Pattern Recognition 46 (2013) 628–641

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Facial age estimation based on label-sensitive learning and age-oriented regression

Wei-Lun Chao*, Jun-Zuo Liu, Jian-Jiun Ding

Graduate Institute of Communication Engineering, National Taiwan University, No. 1, Section 4, Roosevelt Rd., Taipei 10617, Taiwan

ARTICLE INFO

Article history:
Received 16 March 2012
Received in revised form
9 September 2012
Accepted 17 September 2012
Available online 24 September 2012

Keywords:
Machine learning
Pattern recognition
Manifold learning
Dimensionality reduction
Distance metric learning
Local regression
Age estimation

ABSTRACT

This paper provides a new age estimation approach, which distinguishes itself with the following three contributions. First, we combine distance metric learning and dimensionality reduction to better explore the connections between facial features and age labels. Second, to exploit the intrinsic ordinal relationship among human ages and overcome the potential data imbalance problem, a label-sensitive concept and several imbalance treatments are introduced in the system training phase. Finally, an age-oriented local regression is presented to capture the complicated facial aging process for age determination. The simulation results show that our approach achieves the lowest estimation error against existing methods.

1. Introduction

In the last few decades, with the increasing need of automatic recognition and surveillance systems, researches on human faces – including face detection, face recognition, gender classification, and facial expression recognition – have attracted significant attention in both computer vision and pattern recognition. Compared to these face-related researches, facial age estimation is a relatively new topic yet with several interesting and important potential usage. For example, an automatic age estimation system can not only facilitate the human–computer interface, but also prevent under ages from accessing cigarettes, bears, and pornographic websites. In addition, the age attribute has been applied in face verification and retrieval [24] to enhance the overall performance.

Estimating human ages is intrinsically a challenging task due to its *multi-class nature*: An age label can be seen as an individual class. This nature makes age estimation much harder than gender classification and face detection from the perspective of machine learning, where well-studied binary classifiers cannot be directly applied. This nature also makes age estimation easily suffer the over-fitting problem when the size of the training set is insufficient. Moreover, because of the *diversity of facial aging processes* across people, it is very difficult to design and determine the type of facial features that can directly represents human ages. To solve these problems, several age estimation algorithms have been published in the past decade; these algorithms are generally composed of two steps: feature extraction [10,12,14,20,22,25,26,31,35,46,47] and age determination [5,6,18,19,32,40,42–44,51].

Apart from the challenges mentioned above, three important factors should also be regarded for building a robust age estimation system. First, there exist the ordinal relationship and correlations among age labels; for example, age 30 is closer to age 25 than to age 10. This relationship leads age estimation to a more difficult learning task than the traditional multi-class classification problem, which assumes no correlations among classes. Although techniques like regression and cost-sensitive learning could model the ordinal relationship in their objective functions, most of the distance metric learning and dimensionality reduction algorithms are designed only for the traditional classification case and ignore the correlations among labels. Second, as mentioned in [18], the aging process on human faces is rather complicated and may not be fully captured by a single classifier or regressor. Finally in age databases, the number of images of each age label can be highly different, which would result in the serious imbalance problem during the training phase and degrade the overall performance of age estimation.

In this paper, a new facial age estimation approach is proposed, which takes all the above factors and challenges into consideration:

- To avoid over-fitting and explore the connections between facial features and age labels, the locality preserving projection (LPP) [21] is utilized to reduce the dimensionality of features

* Corresponding author. Tel./fax: +886 2 33663662.
E-mail addresses: weilunchao760414@gmail.com (W.-L. Chao), jinzuo2011@gmail.com (J.-Z. Liu), djj@cc.ee.ntu.edu.tw (J.-J. Ding).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.09.011

drastically and preserve the most important information for age estimation.
- To better exploit the ordinal relationship among age labels, a *label-sensitive* concept is introduced, which regards the label similarity during the training phase of LPP.
- To capture the complicated facial aging process, an age-oriented local regression algorithm named *K*NN-SVR (*K* nearest neighbors-support vector regression) is presented.
- To alleviate the imbalance problem, several imbalance treatments are proposed for some steps in our approach.

In addition, to further enhance the performance of LPP, a distance metric adjustment step is introduced (just before LPP) to achieve a suitable space for neighbor searching, an essential operation in LPP. The *label-sensitive* concept could also be employed in this step.

The proposed approach is examined under several experimental settings and evaluation criteria on the most widely-used FG-NET aging database [53], and the simulation results demonstrate the effectiveness and efficiency of our approach. Besides, to further illustrate the availability of the *label-sensitive* concept on dimensionality reduction algorithms other than LPP, we apply this concept to another popular algorithm called the marginal fisher analysis (MFA) [45]; the results also show significant improvements in the age estimation performance.

This paper is organized as follows: In Section 2, we broadly review the previous work on facial age estimation and some related techniques of our approach. In Section 3, an overview of the proposed approach is presented, and the *label-sensitive* concept is introduced in Section 4. In Sections 5 and 6, the proposed algorithms of distance metric adjustment and dimensionality reduction as well as their corresponding imbalance treatments are described; the proposed age-oriented local regression *K*NN-SVR and a short summary of our approach are presented in Sections 7 and 8. Finally, the simulation results and conclusion are given in Sections 9 and 10.

2. Previous and related work

2.1. Previous work on facial age estimation

There were several age estimation algorithms published in the last decade, and the goals of these algorithms can be separated into two categories: One is to estimate the actual age (e.g., 30-year old); the other is to classify a person into an age range, such as baby, teenager, middle-ager, or elder. In this paper, we aim at the first goal, which is the main focus in previous work.

An age estimation algorithm can be simply divided into two steps: *feature extraction* and *age determination*. In the first step, facial features related to human ages or facial appearance change are extracted from human faces for compact representation; in the second step, an age determination function is built to estimate the age based on the extracted features. In the following, we give a broad review of the previous work in each step respectively.

2.1.1. Feature extraction

Among all kinds of facial features, the first one utilized in age estimation is probably the anthropometric model, which is based on the domain knowledge of facial aging processes, such as the occurrence of wrinkles and the change of face shapes. In [25], Kwon et al. exploited the snake algorithm [23] for wrinkle detection, and combined this information with some measures of facial geometry for age range classification. The anthropometric model was also applied in [22,31] for the same task and resulted in acceptable accuracy. However, according to the experiments presented in [13], this model was claimed not suitable for actual age estimation.

The active appearance model (AAM) [7], originally proposed for face detection, was first exploited for age estimation by Lanitis et al. [26] in 2002. AAM can jointly extract the shape and texture variations from human faces – the variations that indeed show some clues of the facial aging process – so it soon became the most widely-used feature type in age estimation [5,6,17–19,27,29,32,40,42–44,48,51]. Later according to AAM, Geng et al. proposed the aging pattern subspace (AGES) [14–16], which further considers the identity information and the ordinal relationship of ages during feature extraction.

To enhance the extraction of local facial information, Yan et al. [46] proposed the spatially flexible patch (SFP), which encodes the age label, the local appearance, and the corresponding spatial coordinate into a SFP vector; a unified Gaussian mixture model (GMM) is then trained to model the variations of all SFP vectors. The SFP method is claimed robust against slight pose change and face misalignment [13], whereas a unified probability model is unlikely to represent all possible local appearances. This problem was later considered and alleviated in [47,50].

Manifold learning has also been applied for feature extraction on human faces. In [10,12], Fu et al. directly performed manifold learning algorithms [11,21] on the resized face images to extract compact facial features. This approach shows plausible performance on the private UIUC-IFP database [10], but is not suitable for the FG-NET database [53], which contains significant pose and expression variations. In our approach, we first extract the AAM features, which could effectively record the pose and expression variations on faces; manifold learning algorithms are then applied for dimensionality reduction and learning the connections between facial features and age labels, not for direct feature extraction from face images.

Besides the feature extraction methods mentioned above, well-known texture descriptors like the Gabor wavelets [20] and the local binary patterns (LBP) [9] have been utilized for age estimation as well.

2.1.2. Age determination

As described in Section 1, age estimation can be seen as a multi-class classification problem, so traditional classification algorithms such as the nearest centroid classifier [26,27,50] and the support vector machine (SVM) [18–20] can be directly applied for age determination. These algorithms, nevertheless, do not take the ordinal relationship and correlations among age labels into account. Namely, they assign in their objective functions a uniform penalty to any misclassification case; however, in age estimation, wrongly predicting a 30-year-old person as 10-year-old is much more serious than predicting him/her as 25-year-old.

Regressors, on the other hand, indigenously consider the ordinal relationship and could probably result in better estimation results. Regression algorithms such as the least square regression [10,12], the kernel regression [42,47], the Gaussian process regression (GPR) [32,51], and the support vector regression (SVR) [18–20] all have been applied and examined in previous work. In [43], Yan et al. proposed an auto-structured regression according to the claim: The age label given to each image should be a nonnegative interval, rather than a fixed value. This regression model was later improved in [44] by utilizing the expectation maximization (EM) algorithm for optimization, leading to better age estimation performance than in [43].

Compared to using a single regressor or classifier for age determination, combining several regressors and classifiers could better capture the complicated aging process and improve the estimation performance [18]. In [26,27], Lanitis et al. proposed the appearance- and age-specific combinations, which first classify a

face into a similar-appearance group or an age group, and then utilize the age determination function built specifically for that group to predict the age. The age-specific combination was also applied in [18,19,29] by concatenating SVM and SVR.

Recently, some new techniques of machine learning have been brought into age estimation. In [5,52], the ordinal regression technique is exploited to model the ordinal relationship among age labels; and in [6], Chang et al. applied the cost-sensitive classification technique, where the correlations among age labels can be flexibly modeled according to different evaluation criteria.

2.2. Related techniques of the proposed approach

2.2.1. Manifold learning

The proposed approach improves LPP [21], a *linearized* manifold learning algorithm, for dimensionality reduction on the extracted facial features, as mentioned in Section 1. Manifold learning has been experimented to successfully discover the intrinsic and latent variables from data samples [28,39]; since the number of these variables is generally much smaller than that of the input features, manifold learning could therefore be used for dimensionality reduction.

Algorithms of manifold learning are generally composed of three steps: The first two steps, *neighbor searching* and *local geometry modeling*, explore certain kinds of nonlinear properties from the input data samples; the third step, *embedding computation*, then embeds these input samples into the output feature space (with the reduced dimensionality) through optimizing the objective function built on such explored properties. In early manifold learning algorithms (such as [4,33,36]), the *embedding computation* step directly generates the output sample for each input sample in a given training set, but cannot handle the testing (or unseen) data. To solve this problem, linearization – a modification that restricts the input–output relationship by a linear projection matrix – is proposed in [21], where LPP is the linearized form of the Laplacian eigenmap [4]. Based on the learned linear matrix, a testing sample can thus be projected into the output space for dimensionality reduction. Linearization was later applied in many manifold learning algorithms [11,45], and in the next subsection, the algorithm of LPP is presented.

