romano. N.C, "Electronic customer relationship management," *Business Process Management Journal*, Vol. 9 No. 5, pp. 572-591, 2003.Article (CrossRef Link)

[3] Jain. A.K, Dass. S.C, and Nandakumar. K, "Soft biometric traits for personal recognition systems," in *Proc. of International conference on biometric authentication, Springer*, pp. 731–738, 2004.Article (CrossRef Link)

[4] Chen. Y, Xu. W, Zuo. J, and Yang. K, "The fire recognition algorithm using dynamic feature fusion and IV-SVM classifier," *Cluster Computing*, vol. 22, pp. 7665–7675, 2018. Article (CrossRef Link)

[5] Torres. H.R, Oliveira. B, Veloso. F, Ruediger. M, Burkhardt. W, Moreira. A, Dias. N, Morais. P, Fonseca. J.C, and Vilaca. J.L, "Deep learning-based detection of anthropometric landmarks in 3d infants head models," in *Proc. of Medical Imaging 2019: Computer-Aided Diagnosis, International Society for Optics and Photonics*, p. 1095034, 2019.Article (CrossRef Link)

[6] Liu. J, Shen. C, Liu. T, Aguilera. N and Tam. J, "Active appearance model induced generative adversarial network for controlled data augmentation," in *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer*, pp.201–208, 2019.Article (CrossRef Link)

[7] Zhao. H, Zhan. Z.H, Lin. Y, Chen. X, Luo. X.N, Zhang. J, Kwong. S, and Zhang. J, "Local binary pattern-based adaptive differential evolution for multimodal optimization problems," *IEEE transactions on cybernetics*, 2019.Article (CrossRef Link)

[8] Geng. X, Smith-Miles. K, and Zhou. Z.H, "Facial age estimation by nonlinear aging pattern subspace," in *Proc. of the 16th ACM international conference on Multimedia, ACM*, pp. 721–724, 2008.Article (CrossRef Link)

[9] Kim. K.I, Jung. K, and Kim. H.J, "Face recognition using kernel principal component analysis," *IEEE signal processing letters*, vol. 9, pp. 40–42, 2002.Article (CrossRef Link)

[10] Tian. Q, and Chen. S, "Cross-heterogeneous-database age estimation through correlation representaion learning," *Neurocomputing*, vol. 238, pp. 286–295, 2017.Article (CrossRef Link)

[11] Alnajar. F, Shan. C, Gevers. T, and Geusebroek. J.M, "Learning-based encoding with soft assign-ment for age estimation under unconstrained imaging conditions," *Image and Vision Computing* vol. 30, pp. 946–953, 2012.Article (CrossRef Link)

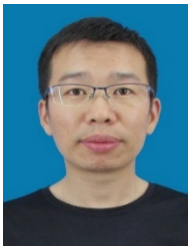
[12] Sai. P.K, Wang. J.G, and Teoh. E.K, "Facial age range estimation with extreme learning machines," *Neurocomputing*, vol. 149, pp. 364–372, 2015.Article (CrossRef Link)

[13] Tian. Q, Cao. M, and Ma. T, "Feature relationships learning incorporated age estimation assisted by cumulative attribute encoding," *Computers, Materials & Continua*, vol. 56, no. 3, pp. 467–482, 2018.Article (CrossRef Link)

[14] Lanitis. A, Draganova. C, and Christodoulou. C, "Comparing different classifiers for automatic age estimation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, pp. 621–628, 2004.Article (CrossRef Link)

