CrossMark

Age estimation algorithm of facial images based on multi-label sorting

Zijiang Zhu[1,3*], Hang Chen[2,3], Yi Hu[1,3] and Junshan Li[1,3]

Abstract

Multi-label sorting learning has been successful in many fields. It can not only express the complex semantic information of learning objects, but also present good generalization ability in dealing with complex things. This paper proposes age estimation algorithm of facial images based on multi-label sorting. This estimation algorithm is for the lack of facial age dataset, and it changes the traditional multi-valued classification method, simplified the problem of tedious steps to estimate age and shortened the time for model training. A series of experiments on two age datasets shows that the algorithm has achieved very good results in evaluating indicators, and these indicators include MAE (mean absolute error), CS (cumulative score), and convergence rate. When compared with some classic algorithms of age estimation, the efficiency and accuracy of the algorithm are verified.

Keywords: Multi-label sorting, Age estimation of facial images, Mean absolute error, Cumulative score

1 Introduction

In recent years, multi-label sorting learning technology has been widely used in the research fields of document classification, image recognition, gene function prediction, and so on. However, the technology is relatively less applied in the field of age estimation of facial images. In the age estimation dataset, the annotation method is usually a facial image corresponding to an accurate age value, but there are many problems with such a simple annotation method. The most important of these problems is that using an accurate age value to represent the true age of a facial image is unreliable and unstable, due to the slow changes in the face's appearance and the slight differences in facial images between similar ages as the age increases; it is easy to be confused with age classification. In addition, how to use the limited facial dataset to establish a good age estimation model has always been a problem followed with interest in the technical research process of facial age estimation, and the reason why this problem has not been resolved is due to the very small number of samples in the age dataset of facial images.

In the research of face age estimation, many scholars put forward a conclusion of classical algorithm. Geng et al. proposed Aging Patterns Subspace (AGES) algorithm [1]; by constructing a representative subspace, the aging pattern of human face is modeled, but the modeling time is longer. Guo et al. proposed the support vector machine (SVM) algorithm [2, 3] and support vector regression (SVR) algorithm [4, 5]; their algorithm is simple and robust, but it will consume a lot of machine memory and computing time. Hong et al. proposed k-nearest neighbor sorting (kNN) algorithm [6]; this algorithm is simple and accurate, but it needs a lot of calculation. In addition, Li et al. proposed Ordinal Hyperplanes Ranker (OHRank) algorithm [7], Geng et al. proposed Improved Iterative Scaling-Learning from Label Distribution (IIS-LLD) algorithm [8], and Yin et al. proposed Conditional Probability Neural Network (CPN) algorithm [9]. Among them, IIS-LLD and CPNN are one of the label distribution (LD) algorithms [10]. All the algorithms in reference [7–9] can be used to realize face age recognition. However, a large number of training samples are required, and the number of training samples directly affects the accuracy of recognition. Liao et al. proposed a face age feature extraction method based on deep convolution neural network [11]; it has strong discrimination and robustness, but the implementation of neural network is more complex, and the training time of model is longer.

* Correspondence: zzjgwng@sina.com
[1]School of Information Science and Technology, South China Business College, Guangdong University of Foreign Studies, No. 181 Liangtian Zhong Road, Baiyun District, Guangzhou 510545, China
[3]Institute of Intelligent Information Processing, South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, China
Full list of author information is available at the end of the article

In this paper, a face age estimation algorithm based on multi-label sorting is proposed, which can solve the problems of insufficient training samples and long training time of the model. At the same time, face age recognition of different races and different genders also has a better discrimination. Compared with other classical algorithms, it can achieve better face age recognition effect.

2 A review of multi-label sorting learning
2.1 Overview of multi-label sorting learning
Multi-label sorting learning can naturally express the complex semantic information of learning objects and has been successful in many fields. It was originally applied in the field of text processing; for example, it often has wrong classification due to the ambiguity of words in the document classification task. In addressing this challenge, at first, Schapire and Singer [12] and others improved the AdaBoost algorithm and achieved significant results in BoostTexter, a Boosting-based multi-label document classification system. After that, multi-label learning began to attract the attention of more and more researchers and often appeared in many research fields such as content identification of multi-media, bioinformatics, and information retrieval.