2.2.2. Locality preserving projection (LPP)

Given a training set $X=\{\boldsymbol{x}^{(n)}\in R^d\}_{n=1}^N$ with N d-dimensional samples and a desired output dimensionality p, LPP aims to learn a matrix $W_{LPP}\in R^{d\times p}$ that minimizes the average neighbor distance among the projected output samples $Z=\{\boldsymbol{z}^{(n)}\in R^p\}_{n=1}^N$, where $\boldsymbol{z}^{(n)}=W_{LPP}^T\boldsymbol{x}^{(n)}\in R^p$.

The algorithm of LPP comprises the three steps mentioned in Section 2.2.1. In the first step, each sample in X searches its k_1 nearest neighbors via the Euclidean distance. And in the second step, an $N\times N$ matrix $B^+=[b_{ij}^+]_{1\le i,j\le N}$ is constructed to record the feature similarity

$$b_{ij}^+=\exp(-\|\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}\|^2/t) \tag{1}$$

between any pair of neighboring samples for local geometry modeling; if samples $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are not neighbors, $b_{ij}^+=b_{ji}^+=0$. Finally in the third step, the projection matrix W_{LPP} is reached through minimizing the following objective function $E_{LPP}(W)$:

$$E_{LPP}(W)=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^+\times\|W^T(\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)})\|^2=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^+\times\|\boldsymbol{z}^{(i)}-\boldsymbol{z}^{(j)}\|^2, \tag{2}$$

with the constraint $W^TXD^+X^TW=I$ (for preventing trivial solutions), where X is a $d\times N$ matrix containing each training sample in the corresponding column, and D^+ is a diagonal matrix with $d_{ii}^+=\sum_j b_{ij}^+$. To sum up, LPP tries to bring the neighbors (searched in X) as closer as possible in the output p-dimensional space, according to the neighbor similarity B^+. Furthermore, the learned matrix W_{LPP} can be used directly for dimensionality reduction on the testing data.

When the class label $y^{(n)}$ of each training sample $\boldsymbol{x}^{(n)}$ is provided in the training phase, the algorithm of LPP could be adjusted into the supervised version by searching neighbors with only the same class label in the first step [45]; the other parts of the LPP algorithm remained unchanged. Since now the neighbors are defined not only by feature similarity, but also by the same-label constraint, the connection between features and labels can be explicitly linked, hence leading to more representative features for classification after dimensionality reduction. The implementation details of LPP could be traced from the proposed algorithm in Table 3.

3. The proposed age estimation approach

3.1. The problem setting of age estimation

In this section, we give an overview about the proposed age estimation approach. The problem setting used in this paper is defined as follows: Given a training set $\{\boldsymbol{i}^{(n)}\}_{n=1}^N$ with N face images, and its corresponding label set $Y=\{y^{(n)}\in L\}_{n=1}^N$ with $L=\{l_1,\ldots,l_c\}$, building an age estimation system can be modeled as a *supervised machine learning task*, where the symbol c indicates the total number of age labels concerned.

3.2. The overview of the proposed approach

The proposed approach basically adopts the traditional two-step framework of age estimation algorithms (*feature extraction+age determination*), but further include two new steps – ***distance metric adjustment*** and ***dimensionality reduction*** – as illustrated in Fig. 1. There are two main purposes for performing dimensionality reduction on the extracted features before feeding them into the age determination step: One is to alleviate the overfitting problem in training the age determination function; the other is to better explore the connection between facial features and age labels—via the supervised version of dimensionality reduction.

Suggested by the previous work, the active appearance model (AAM) [7], which jointly considers the texture and shape information from human faces, is utilized for feature extraction in our approach and outputs a d-dimensional feature vector $\boldsymbol{x}\in R^d$ for an input image $\boldsymbol{i}$. Then, a distance metric adjustment step is applied to transfer the AAM feature space into a suitable space that can enhance the performance of the following dimensionality reduction step. The resulting feature vectors after distance metric adjustment and dimensionality reduction are denoted as $\boldsymbol{x}_{adjust}\in R^d$ and $\boldsymbol{z}\in R^p$ respectively; $\boldsymbol{x}_{adjust}$ has the same dimensionality as the extracted AAM features and $\boldsymbol{z}$ has a lower dimensionality p. Finally, an age determination function is built to estimate the age label $\hat{y}$ (for the input image $\boldsymbol{i}$) based on the p-dimensional vector $\boldsymbol{z}\in R^p$.

In our approach, we improve the relevant component analysis (RCA) [3] and LPP [21] for distance metric adjustment and dimensionality reduction, respectively; both of them are machine learning algorithms, and each requires a training phase to generate a projection matrix that can be directly applied on the testing data. To our best knowledge, this is the first time to combine supervised distance metric learning and dimensionality

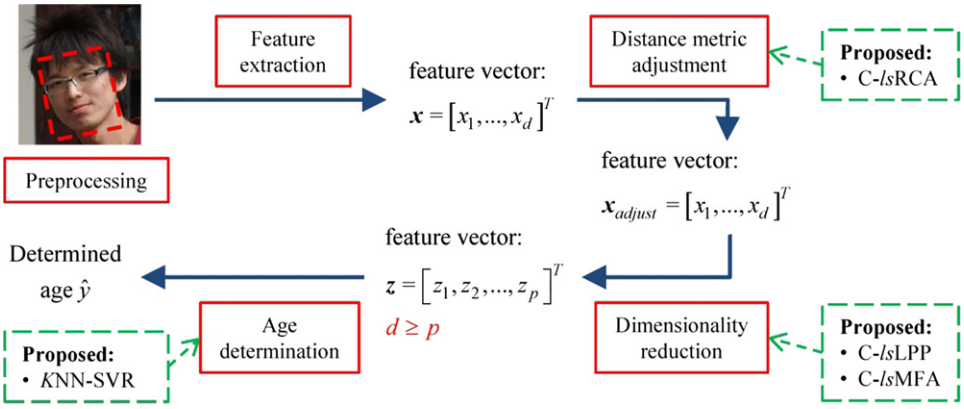

Fig. 1. The flowchart of our approach: The proposed algorithms will be introduced in later sections.

reduction algorithms for facial age estimation; the motivation will be described in Section 5.

3.3. Main concerns and contributions

Aside from the combinations of distance metric learning and dimensionality reduction, we also emphasize the following concerns and propose the corresponding treatments:

- The *label-sensitive* concept is proposed (in Section 4) to take the ordinal relationship among age labels into account in distance metric adjustment and dimensionality reduction.
- As mentioned in [18], the facial aging process is rather complicated. To overcome this problem and also fit the criterion used for evaluating the performance of age estimation, an *age-oriented local regression algorithm*, called *K*NN-SVR, is proposed.
- In age databases, the number of images of each age label can be highly different; if these databases are directly used for training, the resulting age estimation system will be biased by the labels with relatively more samples, leading to the imbalanced learning problem. To alleviate this problem, several *imbalance treatments* are considered in our approach.
- Through incorporating the *label-sensitive* concept and the imbalance treatments, new algorithms for the 2nd and 3rd steps in our approach are presented and marked in Fig. 1.

4. The label-sensitive concept

Before describing in more detail about the proposed approach, we first introduce the *label-sensitive* concept. In traditional multi-class classification, a class is treated independently of the other classes, and a uniform penalty is given when a sample is misclassified into any other class. This condition also occurs in training the conventional supervised dimensionality reduction and distance metric learning algorithms, where a uniform penalty is given to each different-label sample pair and only samples with exactly the same label are to be pulled closer. However, in the task of age estimation, there intrinsically exists the ordinal relationship and correlations among age labels; therefore, the penalties should be varied according to different misclassification cases and different degrees of label dissimilarity.

In the existing literature, there are indeed some algorithms that can handle the ordinal relationship in the age determination step, such as the regression algorithms (as described in Section 2.1.2) and the modified forms of multi-class classification [2,5,6,8,37]. For the distance metric learning and dimensionality reduction steps, unfortunately, few previous algorithms (e.g. [30]) have taken the ordinal relationship into account. To achieve this, the *label-sensitive* concept is proposed in paper, and in Section 9.7, the differences between our algorithms and the one in [30] are compared and discussed.

In the training phases of supervised distance metric learning and dimensionality reduction, several statistical quantities (e.g., the scatter matrix) are required to compute for every single class (age label), as will be mentioned in the next two sections. Instead of treating each class independently with only the same-label samples, the proposed *label-sensitive* concept claims that samples with similar class labels (defined from the ordinal relationship) should also be considered; the weight of each sample in computing the quantities of a specific class is assigned based on the label similarity. For example, when computing the scatter matrix of age 30, samples with ages around 30 are also regarded. In the following two sections, we show how to embed the *label-sensitive* concept into distance metric learning and dimensionality reduction.

5. Distance metric adjustment

5.1. The drawbacks of AAM features and manifold learning algorithms

The shape and texture variations extracted from human faces by AAM may not directly reflect the corresponding age labels: These variations may result from different personalities, poses, genders, races, expressions, and living environments, not just from ages. Besides, the dimensionality of AAM features is usually too high (simply over a hundred) to train a robust age classifier or regressor. To overcome these problems, we apply LPP [21] to explore the connections between facial features and age labels, and drastically reduce the dimensionality of features. LPP, as introduced in Section 2.2.2, is a manifold learning algorithm and composed of three steps in the training phase. In the first step, LPP and most manifold learning algorithms [21,45] assumes the input space to be locally Euclidean, and utilizes the Euclidean metric for neighbor searching; this distance metric, however, may not be suitable in practical cases.

5.2. The reason for performing distance metric adjustment

In age estimation, the desired neighbors of a given sample are samples with *similar facial appearance change caused by human ages*. To judge the availability of the Euclidean metric for

Table 1
The algorithm of RCA (relevant component analysis) [3].

Presetting

- Training set: $X=\{\boldsymbol{x}^{(n)}\in R^d\}_{n=1}^N$, $Y=\{y^{(n)}\in L\}_{n=1}^N$
- Define X_i as the feature set containing all feature samples with age label l_i. The number of samples in $X_i$is denoted as N_i.
- The goal of RCA is to find the projection matrix $W_{RCA}\in R^{d\times d}$; then $\boldsymbol{x}_{adjust}=W_{RCA}^T\boldsymbol{x}\in R^d$.

Algorithm

- For each age label l_i, compute the mean vector $\boldsymbol{\mu}_i=\frac{1}{N_i}\sum_{\boldsymbol{x}^{(n)}\in X_i}\boldsymbol{x}^{(n)}$ and the intra-class scatter matrix $S_i=\frac{1}{N_i}\sum_{\boldsymbol{x}^{(n)}\in X_i}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i)(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i)^T$.
- Compute the total intra-class scatter matrix$S=\frac{1}{N}\sum_{i=1}^{c}N_i\times S_i$.
- Perform the eigendecomposition: $S=V\Lambda V^T$; then $W_{RCA}=V\Lambda^{(-1/2)}$.