[15] Deng. Z, Zhu. X, Cheng. D, Zong. M, and Zhang. S, "Efficient classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.Article (CrossRef Link)
[16] Ueki. K, Hayashida. T, and Kobayashi. T, "Subspace-based age-group classification using facial images under various lighting conditions," in *Proc. of 7th International Conference on Automatic Face and Gesture Recognition (FGR06), IEEE*, pp. 6–pp, 2006.Article (CrossRef Link)
[17] Geng. X, Yin. C, and Zhou. Z.H, "Facial age estimation by learning from label distributions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 2401–2412, 2013.Article (CrossRef Link)
[18] Li. Z, Fu. Y, and Huang. T.S, "A robust framework for multiview age estimation," in *Proc. of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE*, pp. 9–16, 2010.Article (CrossRef Link)
[19] Torrisi. A, Farinella. G.M, Puglisi. G, and Battiato. S, "Selecting discriminative clbp patterns for age estimation," in *Proc. of 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE*, pp. 1–6, 2015.Article (CrossRef Link)
[20] Fu. Y, Xu. Y, and Huang. T.S, "Estimating human age by manifold analysis of face pictures and regression on aging features," in *Proc. of 2007 IEEE International Conference on Multimedia and Expo, IEEE*, pp. 1383–1386, 2007.Article (CrossRef Link)
[21] Yan. S, Wang. H, Tang. X, and Huang. T.S, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. of 2007 IEEE 11th International Conference on Computer Vision, IEEE*, pp. 1–8, 2007.Article (CrossRef Link)
[22] Luu. K, Ricanek. K, Bui. T.D, and Suen. C.Y, "Age estimation using active appearance models and support vector machine regression," in *Proc. of 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, IEEE*, pp. 1–5, 2009.Article (CrossRef Link)
[23] Yan. S, Wang. H, Huang. T.S, Yang. Q, and Tang. X, "Ranking with uncertain labels," in *Proc. of 2007 IEEE International Conference on Multimedia and Expo, IEEE*, pp. 96–99, 2007.
Article (CrossRef Link)
[24] Lanitis. A, Taylor. C.J, and Cootes. T.F, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.442–455, 2002.Article (CrossRef Link)
[25] Lu. J, and Tan. Y.P, "Fusing shape and texture information for facial age estimation," in *Proc. of* 2011 *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE*, pp. 1477–1480, 2011.Article (CrossRef Link)
[26] Lu. J, Tan. Y.P, "Ordinary preserving manifold analysis for human age and head pose estimation," *IEEE Transactions on Human-Machine Systems*, vol. 43, pp. 249–258, 2012.
Article (CrossRef Link)
[27] Li. K, Xing. J, Hu. W, and Maybank. S.J, "D2c: Deep cumulatively and comparatively learning for human age estimation," *Pattern Recognition*, vol. 66, pp. 95–105, 2017.Article (CrossRef Link)
[28] Rothe. R, Timofte. R, and Van Gool. L, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, pp. 144–157, 2018.Article (CrossRef Link)
[29] Shen. W, Guo. Y, Wang. Y, Zhao. K, Wang. B, and Yuille. A.L, "Deep regression forests for age estimation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2304–2313, 2018.Article (CrossRef Link)
[30] Taheri. S, and Toygar. O, "On the use of dag-cnn architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, 2019.Article (CrossRef Link)
[31] Gui. Y, and Zeng. G, "Joint learning of visual and spatial features for edit propagation from a single image," *The Visual Computer*, vol. 36, pp. 469–482, 2020.Article (CrossRef Link)
[32] Niu. Z, Zhou. M, Wang. L, Gao. X, and Hua. G, "Ordinal regression with multiple output cnn for age estimation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 4920–4928, 2016.Article (CrossRef Link)
[33] Zhang. Y, Wang. Q, Li. Y, and Wu. X, "Sentiment classification based on piecewise pooling convolutional neural network," *Computers, Materials & Continua*, 2018.

[34] Chen. K, Gong. S, Xiang. T, and Change Loy. C, "Cumulative attribute space for age and crowd density estimation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 2467–2474, 2013.Article (CrossRef Link)
[35] An. S, Liu. W, and Venkatesh. S, "Face recognition using kernel ridge regression," in *Proc. of 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE*, pp. 1–7, 2007.
Article (CrossRef Link)
[36] Tian. Q, and Chen. S, "Cumulative attribute relation regularization learning for human age estimation," *Neurocomputing*, vol. 165, pp. 456–467, 2015.Article (CrossRef Link)
[37] Coates. A, and Ng. A.Y, "The importance of encoding versus training with sparse coding and vector quantization," in *Proc. of the 28th international conference on machine learning (ICML-11)*, pp. 921–928, 2011.Article (CrossRef Link)
[38] Liu. J, and Ye. J, "Efficient l1/lq norm regularization," *arXiv preprint arXiv:1009.4766*, 2010.Article (CrossRef Link)
[39] Yosinski. J, Clune. J, Bengio. Y, and Lipson. H, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, pp. 3320–3328, 2014.
Article (CrossRef Link)
[40] Zhang. J, Jin. X, Sun. J, Wang. J, and Sangaish. A. K, "Spatial and semantic convolutional features for robust visual object tracking," *Multimedia Tools and Applications*, 2018.
Article (CrossRef Link)



Qing Tian received his Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, China, in 2016. He is currently an Associate Professor in the School of Computer and Software, Nanjing University of Information Science and Technology, China. He was an Academic Visitor at the University of Manchester, UK, from 2018 to 2019. He was the recipient of the National PhD Scholarship Award of China, the Best Scientific Paper Award of ICPR, the Excellent Doctoral Dissertation Award of Jiangsu Province of China, etc. He has published nearly 30 peer-reviewed scientific papers and served as PC member for prestigious international conferences, such as IJCAI, PRICAI, and reviewer for prestigious journals and conferences, such as IEEE TPAMI, IEEE TNNLS, IEEE TCYB, IEEE TIFS, IJCAI, ICDM, CVPR. His research interests include machine learning and pattern recognition.

Heyang Sun received his B.S. degree in computer science from Nanjing University of Information Science and Technology (NUIST) in 2019, China. He is currently pursuing his master degree at the NUIST. His research interests include machine learning and pattern recognition.

Chuang Ma received his B.S. degree in computer science from Nanjing University of Information Science and Technology (NUIST) in 2018, China. He is currently pursuing his master degree at the NUIST. His research interests include machine learning and pattern recognition.

Meng Cao received his B.S. degree in computer science from Nanjing University of Information Science and Technology (NUIST) in 2017, China. He is currently pursuing his master degree at the NUIST. His research interests include machine learning and pattern recognition.

Yi Chu is currently pursuing his B.S. degree at Nanjing University of Information Science and Technology (NUIST). His research interests include machine learning and pattern recognition.