Table 2
The proposed *ls*RCA algorithm (without the imbalance treatment).

- Define the sample-to-class weight (from sample n to class i;σ and ε are tunable):

$$e_i^{(n)}=\begin{cases}\exp(-(y^{(n)}-l_i)^2/\sigma) & \text{if}|y^{(n)}-l_i|\le\varepsilon\\ 0 & \text{otherwise}\end{cases}$$

where ε is the *label-sensitive* threshold to define the range of similar labels.
- Modify the computation of $\boldsymbol{\mu}_i$,S_i, and S as:

$$\boldsymbol{\mu}_i=\frac{1}{\omega_i}\sum_{n=1}^{N}e_i^{(n)}\boldsymbol{x}^{(n)},\ S_i=\frac{1}{\omega_i}\sum_{n=1}^{N}e_i^{(n)}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i)(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_i)^T,\ \text{and}\ S=\sum_{i=1}^{c}\omega_iS_i\Big/\sum_{i=1}^{c}\omega_i,$$

where $\omega_i=\sum_{n=1}^{N}e_i^{(n)}$ is the sum of sample-to-class weights from all the samples to the label i.

- Follow the last step of RCA to compute $W_{RCA}\in R^{d\times d}$.

searching neighbors in the AAM feature space, we simply check the standard deviation (STD) of each feature: AAM utilizes the principal component analysis (PCA) [38] for features generation, and each STD value (equivalent to the square root of the eigenvalue in PCA) represents how sharply the corresponding feature varies across different face images. Besides, the usage of PCA also makes the AAM features *uncorrelated*, suggesting that each feature could be independently considered. Namely, the higher the STD is, the stronger influence the feature has in computing the Euclidean distance.

The STDs of the AAM features extracted from the FG-NET database [53] are shown in Fig. 2; as presented, there exists a strong variation among the STD values. However, from both our observations and the experiments in [7,26,27], most of the AAM features with higher STDs are not relevant to human ages, but to poses and expressions, indicating the problem of applying the Euclidean metric for neighbor searching. That is, the searched neighbors are mainly with similar poses or expressions, but not with similar change of appearances caused by human ages. This problem would degrade not only the performance of LPP, but also the overall accuracy of age estimation.

To deal with the above issue, a distance metric adjustment step is introduced in our work (just before LPP) for re-weighing the strengths of features.

5.3. The proposed usage of distance metric learning algorithms

In the existing literature of pattern recognition, normalizing or scaling the STDs of features is the most widely-used method to achieve distance metric adjustment. Nevertheless, without regarding the label information, this method again could not produce a suitable space for searching the desired neighbors via the Euclidean distance. In our approach, we adopt the relevant component analysis (RCA) [3] for distance metric adjustment because of its efficiency and supervised nature: RCA is a supervised distance metric learning algorithm.

Given a training set $X=\{\boldsymbol{x}^{(n)}\in R^d\}_{n=1}^N$ with the label set $Y=\{y^{(n)}\in L\}_{n=1}^N$, RCA first computes the intra-class scatter matrix S_i of each class l_i, and then performs the weighted sum on these matrices to produce the total intra-class scatter matrix S. Finally, the $d\times d$ matrix W that leads to $W^TSW=I$ is selected as the adjustment matrix W_{RCA}; the sample after distance metric adjustment is denoted as $\boldsymbol{x}_{adjust}$, where $\boldsymbol{x}_{adjust}=W_{RCA}^T\boldsymbol{x}\in R^d$.

The objective function of RCA, $W^TSW=I$, is a whitening operation; in other words, RCA aims to whiten the intra-class scatter matrices of all classes simultaneously. The resulting feature space after RCA, though with no theoretical guarantee as the desired space for neighbor searching, does lead to a (nearly) globally Euclidean space inside each age class—through the whitening operation in RCA with the label information. Therefore, when fed with this feature space, the neighbor searching step of supervised LPP (with the same-label constraint) now explores the k_1 nearest neighbors in a globally Euclidean space, which better matches the usage of the Euclidean metric for neighbor searching, hence improving the overall performance (as shown in Section 9). The detailed algorithm of RCA is presented in Table 1; to be noticed, $\boldsymbol{x}_{adjust}$ has the same dimensionality as the input AAM features.

5.4. The proposed label-sensitive relevant component analysis

To further exploit the ordinal relationship among age labels, the *label-sensitive* concept is applied in RCA: When computing the intra-class scatter matrix S_i of class l_i, samples with labels similar to l_i are also involved. The weight of each sample to l_i is modeled through a radial basis function, and the range of similar labels is defined by a *label-sensitive* threshold ε: Two ages with the gap no larger than ε are viewed as similar ages. The resulting *label-sensitive* form of RCA, called *ls*RCA, is summarized in Table 2, where the formulas of the mean vector $\boldsymbol{\mu}_i$, the intra-class scatter S_i, and the total intra-class scatter S defined in Table 1 are modified.

Eventually, to balance the influence of each label in *ls*RCA – the age labels with more samples are likely to dominate the training of *ls*RCA – an imbalance-compensated version called C-*ls*RCA is proposed, where the formula of S in Table 2 is replaced by

$$S=\frac{1}{c}\sum_{i=1}^{c}S_i. \tag{3}$$

This modification equalizes the influence of each intra-class scatter matrix on the total intra-class scatter matrix.

6. Dimensionality reduction

6.1. The proposed label-sensitive locality preserving projection

RCA or its modified versions result in a suitable feature set $\{\boldsymbol{x}_{adjust}^{(n)}=W_{RCA}^T\boldsymbol{x}^{(n)}\in R^d\}_{n=1}^N$, where the Euclidean metric now can be utilized in the neighbor searching step of LPP. As mentioned in Section 2.2.2, LPP is originally an unsupervised technique, yet can be extended into the supervised version by searching

Table 3
The proposed *ls*LPP algorithm (*label-sensitive* LPP).

Presetting

- Training set: $X=\{\boldsymbol{x}_{adjust}^{(n)}\in R^d\}_{n=1}^N$, $Y=\{y^{(n)}\in L\}_{n=1}^N$ (X is represented as a $d\times N$ matrix)
- Define the similar-label set $N^+(i)$ for each sample $\boldsymbol{x}_{adjust}^{(i)}$:

$$N^+(i)=\{\boldsymbol{x}_{adjust}^{(j)}\,\big|\,|y^{(j)}-y^{(i)}|\le\varepsilon, j\ne i\},$$

where ε is the *label-sensitive* threshold to define the range of similar labels.
- Create an $N\times N$ sample similarity matrix $B^+=[b_{ij}^+=0]_{1\le i,j\le N}$.
- The goal of *ls*LPP is to find the projection matrix $W_{LPP}\in R^{d\times p}$; then $\boldsymbol{z}=W_{LPP}^T\boldsymbol{x}_{adjust}\in R^p$.

Algorithm

- For each sample $\boldsymbol{x}_{adjust}^{(i)}$, find the k_1-nearestsamples in $N^+(i)$, and denote these samples as $KNN^+(i)$. The parameter k_1 defines the number of neighboring samples.
- For each sample pair $\{\boldsymbol{x}_{adjust}^{(i)},\boldsymbol{x}_{adjust}^{(j)}\}$, if $\boldsymbol{x}_{adjust}^{(j)}\in KNN^+(i)$or $\boldsymbol{x}_{adjust}^{(i)}\in KNN^+(j)$, set:

$$b_{ij}^+=\exp(-\|\boldsymbol{x}_{adjust}^{(i)}-\boldsymbol{x}_{adjust}^{(j)}\|^2/t)\times\exp(-(y^{(i)}-y^{(j)})^2/\sigma).$$

- Compute $L^+=D^+-B^-$, where D^+ is a diagonal matrix with $d_{ii}^+=\sum_j b_{ij}^+$.
- Solve the generalized eigendecomposition problem:

$$\left(XL^+X^T\right)\boldsymbol{v}^{(i)}=\lambda^{(i)}\left(XD^+X^T\right)\boldsymbol{v}^{(i)}\rightarrow\left(XL^+X^T\right)V=\left(XD^+X^T\right)V\Lambda,$$

Where Λ is arranged in the descending order.
- $W_{LPP}=\left[\boldsymbol{v}^{(N-p+1)},\boldsymbol{v}^{(N-p+2)},\ldots,\boldsymbol{v}^{(N)}\right]=V\left[O_{p\times(N-p)}|I_{p\times p}\right]^T$.

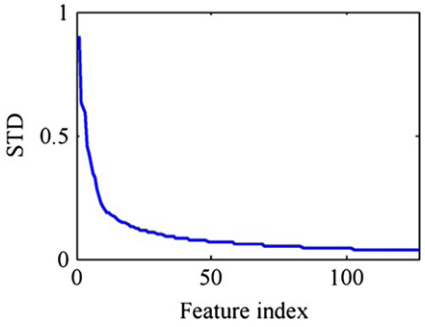

Fig. 2. The STD of each AAM feature (127 features are extracted from the FG-NET database).

neighbors with the same class label. To further take the ordinal relationship among age labels into consideration, the *label-sensitive* concept is applied in LPP to achieve an improved version named *label sensitive* LPP (*ls*LPP), where the same-label constraint is replaced by a similar-label one (according to the *label-sensitive* threshold ε). That is, a training sample now searches the k_1 nearest neighbors with the similar labels.

In addition, *ls*LPP defines a new neighbor weighting function:

$$b_{ij}^+=\exp(-\|\boldsymbol{x}_{adjust}^{(i)}-\boldsymbol{x}_{adjust}^{(j)}\|^2/t)\times\exp(-(y^{(i)}-y^{(j)})^2/\sigma), \tag{4}$$

which regards both the *feature similarity* (former part) and the *label similarity* (later part) between neighbors: The degree of label similarity is again modeled through a radial basis function. In Fig. 3, we show the schematic illustration of *ls*LPP compared to the conventional supervised extension of LPP; the detailed algorithm of *ls*LPP is described in Table 3, where ε, σ, and t are tunable parameters for defining the feature and label similarity. The learned matrix W_{LPP} by *lsLPP* then can be applied to the unseen or testing data.

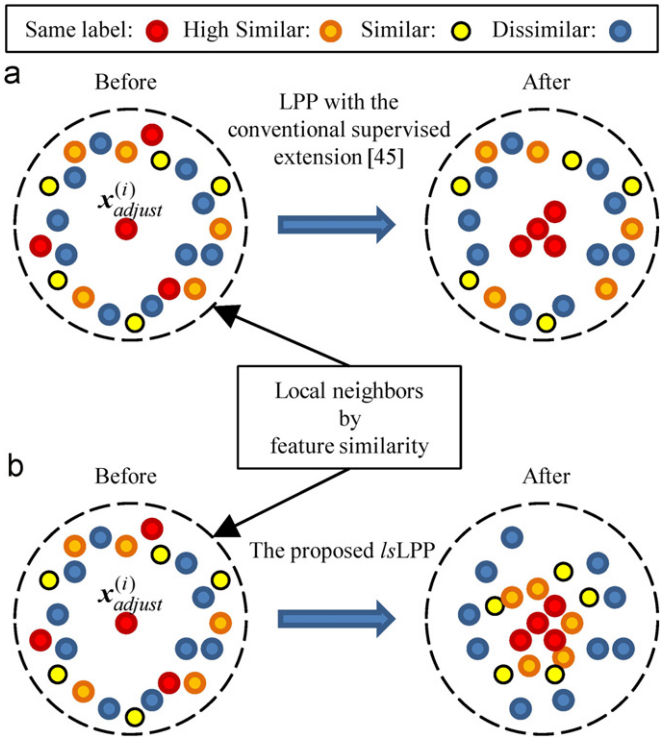

Fig. 3. The schematic illustration of the proposed *ls*LPP compared to the conventional supervised extension of LPP [45]: Here we only take $\boldsymbol{x}_{adjust}^{(i)}$'s neighborhood as an example. In both (a) and (b), the left ones are the distributions of neighboring samples **before** dimensionality reduction; the right ones depict the distributions **after** applying supervised LPP and *ls*LPP. As shown, *ls*LPP can not only pull samples with similar labels together, but also preserve the order of label similarity with respect to $\boldsymbol{x}_{adjust}^{(i)}$.

6.2. The proposed label-sensitive marginal fisher analysis

To illustrate the availability of the *label-sensitive* concept on dimensionality reduction algorithms other than LPP, we also apply this concept to another popular algorithm called the marginal fisher analysis (MFA) [45]. MFA, contrary to LPP, is intrinsically a supervised manifold learning algorithm. In the neighbor searching step, MFA searches the k_1 same-label and k_2 different-label nearest neighbors for each sample via the Euclidean metric. Then in the local geometry modeling step, MFA builds two $N\times N$ matrices, an intrinsic matrix B^+ and a penalty matrix B^-, to record the geometrical information for each type of neighbors, respectively; the entries in B^+ and B^- are defined by (1). Finally in the embedding computation step, MFA seeks the matrix $W_{MFA}\in R^{d\times p}$ that minimizes $E_{MFA}^+(W)$ and maximizes $E_{MFA}^-(W)$, as defined below, simultaneously:

$$E_{MFA}^+(W)=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^+\times\|W^T\left(\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}\right)\|^2=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^+\times\|\boldsymbol{z}^{(i)}-\boldsymbol{z}^{(j)}\|^2, \tag{5}$$

$$E_{MFA}^-(W)=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^-\times\|W^T\left(\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}\right)\|^2=\sum_{i=1}^N\sum_{j=1}^N b_{ij}^-\times\|\boldsymbol{z}^{(i)}-\boldsymbol{z}^{(j)}\|^2. \tag{6}$$

In other words, MFA tries to pull the same-label neighbors closer while push the different-label neighbors away for each sample. The last step of MFA, just like LPP, can be formulated as a generalized eigendecomposition problem.

Following the modifications proposed in Section 6.1 for LPP, we replace the constraints of the same- and different-label neighbors in MFA by the similar- and dissimilar-label neighbors (again based on the label-sensitive threshold ε). Moreover, the degrees of label similarity and label dissimilarity, which are used in computing the modified neighbor weights b_{ij}^{+} and b_{ij}^{-} as in (4), are modeled by a radial basis and a sigmoid function, respectively (shown in Fig. 4). The improved algorithm is called the *label-sensitive* MFA (*ls*MFA) and summarized in Table 4, where ε, σ, γ, η, and t are tunable parameters for defining the feature and label similarity.

6.3. The proposed imbalance treatments for dimensionality reduction

To alleviate the imbalance problems when training *ls*LPP and *ls*MFA, three imbalance treatments are proposed: **sample weight modification (SWM)**, **neighbor size modification (NSM)**, and **neighbor range modification (NRM)**.

The **sample weight modification (SWM)** is to balance the influence of each label in the optimization problems of *ls*LPP and *ls*MFA. Namely, when computing b_{ij}^{+},b_{ij}^{-} in Tables 3 and 4, the label influence of $y^{(i)}$and $y^{(j)}$ should also be considered. For this purpose, an SWM weighting function $f_{SWM}(y)$ is defined, which gives higher weights to age labels with fewer samples. The computation of b_{ij}^{+},b_{ij}^{-} are then modified as

$$\{b_{ij}^{+},b_{ij}^{-}\} := \{b_{ij}^{+},b_{ij}^{-}\} \times f_{SWM}(y^{(i)}) \times f_{SWM}(y^{(j)}). \tag{7}$$

The **neighbor size modification (NSM)** is another way to balance the influence of each label in the optimization problems of *ls*LPP and *ls*MFA. Instead of directly changing the values of b_{ij}^{+} and b_{ij}^{-}, NSM changes the number of neighbors searched for each sample based on the corresponding label: The more neighbors are searched, the higher influence that sample has. In our implementation, an NSM function is defined, which also gives higher weights to labels with fewer samples. The computation of k_1 and k_2 in Tables 3 and 4 are then modified as

$$\{k_1(i),k_2(i)\} := \{k_1,k_2\} \times f_{NSM}(y^{(i)}). \tag{8}$$

In addition to the influence of each label, the quality of the searched neighbors, which define the local geometry to be preserved, is also an important issue of *ls*LPP and *ls*MFA. That is, the searched neighbors should be close (according to the feature similarity) to the target sample to represent the local geometry faithfully. For a sample $\boldsymbol{x}_{adjust}^{(i)}$, if the similar-label set $N^{+}(i)$ defined in Tables 3 and 4 contains sufficient samples, it is more likely to find the k_1 neighbors close to $\boldsymbol{x}_{adjust}^{(i)}$; on contrary, if $N^{+}(i)$ contains insufficient samples, the k_1 searched neighbors may not be close to $\boldsymbol{x}_{adjust}^{(i)}$, resulting in poor local geometry. To compensate this problem, we propose the **neighbor range modification (NRM)**, which sets the *label-sensitive* threshold ε and the parameter σ of the radial basis function in Tables 3 and 4 individually for each sample based on the corresponding label:

$$\{\varepsilon(i),\sigma(i)\} := \{\varepsilon,\sigma\} \times f_{NRM}(y^{(i)}). \tag{9}$$

The weighting function $f_{NRM}(y)$ also gives higher weights to labels with fewer samples, making the size of $N^{+}(i)$ much stable. Fig. 5 illustrates the concept of NRM.

In our implementation, we apply NSM+NRM to *ls*LPP and SWM+NRM to *ls*MFA according to the experimental results. The resulting imbalance-compensated versions are then called C-*ls*LPP and C-*ls*MFA respectively. To be noticed, the matrices B^{+}, B^{-}, L^{+}, and L^{-} defined in Tables 3 and 4 may become asymmetric after these three treatments, disobeying the standard form of LPP and MFA. To deal with this problem, we replace the entries of B^{+} and B^{-} by the entries of $(B^{+}+B^{+T})/2$and$(B^{-}+B^{-T})/2$ before computing D^{+}, D^{-} and L^{+}, L^{-}.

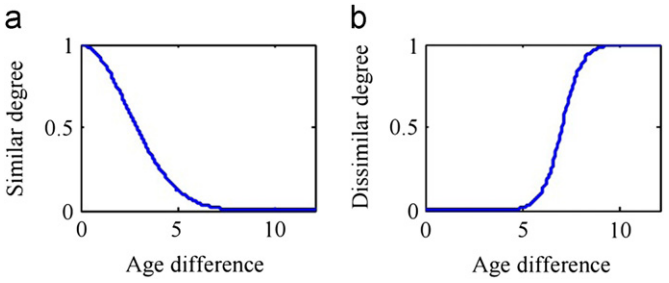

Fig. 4. Illustrations of the degrees of label similarity and label dissimilarity defined in Table 4 for *ls*MFA. (a) The degree of label similarity is modeled by a radial basis function (e.g. with $\sigma=12$). (b) The degree of label dissimilarity is modeled by a sigmoid function (e.g. with $\varepsilon=5$, $\gamma=2$, and $\eta=2$).

7. Age determination

After dimensionality reduction, the resulting p-dimensional feature set $Z=\{\boldsymbol{z}^{(n)}\in R^{p}\}_{n=1}^{N}$ ($z^{(n)}=W_{LPP}^{T}x_{adjust}^{(n)}$ or $W_{MFA}^{T}x_{adjust}^{(n)}$; p is usually much smaller than d) now can be used for training the age determination function. Instead of applying the global regression techniques for age determination, we propose to utilize local regression mainly for two reasons: Local regression can generate comparably sophisticated mapping functions; hence is more likely to capture the complicated facial aging process. Besides, from the perspective of manifold learning [21,45], the output space of LPP, MFA, or their modified versions is locally Euclidean rather than globally Euclidean, meaning that only the local statistical properties are reliable to exploit.

Inspired by the local classification algorithm in [49] and the L_1 loss of support vector regression (SVR) emphasized in [18], an age-oriented local regression algorithm named *K*NN-SVR is proposed: Given a testing sample $\boldsymbol{z}$, *K*NN-SVR first searches the k-nearest neighbors of $\boldsymbol{z}$ in the training set $Z=\{\boldsymbol{z}^{(n)}\in R^{p}\}_{n=1}^{N}$, and trains an RBF-kernel SVR regressor based on these neighboring samples; the learned regressor is then applied to the testing sample $\boldsymbol{z}$ for age estimation. The algorithm of *K*NN-SVR is summarized in Table 5. Notice that, the neighbor searching step in *K*NN-SVR is with no prior distance metric adjustment step.

8. Summarization of the proposed approach

In this section, a short summary of the proposed approach is presented, where the training phase for building the age estimation system and the testing phase for predicting the age of a testing image are described separately.

8.1. The training phase

Given a training set $\{\boldsymbol{i}^{(n)}\}_{n=1}^{N}$ with N face images, our approach first applies AAM to extract the facial features from each image, resulting in a d-dimensional feature set $X=\{\boldsymbol{x}^{(n)}\in R^{d}\}_{n=1}^{N}$. Then, through jointly considering X with its corresponding label set $Y=\{y^{(n)}\in L\}_{n=1}^{N}$, a $d\times d$ matrix W_{RCA} is learned by the proposed C-lsRCA algorithm for distance metric adjustment; the adjusted feature set is denoted as $X=\{\boldsymbol{x}_{adjust}^{(n)}=W_{RCA}^{T}\boldsymbol{x}^{(n)}\in R^{d}\}_{n=1}^{N}$. With this adjusted feature set X and the label set Y, the proposed C-*ls*LPP

Table 4
The proposed *ls*MFA algorithm (*label-sensitive* MFA).

Presetting

- Training set: $X=\{\boldsymbol{x}_{adjust}^{(n)}\in R^d\}_{n=1}^N$, $Y=\{y^{(n)}\in L\}_{n=1}^N$ (X is represented as a $d\times N$ matrix)
- Define the similar-label set $N^+(i)$ and dissimilar-label set $N^-(i)$ for each sample $\boldsymbol{x}_{adjust}^{(i)}$:

$$N^+(i)=\{\boldsymbol{x}_{adjust}^{(j)}\,\big|\,|y^{(j)}-y^{(i)}|\le\varepsilon, j\ne i\}\ \text{ and }\ N^-(i)=\{\boldsymbol{x}_{adjust}^{(j)}\,\big|\,|y^{(j)}-y^{(i)}|>\varepsilon, j\ne i\},$$

  where ε is the *label-sensitive* threshold to define the range of similar labels.
- Create an $N\times N$ intrinsic matrix $B^+=[b_{ij}^+=0]_{1\le i,j\le N}$ and a penalty one $B^-=[b_{ij}^-=0]_{1\le i,j\le N}$.
- The goal of *ls*MFA is to find the projection matrix$W_{MFA}\in R^{d\times p}$, then $\boldsymbol{z}=W_{MFA}^T\boldsymbol{x}_{adjust}\in R^p$.

Algorithm

- For each sample $\boldsymbol{x}_{adjust}^{(i)}$, find the k_1- and k_2-nearestsamples in $N^+(i)$ and $N^-(i)$, and denote these samples as $KNN^+(i)$ and $KNN^-(i)$.
- For each sample pair $\{\boldsymbol{x}_{adjust}^{(i)},\boldsymbol{x}_{adjust}^{(j)}\}$, if $\boldsymbol{x}_{adjust}^{(j)}\in KNN^+(i)$or $\boldsymbol{x}_{adjust}^{(i)}\in KNN^+(j)$, set:

$$b_{ij}^+=\exp(-\|\boldsymbol{x}_{adjust}^{(i)}-\boldsymbol{x}_{adjust}^{(j)}\|^2/t)\times\exp(-(y^{(i)}-y^{(j)})^2/\sigma).$$

- For each sample pair $\{\boldsymbol{x}_{adjust}^{(i)},\boldsymbol{x}_{adjust}^{(j)}\}$, if $\boldsymbol{x}_{adjust}^{(j)}\in KNN^-(i)$or $\boldsymbol{x}_{adjust}^{(i)}\in KNN^-(j)$, set:

$$b_{ij}^-=\exp(-\|\boldsymbol{x}_{adjust}^{(i)}-\boldsymbol{x}_{adjust}^{(j)}\|^2/t)\times\left(1+\exp\left(-\gamma\left(|y^{(i)}-y^{(j)}|-(\varepsilon+\eta)\right)\right)\right)^{-1}.$$

- Compute $L^+=D^+-B^-$ and $L^-=D^--B^-$, where D^+ and D^- are both diagonal matrices with $d_{ii}^+=\sum_j b_{ij}^+$ and $d_{ii}^-=\sum_j b_{ij}^-$.
- Solve the generalized eigendecomposition problem:

$$\left(XL^+X^T\right)\boldsymbol{v}^{(i)}=\lambda^{(i)}\left(XL^-X^T\right)\boldsymbol{v}^{(i)}\rightarrow\left(XL^+X^T\right)V=\left(XL^-X^T\right)V\Lambda,$$

  where Λ is arranged in the descending order.
- $W_{MFA}=\left[\boldsymbol{v}^{(N-p+1)},\boldsymbol{v}^{(N-p+2)},\ldots,\boldsymbol{v}^{(N)}\right]=V\left[O_{p\times(N-p)}\,|\,I_{p\times p}\right]^T$.

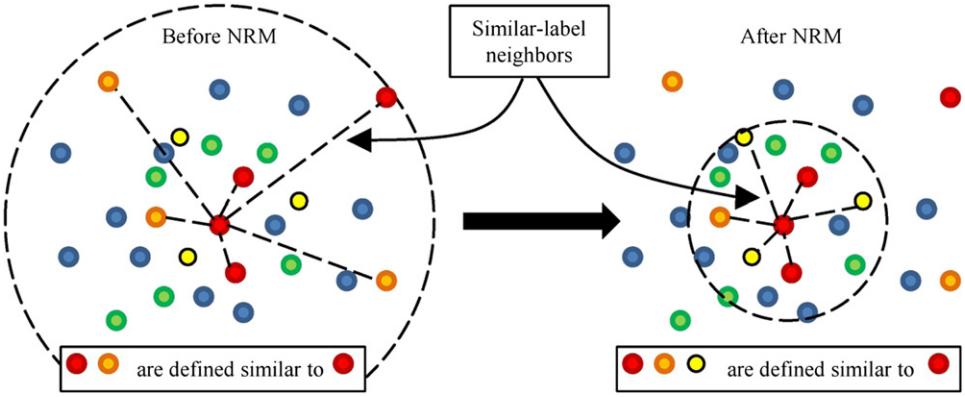

Fig. 5. The schematic illustration of the neighbor range modification (NRM): Before NRM, the 6 nearest samples with similar labels are sparsely distributed; after NRM, the yellow samples are considered similar to the red sample, and the 6 nearest samples can be searched in a much smaller area around the target red sample, resulting in a much faithful local geometry. (For interpretation of the reference to color in this figure, the reader is referred to the web version of this article.)

algorithm (or C-*ls*MFA) is performed to achieve a $d\times p$ matrix W_{LPP} (or W_{MFA}) for dimensionality reduction, leading to the output feature set $Z=\{\boldsymbol{z}^{(n)}=W_{LPP}^T\boldsymbol{x}_{adjust}^{(n)}\in R^p\}_{n=1}^N$ (or $Z=\{\boldsymbol{z}^{(n)}=W_{MFA}^T\boldsymbol{x}_{adjust}^{(n)}\in R^p\}_{n=1}^N$) with dimensionality p.

8.2. The testing phase

Now given a testing image $\boldsymbol{i}$, the AAM features $\boldsymbol{x}$ are again extracted at the first step. Then based on the learned matrices W_{RCA} and W_{LPP} (or W_{MFA}), both the distance metric adjustment and dimensionality reduction steps can be simply accomplished by linear projection, resulting in the p-dimensional vector $\boldsymbol{z}$: $\boldsymbol{z}=W_{LPP}^T(W_{RCA}^T\boldsymbol{x})\in R^p$ or $\boldsymbol{z}=W_{MFA}^T(W_{RCA}^T\boldsymbol{x})\in R^p$. Finally, according to the testing sample $\boldsymbol{z}$ and the feature set Z, which is acquired in the training phase, the proposed *K*NN-SVR is performed to estimate the actual age for the input face image $\boldsymbol{i}$. Notice that, there is also an neighbor searching step in *K*NN-SVR, just like LPP and MFA; however, this neighbor searching step is performed on the p-dimensional data in the testing phase.

9. Simulation results

9.1. Database

The age estimation experiments are performed on the most widely-used FG-NET aging database [53], which contains 1002

face images from 82 individuals and provides 68 landmarks on each face. These images are ranging from age 0 to age 69, but more than 700 of them are under age 20, making the FG-NET aging database highly imbalanced. Fig. 6 shows some example images of the FG-NET database, and the number of images in each age range is listed in Table 10.

9.2. Experimental settings and evaluation criteria

Suggested by the experimental setups in previous work, the leave-one-person-out (LOPO) testing strategy is adopted, where the age estimation algorithm is repeatedly trained on images of 81 people and tested on images of the remaining person. After repeating 82 times, each image will be the testing sample for once and receive an estimated age.

To evaluate the performance, two popular measures, the mean absolute error (MAE) and the cumulative score (CS), proposed in [14] are computed in our experiments. MAE stands for the average L_1 loss during testing, *which fits the loss function of SVR*, and that is why we adopt SVR in the proposed local regression algorithm; CS calculates the percentage of images with L_1 losses lower than a given threshold. The formulations of MAE and CS are defined in Table 6.

9.3. Implementation details of the proposed approach

The proposed age estimation approach is composed of four steps as illustrated in Fig. 1. In the feature extraction step, the 68 landmark points provided on each FG-NET face image are used for AAM training (implemented by the AAM-API tool [34]). Under the memory limitation during our implementation, all face images are first down-sampled, and totally 127 features are extracted to represent each image.

In the subsequent three steps, besides the proposed algorithms, other existing algorithms (e.g., SVR for age determination) could also be applied; therefore, we perform different algorithm combinations to demonstrate the improvements achieved by the usage of RCA and the proposed C-*ls*RCA, C-*ls*LPP, C-*ls*MFA, and *K*NN-SVR algorithms. The tunable parameters of the proposed algorithms, such as the *label-sensitive* threshold ε and the number of neighbors, are selected via cross validation. The three functions for imbalance treatments proposed in Section 6.3 are simply defined in Table 7 to balance the influence of each age label.

The LOPO MAE results of different algorithm combinations in our four-step framework are listed in Table 8. As presented, the usage of RCA outperforms the widely-used standard deviation normalization (STDN) for distance metric adjustment; the improved version C-*ls*RCA further reduces the MAEs in most of the combinations. The proposed *ls*LPP and *ls*MFA with the *label-sensitive* concept obviously surpass their original versions, and the imbalance treatments (NSM+NRM for *ls*LPP and SWM+NRM for *ls*MFA) improve the performance in several combinations. Finally, the proposed *K*NN-SVR definitely outperforms SVR in all the cases, demonstrating the usage of local regression after performing manifold learning algorithms. To sum up, the combinations with our propositions and modifications – including distance metric adjustment, local regression, the imbalance treatments, and the *label-sensitive* concept – achieve the lowest age estimation errors against other algorithm combinations.

Furthermore, the optimal output dimensionality p of C-*ls*LPP, C-*ls*MFA and the parameter k of *K*NN-SVR are reached around 10 and 15 (as shown in Fig. 7), illustrating the effectiveness of dimensionality reduction and local statistics in age estimation: Fig. 7 shows the MAEs achieved by the two best combinations in Table 8 – C-*ls*RCA+C-*ls*LPP+*K*NN-SVR and C-*ls*RCA+C-*ls*MFA+*K*NN-SVR – with different choices of dimensionality p and the parameter k of *K*NN-SVR.

9.4. Performance comparisons with existing algorithms

We further compare the two best combinations, C-*ls*RCA+C-*ls*LPP+*K*NN-SVR and C-*ls*RCA+C-*ls*MFA+*K*NN-SVR, with other existing age estimation algorithms; the LOPO MAE results of the existing and the proposed algorithms are listed in Table 9, along with brief algorithm descriptions. As presented, both the two proposed combinations reach the lowest MAEs, even better than

Table 6
Definitions of MAE and CS ($\hat{y}$: the predicted age).

Testing set	$D_{test}: \{\mathbf{f}^{(n)}\}_{n=1}^{N_t}, Y=\{y^{(n)}\in L\}_{n=1}^{N_t}$
MAE	$MAE=\frac{1}{N_t}\sum_{n=1}^{N_t}\left\lvert\hat{y}^{(n)}-y^{(n)}\right\rvert$
CS	$CS(j)=\frac{1}{N_t}\sum_{n=1}^{N_t}\mathbf{1}\left[\left\lvert\hat{y}^{(n)}-y^{(n)}\right\rvert\le j\right]$, where $\begin{cases}\mathbf{1}[true]=1\\ \mathbf{1}[false]=0\end{cases}$

Table 5
The proposed *K*NN-SVR algorithm.

Presetting
- Training set: $Z=\{\mathbf{z}^{(n)}\in R^p\}_{n=1}^{N}$, $Y=\{y^{(n)}\in L\}_{n=1}^{N}$

Algorithm: (k is tunable)
- Given an input query $\mathbf{z}$, find its k-nearest Euclidean neighbors $\{\mathbf{z}_{KNN}^{(i)}, y_{KNN}^{(i)}\}_{n=1}^{k}$ in Z.
- Train an RBF-kernel SVR regressor based on $\{\mathbf{z}_{KNN}^{(i)}, y_{KNN}^{(i)}\}_{n=1}^{k}$ to predict the age of $\mathbf{z}$.

Fig. 6. Example images of the FG-NET aging database: Each row shows part of the face images of one subject from younger to older.

Table 7
Definitions of the three imbalance treatments proposed in Section 6.3.

SWM	NSM	NRM
$f_{SWM}(y)=\begin{cases}1, \text{for } y\le 20\\ 1+\frac{(y-20)}{50}, \text{for } y>20\end{cases}$	$f_{NSM}(y)=\begin{cases}1.0, \text{for } y\le 20\\ 1.3, \text{for } 20<y\le 30\\ 1.8, \text{for } 30<y\le 40\\ 2.5, \text{for } y>40\end{cases}$	$f_{NRM}(y)=\begin{cases}1.0, \text{for } y\le 20\\ 1.2, \text{for } 20<y\le 30\\ 1.6, \text{for } 30<y\le 40\\ 2.0, \text{for } y>40\end{cases}$

Table 8
Comparisons of the LOPO MAEs with different algorithm combinations in our four-step framework. For the distance metric adjustment step, four different algorithms are considered: No change, STDN, RCA, and C-IsRCA. For the dimensionality reduction step, six algorithms are compared: LPP, *Is*LPP, C-*Is*LPP, MFA, *Is*MFA, and C-*Is*MFA. For the age determination step, SVR and *K*NN-SVR are used and compared.

	sLPP [a]	***Is*LPP**	**C-*Is*LPP**	MFA	***Is*MFA**	**C-*Is*MFA**
SVR						
No change [b]	7.35	6.95	7.28	8.07	7.62	7.58
STDN [c]	5.79	5.82	5.85	5.62	5.71	5.84
RCA	5.43	5.38	5.44	5.49	5.43	5.48
C-*Is*RCA	5.55	5.32	5.34	5.89	5.46	5.37
***K*NN-SVR**						
No change	4.84	4.85	4.81	5.06	4.61	4.68
STDN	5.10	4.95	4.75	5.07	4.97	5.12
RCA	4.67	4.53	4.49	4.95	4.60	4.62
C-*Is*RCA	4.74	4.43	**4.38**	5.40	4.67	**4.44**

[a] sLPP: supervised LPP.
[b] No change: no distance metric adjustment.
[c] STDN: STD normalization.

Table 9
Comparisons of the LOPO MAE results on the FG-NET database.

Category	**Algorithm name**	**MAE**	**Algorithm description**
AAM+Classifiers/Regressors	*K*NN	8.24	AAM+*K* nearest neighbors
	SVM	7.25	AAM+Support vector machine
	MLP	6.95	AAM+Multi-layer perceptron
	*K*NN Regression	6.44	AAM+*K*NN Regression
	RUN1 [43]	5.78	AAM+RUN1[a]
	Gaussian IIS-LLD [17]	5.77	AAM+Learning from label distribution
	SVR	5.68	AAM+Support vector regression
	RUN2 [44]	5.33	AAM+RUN2
	RED-SVM [5]	5.24	AAM+RED-SVM
	MHR [6]	4.87	AAM+Multiple hyperplanes ranker
	OHR [6]	4.48	AAM +Ordinal hyperplanes ranker
AAM+Hybrid combination	WAS [26]	8.06	AAM+Weighted appearance-specific
	LARR [18]	5.07	AAM+LARR[b]
	PFA [19]	4.97	AAM+Probabilistic fusion approach
AGES	AGES [14]	6.77	AAM+Aging pattern subspace
	AGES with LDA [14]	6.22	AAM+AGES with LDA[c]
	KAGES [15]	6.18	AAM+Kernel AGES
	MSA [16]	5.36	AAM+Multi-linear subspace analysis
Manifold learning	LEA [48]	7.65	AAM+Locally embedded analysis
	SSE [48]	5.21	AAM+SSE[d]
Distance metric learning	mGPR [32]	5.08	AAM+DML[e]+GPR[f]
	m*K*NN [42]	4.93	AAM+DML+*K*NN Regression
Gaussian process regression	WGP [51]	4.95	AAM+Warped GPR
	MTWGP [51]	4.83	AAM+Multi-task warped GPR
Other features	BIF [20]	4.77	BIF[g]+PCA+Support vector regression
	RPK [47]	4.95	SFP[h]+Patch kernel+Kernel regression
Proposed	**C-*Is*RCA+C-*Is*MFA (SWM+NRM)**	**4.44**	**AAM+C-*Is*RCA+C-*Is*MFA+*K*NN-SVR**
	C-*Is*RCA+C-*Is*LPP (NSM+NRM)	**4.38**	**AAM+C-*Is*RCA+C-*Is*LPP+*K*NN-SVR**

[a] RUN: regression with uncertain nonnegative label.
[b] LARR: locally adjusted robust regression.
[c] LDA: linear discriminant analysis.
[d] SSE: synchronized submanifold embedding.
[e] DML: distance metric learning.
[f] GPR: Gaussian process regression.
[g] BIF: biologically inspired features.
[h] SFP: Spatially flexible patch.

Table 10
Comparisons of the LOPO MAEs at different age ranges on the FG-Net database.

Range	# images	LEA [48]	MLP	QF	*K*NN Reg. [a]	RUN1 [43]	SSE [48]
0–9	371	3.89	5.25	5.67	4.76	2.51	2.06
10–19	339	4.85	5.24	5.54	3.43	3.76	3.26
20–29	144	8.67	5.85	5.92	5.46	6.38	6.03
30–39	70	13.02	11.29	10.27	13.51	12.51	9.53
40–49	46	19.46	16.48	12.24	22.2	20.09	11.17
50–59	15	26.13	28.80	18.60	31.17	28.07	16.00
60–69	8	39.00	39.50	28.00	43.47	42.50	26.88
Average	1002	7.65	6.95	6.70	6.44	5.78	5.21
Range	# images	mGPR[32]	RPK[47]	m*K*NN[42]	BIF[20]	**C-*ls*MFA**	**C-*ls*LPP**
0–9	371	2.99	2.3	2.29	2.99	**1.965**	**1.911**
10–19	339	4.19	4.86	3.65	3.39	**3.7021**	**3.5264**
20–29	144	5.34	4.02	5.44	4.30	**5.5278**	**5.326**
30–39	70	9.28	7.32	10.55	8.24	**10.06**	**10.67**
40–49	46	13.52	15.24	15.81	14.98	**10.80**	**10.11**
50–59	15	17.79	22.2	25.18	20.49	**14.53**	**15.07**
60–69	8	22.68	33.15	36.80	31.62	**22.25**	**23.37**
Average	1002	5.08	4.95	4.93	4.77	**4.44**	**4.38**

[a] *K*NN Reg.: *K*NN regression.

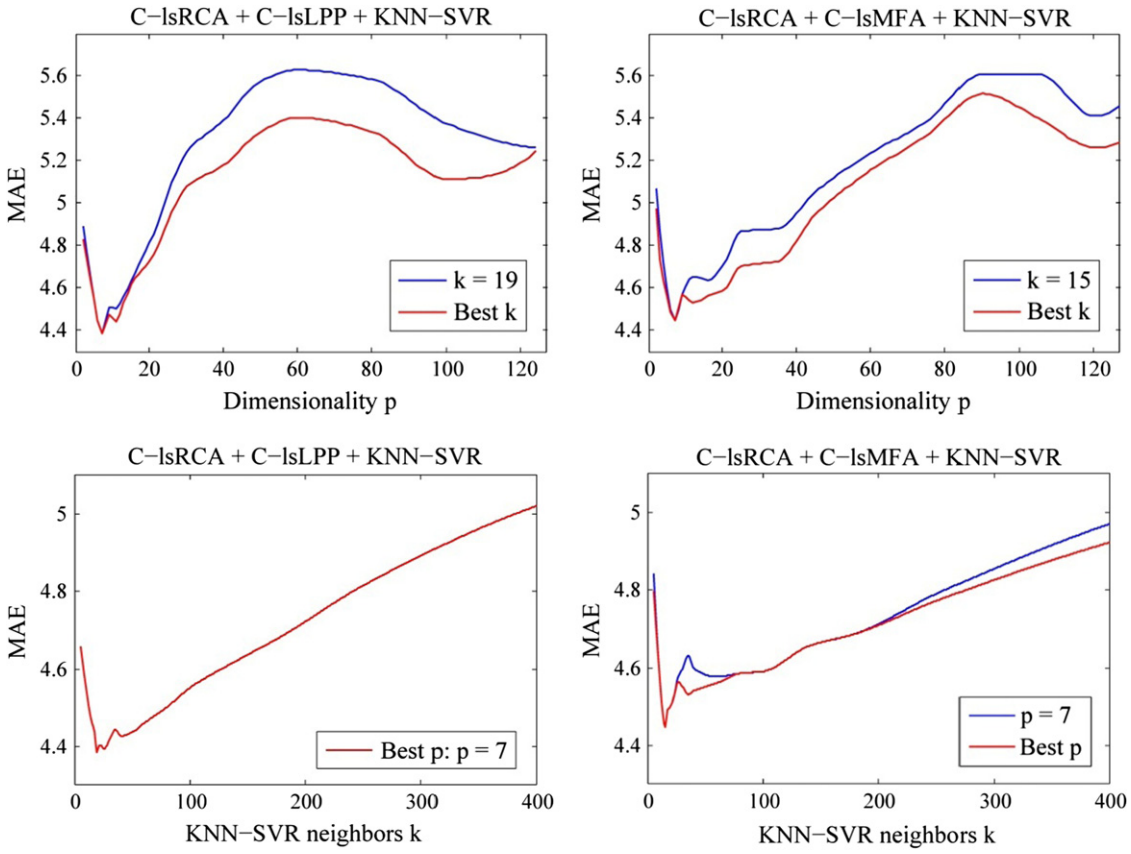

Fig. 7. The LOPO MAEs achieved by C-*ls*RCA+C-*ls*LPP+*K*NN-SVR and C-*ls*RCA+C-*ls*MFA+*K*NN-SVR with different choices of dimensionality p and the parameter k of *K*NN-SVR. The first row shows the MAEs achieved by varied p, with either a fixed k or the best k under the corresponding p. The second row shows the MAEs achieved by varied d, with either a fixed p or the best p under the corresponding k. From our observation, the best p under each varied k in the combination C-*ls*RCA+C-*ls*LPP+*K*NN-SVR is always 7.

the state-of-art ordinal hyperplanes ranker (OHR) proposed by Chang et al. [6]. Also from Table 9, we found that our approach, combining distance metric learning and dimensionality reduction, definitely outperform the work in [32,42,48], which only applies one of these two techniques, supporting the usage of the four-step framework in our approach. To further illustrate the performance improvement achieved by the proposed imbalance treatments, the LOPO MAEs at different age ranges are computed and listed in Table 10; both the proposed combinations reach significantly lower MAEs at higher ages.

Finally in Fig. 8, we show the comparison of the cumulative score (CS), where totally 11 thresholds are examined. As presented, the proposed approaches achieve the highest CSs at each threshold against other algorithms.

9.5. Performance comparisons under other experimental settings

In addition to the LOPO, there are still other experimental settings adopted in previous work. In [9], only the images under age 30 (totally 873 images) are utilized for training and testing with the LOPO evaluation; in [35,52], the whole FG-NET database are randomly divided into 4 folders, and the 4-fold cross validation is performed for the MAE evaluation. This cross validation setting disregards the identity information, where a single person may have images in the training (base) and testing (validation) sets simultaneously.

To justify the effectiveness of the proposed approach, we also perform these two settings for comparison. The MAE results are listed in Tables 11 and 12, and our approaches still achieve the lowest MAEs in both settings. Furthermore, the cross validation case is observed to result in lower MAEs than the LOPO case with the whole FG-NET database, revealing that if images of one person are included in both the training and testing sets, the accuracy of age estimation could be further improved. Besides, this observation also tells that under different experimental settings, an age estimation algorithm would produce different estimation results.

9.6. Further discussions and illustrations

The proposed approach can be efficiently trained and tested. It takes only 6 s for training the C-*ls*RCA and C-*ls*LPP/ C-*ls*MFA matrices by using Matlab on a duo-core PC. Although *K*NN-SVR is an on-line algorithm, in the testing phase it requires only 0.003 s for searching neighbors in the low-dimensional space after dimensionality reduction and training the SVR regressor with only the k nearest neighbors: The training complexity of the dual-form SVR is generally from $O(m)$ to $O(m^2)$, where m is the number of training samples. Compared to other age estimation algorithms, such as the ones requiring nearly a half minute for training the SVR regressor or the multi-class SVM classifier, the computational complexity of our approach seems acceptable. Even with a larger training set, the efficiency of the neighbor searching step in *K*NN-SVR could still be kept by incorporating the indexing techniques [1,41].

Besides, to demonstrate the effectiveness of the proposed dimensionality reduction algorithms on learning the feature-label connection, we depict in Fig. 9 the distributions of the first two C-*ls*LPP and C-*ls*MFA features from the whole FG-NET database. As shown, obvious feature-label dependences have been reached by the proposed algorithms. This result also provides support to the use of local regression: Neighboring samples in the feature space after performing C-*ls*RCA+C-*ls*LPP or C-*ls*RCA+C-*ls*MFA are with similar age labels.

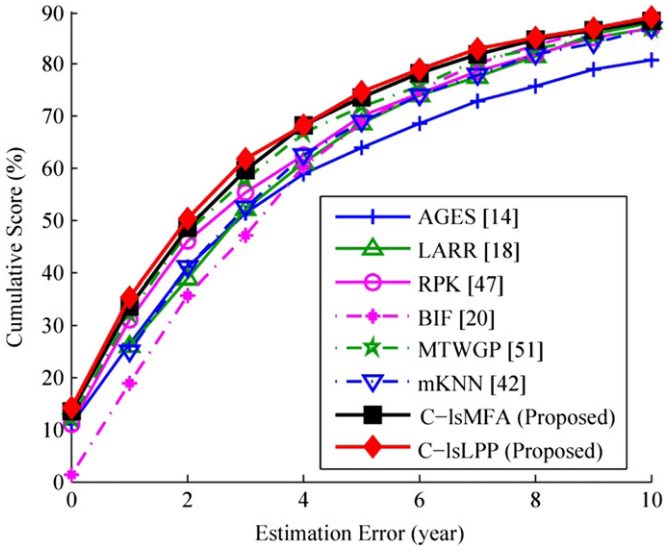

Fig. 8. The comparison of LOPO CS (cumulative score) results on the FG-NET database.

Table 11
Comparisons of the LOPO MAE results on the FG-NET database (with ages smaller than 30).

Algorithm name	MAE	Algorithm description
LBP+OLPP+MLP [9]	4.28	LBP [a]+OLPP[b]+MLP[c]
LBP+OLPP+QF [9]	3.65	LBP+OLPP+Quadratic function
C-*ls*RCA+*ls*MFA (SWM+NRM)	**3.10**	**AAM+C-*ls*RCA+C-*ls*MFA+*K*NN-SVR**
C-*ls*RCA+*ls*LPP (NSM+NRM)	**3.06**	**AAM+C-*ls*RCA+C-*ls*LPP+*K*NN-SVR**

[a] LBP: local binary pattern.
[b] OLPP: orthogonal LPP.
[c] MLP: multi-layer perceptron.

Table 12
Comparisons of the 4-fold cross validation MAE results on the FG-NET database.

Algorithm name	MAE	Algorithm description
HFM [35]	5.97	Hierarchical face model+MLP
RankBoost [52]	5.67	Haar-like features+RankBoost+SVR
C-*ls*RCA+*ls*MFA (SWM+NRM)	**4.23**	**AAM+C-*ls*RCA+C-*ls*MFA+*K*NN-SVR**
C-*ls*RCA+*ls*LPP (NSM+NRM)	**4.11**	**AAM+C-*ls*RCA+C-*ls*LPP+*K*NN-SVR**

Furthermore, to justify the improvement (on the neighbor searching step in manifold learning) achieved by distance metric adjustment, we show in Fig. 10 the searched neighbors (via the Euclidean metric) of an arbitrary target sample, in either the original AAM feature space or the feature space adjusted by C-*ls*RCA, and with or without the similar-label constraint in *ls*LPP. On the condition with no similar-label constraint, the searched neighbors in the AAM feature space (though with similar expressions and poses) are with large age gaps to the target sample; on contrary, the searched neighbors in the adjusted space are of close ages to the target sample and with similar appearance change caused by human ages. Even with the similar-label constraint (e.g., *label-sensitive* threshold $\varepsilon=3$), the searched neighbors in the AAM feature space are still with similar poses and expressions; the influence of these factors can be much suppresses in the adjusted space, demonstrating the effectiveness of C-*ls*RCA on reaching a suitable space for neighbor searching in LPP (MFA), *ls*LPP (*ls*MFA), and C-*ls*LPP (C-*ls*MFA).

9.7. Algorithm comparisons with the class distance based discriminant analysis [30]

The class distance based discriminant analysis (CDDA) presented by Ma et al. [30] also considers the relationship among labels in supervised dimensionality reduction; however, there are some fundamental differences between CDDA and the proposed C-*ls*LPP and C-*ls*MFA. First, though the objective functions of CDDA can be written in the forms of (5) and (6), it is not a manifold algorithm indeed: The *neighbor searching step* for setting the entries of B^+ and B^- (in Tables 3 and 4) are not included in CDDA. Second, any sample pair in CDDA could receive weight in B^+ without considering the *similar-label constraint*. That is, a sample pair with a large label difference (e.g. over 20 years) could still get weight in B^+. Third, the purpose for building the penalty matrix B^- in CDDA is contrary to the one in C-*ls*MFA: According to the claim that similar labels are easily to be

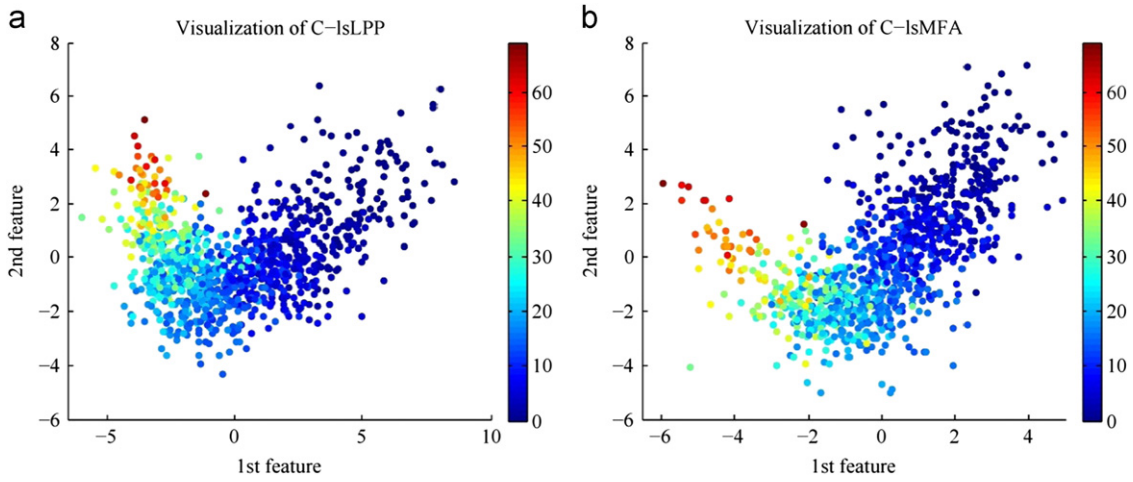

Fig. 9. The dimensionality reduction results after C-*ls*RCA+C-*ls*LPP/C-*ls*MFA based on the whole FG-NET database: The distributions of the 1st and 2nd dimensions (features) are shown for visualization.

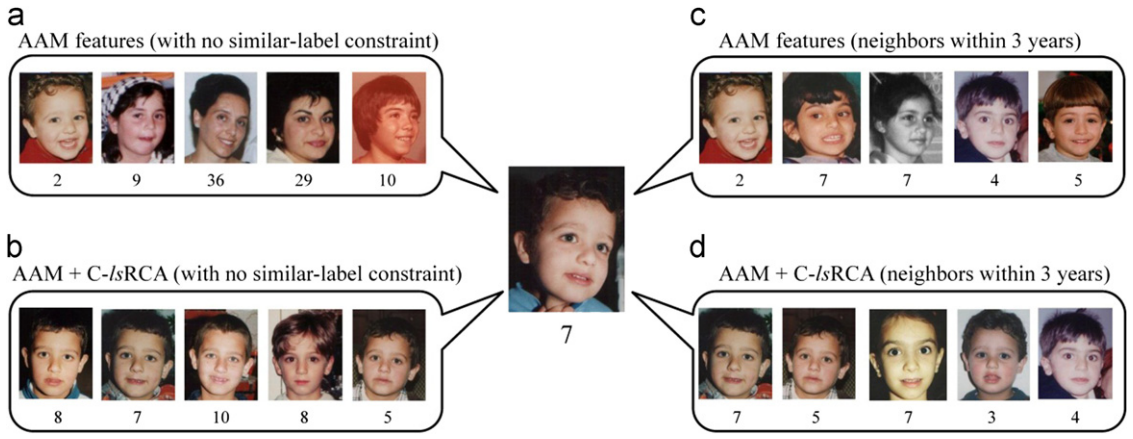

Fig. 10. The influence of distance metric adjustment on the neighbor searching step in manifold learning algorithms: (a), (c) show the 5 nearest neighbors searched in the AAM feature space, either with or without the similar-label constraint; (b), (d) show the 5 nearest neighbors searched in the C-*ls*RCA adjusted space. As presented, the neighbor searching process in the adjusted space is affected less by expressions and poses; therefore, the searched neighbors are with more similar ages and facial aging processes to the target sample than the ones in the AAM feature space.

misclassified, CDDA gives a larger penalty b_{ij}^{-} to the sample pair with a smaller label difference (the largest penalty is given to sample pairs with $|y^{(i)}-y^{(j)}|=1$). Our *label-sensitive* constraint, on the other hand, aims to pull the similar-label neighbors closer while push the dissimilar-label neighbors away for each sample; therefore, the smaller the label difference is, the smaller the penalty is assigned. In more detail, CDDA simultaneously gives the similar-label sample pair a large b_{ij}^{+} and b_{ij}^{-}, whereas C-*ls*MFA assign a small (even no) b_{ij}^{-} to this pair. Finally, the distance metric adjustment step and the potential imbalance problem are not considered in CDDA. Since the age estimation experiments in [30] are not performed on the FG-NET database, we implement CDDA and substitute it for the proposed C-*ls*LPP (C-*ls*MFA) in our four-step framework; the resulting LOPO MAE is 5.80, demonstrating the effectiveness of our *label-sensitive* concept on exploiting the label information.

10. Conclusion

In this paper, a new age estimation approach considering the intrinsic factors of human ages is proposed. After feature extraction, RCA is utilized to achieve a suitable space for neighbor searching. Then based on this adjusted space, LPP and MFA are trained to drastically reduce the feature dimensionality and learn the connections between features and age labels. To further consider the ordinal relationship of human ages as well as the imbalanced learning problem in RCA, LPP, and MFA, the *label-sensitive* concept and several imbalance treatments are proposed, resulting in new algorithms called C-*ls*RCA, C-*ls*LPP, and C-*ls*MFA. In addition, an age-oriented local regression algorithm called *K*NN-SVR is presented to capture the complicated facial aging process for age determination. From the simulation results performed on the most widely-used FG-NET aging database, the proposed approach achieves the lowest MAE against the state-of-art algorithms under several experimental settings.

References

[1] A. Andoni, P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, Communications of the ACM 51 (1) (2008) 117–122.
[2] J. Avigad, R. Sommer, A model-theoretic approach to ordinal analysis, The Bulletin of Symbolic Logic 3 (1997) 17–52.
[3] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning distance functions using equivalence relations, in: Proceedings of the International Conference on Machine Learning (ICML), 2003, pp. 11–18.
[4] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2001, pp. 585–591.

[5] K.Y. Chang, C.S. Chen, Y.P. Hung, A ranking approach for human age estimation based on face images, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2010, pp. 3396–3399.
[6] K.Y. Chang, C.S. Chen, Y.P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2001, pp. 585–592.
[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, in: Proceedings of the European Conference on Computer Vision (ECCV), vol. 2, 1998, pp. 484–498.
[8] J. Coste, E. Walter, D. Wasserman, A. Venot, Optimal discriminant analysis for ordinal responses, Statistics in Medicine 16 (5) (1997) 561–569.
[9] H. Fang, P. Grant, M. Chen, Discriminant feature manifold for facial aging estimation, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2010, pp. 593–596.
[10] Y. Fu, Y. Xu, T.S. Huang, Estimating human ages by manifold analysis of face pictures and regression on aging features, in: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME), 2007, pp. 1383–1386.
[11] Y. Fu, M. Liu, T.S. Huang, Conformal embedding analysis with local graph modeling on the unit hypersphere, in: Proceedings of the Computer Vision and Pattern Recognition Workshop (CVPRW), 2007.
[12] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Transactions on Multimedia (TMM) 10 (4) (2008) 578–584.
[13] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32 (11) (2010) 1955–1976.
[14] X. Geng, Z.H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (12) (2007) 2234–2240.
[15] X. Geng, K. Smith-Miles, Z.H. Zhou, Facial age estimation by nonlinear aging pattern subspace, in: Proceedings of the ACM Multimedia (ACM-MM), 2008, pp. 721–724.
[16] X. Geng, K. Smith-Miles, Facial age estimation by multilinear subspace analysis, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 865–868.
[17] X. Geng, K. Smith-Miles, Z.H. Zhou, Facial age estimation by learning from labeled distributions, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2010, pp. 451–456.
[18] G. Guo, Y. Fu, C. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Transactions on Image Processing (TIP) 17 (7) (2008) 1178–1188.
[19] G. Guo, Y. Fu, T.S. Huang, C. Dyer, A probabilistic fusion approach to human age prediction, in: Proceedings of the Computer Vision and Pattern Recognition Workshop (CVPRW), 2008.
[20] G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio inspired features, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2009, pp. 112–119.
[21] X. He, P. Niyogi, Locality preserving projections, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2003, pp. 153–160.
[22] W.B. Horng, C.P. Lee, C.W. Chen, Classification of age groups based on facial features, Tamkang Journal of Science and Engineering 4 (3) (2001) 183–192.
[23] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, International Journal of Computer Vision (IJCV) 1 (4) (1988) 321–331.
[24] N. Kumar, A. Berg, P.N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 33 (10) (2011) 1962–1977.
[25] Y. Kwon, N. Lobo, Age classification from facial images, Computer Vision and Image Understanding 74 (1) (1999) 1–21.
[26] A. Lanitis, C. Taylor, T. Cootes, Toward automatic simulation of aging effects on face images, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24 (4) (2002) 442–455.
[27] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, IEEE Transactions on Systems, Man, and Cybernetics, Part B 34 (1) (2004) 621–628.
[28] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, first ed., Springer, New York, 2007.
[29] K. Luu, K. Ricanek, T.D. Bui, C.Y. Suen, Age estimation using active appearance models and support vector machine regression, in: Proceedings of the IEEE International Conference on Biometrics, 2009, pp. 314–318.
[30] B. Ma, S. Shan, X. Chen, W. Gao, Discriminant analysis for perceptionally comparable classes, in: Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 2008.
[31] A. Montillo, H. Ling, Age regression from faces using random forests, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2009, pp. 2465–2468.
[32] L. Pan, Human age estimation by metric learning for regression problems, Energy Minimization Methods in Computer Vision and Pattern Recognition (EMM-CVPR) (2009) 455–465.
[33] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[34] M.B. Stegmann, The AAM-API: an open source active appearance model implementation, in: Proceedings of the International Conference on Medical image computing and computer-assisted intervention—MICCAI, 2003, pp. 951–952.
[35] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, W. Gao, Design sparse features for age estimation using hierarchical face model, in: Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–6.
[36] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[37] Y. Toren, Ordinal risk-group classification, arXiv: 1012.5487v4, 2011.
[38] M.A Turk, A.P. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[39] L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality reduction: a comparative review, Technical Report, Tilburg University, 2009.
[40] C. Wang, Y. Su, C. Hsu, C. Lin, H. Liao, Bayesian age estimation on face images, in: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME), 2009, pp. 282–285.
[41] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2008.
[42] B. Xiao, X. Yang, Y. Xu, Learning distance metric for regression by semidefinite programming with application to human age estimation, in: Proceedings of the ACM Multimedia (ACM-MM), 2009, pp. 451–460.
[43] S. Yan, H. Wang, X. Tang, T.S. Huang, Learning auto-structured regressor from uncertain nonnegative labels, in: Proceedings of the International Conference on Computer Vision (ICCV), 2007, pp. 1–8.
[44] S. Yan, H. Wang, T.S. Huang, X. Tang, Ranking with uncertain labels, in: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME), 2007, pp. 96–99.
[45] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extension: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (1) (2007) 40–51.
[46] S. Yan, M. Liu, T.S. Huang, Extracting age information from local spatially flexible patches, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008, pp. 737–740.
[47] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, T.S. Huang, Regression from patch-kernel, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[48] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, T.S. Huang, Synchronized submanifold embedding for person independent pose estimation and beyond, IEEE Transactions on Image Processing (TIP) 18 (1) (2009) 202–210.
[49] H. Zhang, A. Berg, M. Maire, J. Malik, SVM-KNN: discriminative nearest neighbor classification for visual category recognition, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2006, pp. 2126–2136.
[50] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, T.S. Huang, Face age estimation using patch-based Hidden Markov Model supervectors, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2008, pp. 1–4.
[51] Y. Zhang, D. Yeung, Multi-task warped Gaussian process for personalized age estimation, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2622–2629.
[52] P. Yang, L. Zhong, D. Metaxas, Ranking model for facial age estimation, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2010, pp. 3404–3407.
[53] The FG-NET Aging Database, 〈http://www.fgnet.rsunit.com/〉.

Wei-Lun Chao received the M.S. degree from the Graduate Institute of Communication Engineering (GICE), National Taiwan University (NTU), Taiwan, in 2011, and he is currently working towards the Ph.D. degree, also in GICE, NTU. His research interests include machine learning, computer vision, multimedia signal processing, and pattern recognition.

Jun-Zuo Liu received the B.S. degree from the Department of Communication Engineering, National Central University, Taiwan, in 2010. He is currently working towards the M.S. degree in Graduate Institute of Communication Engineering, National Taiwan University, Taiwan. His research interests include machine learning, pattern recognition, and face-related topics.

Jian-Jiun Ding was born in 1973 in Taiwan. He received the Ph.D. degree in 2001. He is currently an assistant professor with the Graduate Institute of Communication Engineering, National Taiwan University. His current research areas include time–frequency analysis, linear canonical transforms, image compression, image processing, integer transforms, pattern recognition, etc.