Learning a good classifier requires a lot of training sample data. However, due to some practical problems, there are very few samples about this kind of markers, so it is impossible to train a highly accurate classification model. For the label between different categories, the traditional classification learning believes that it has almost no correlation to exist independently. However, multi-label learning believes that it may have essential interrelationships for model training. Nevertheless, it is undeniable that making full use of useful information between class labels can effectively reduce the difficulty of learning tasks under limited training data conditions.

2.2 Theoretical content of multi-label sorting learning
The key to multi-label learning is to learn possible relationships [13] between the output label of the training samples, using these relationships to optimize mathematical models and improve model accuracy. First, it requires mathematical representation of the input training samples, where $X \in R^d$ indicates that the feature distribution of each training sample is within the d-dimensional feature space, where $Y = \{y_1, y_2, \cdots, y_m\}$ indicates a collection that covers all training sample category labels. Now, the existing training sample set is $D = \{(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)\}$, in which each sample is represented by a multi-label; $x_i = [x_{i1}, x_{i2}, \cdots, x_{id}]^T$ represents the characteristics of the ith training sample; $Y_i \subset Y$ represents the category of annotation by the ith training sample, which is a subset of the label set. In order to meet the requirements of model training, it is necessary to uniformly train the number of class labels in the sample and use $y_i = [y_{i1}, y_{i2}, \cdots, y_{im}]^T$ to represent the label vector corresponding to the sample. In $y_{ij} \in \{-1, +1\}$。, if $y_{ij} = +1$, sample x_i is marked as category y_j; otherwise, sample x_i is not marked as category y_j. The purpose of multi-label learning is to get the mapping function from input to output: $g : X \rightarrow 2^Y$ according to the training sample set. In the design of mapping functions, it is often converted to optimal solution to the function $f = X \times Y \rightarrow R$ according to the model requirements. $f(x, y)$ indicates the possibility that the sample x is marked as y, and the greater $f(x, y)$ is, the greater the probability that the sample x is marked as class y. Therefore, the purpose of training is to make higher confidence level of category between each training sample and its relevant sample, while the lower confidence level of category is between the training sample and its irrelevant sample.

After learning the mapping function, we need to use an evaluation indicator to measure the performance of it. For the multi-label attribute of the sample and learning the correlation between multiple label at the same time, multi-label sorting learning adopts the RL (ranking loss) function [14] as a standard to measure the function of the model. The specific calculation formula is shown in Eq. (1):

$$\mathrm{RL}(f, D) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|Y_i||\overline{Y_i}|} \left\{ (y, \overline{y}) | f(x_i, y) \leq f(x_i, \overline{y}), (y, \overline{y}) \in Y_i \times \overline{Y_i} \right\}$$

$$(1)$$

where Y_i and $\overline{Y_i}$ represent x_i's relevant label set and irrelevant label set, respectively. This indicator is used to calculate the proportion of errors in the label sorting process caused by relevant and irrelevant tags. Therefore, the smaller the value of RL is, the better. When RL is taken to 0, it means that the irrelevant tags are all behind the relevant tags on all the samples.

3 Method—age estimation mode of multi-label sorting
The primary difficulty faced by the research is the lack of training samples on age estimation of face images. For better age estimation results, multi-label replaces the original single label to represent marked face samples, and the sorting learning is based on the correlation between the age tag and the face image. So age estimation algorithm can well establish the mapping relationship between face image and age based on multi-label sorting learning.

The multi-label sorting learning is undoubtedly an important means to effectively alleviate the specificity of age estimation and the inaccurate age estimates caused by insufficient training data. This kind of multi-label sorting algorithm first expresses facial image with only single age label as a set of vectors, and the size of each element in the vector represents the correlation degree

Zhu *et al. EURASIP Journal on Image and Video Processing* (2018) 2018:114

Page 3 of 10

between the facial image and the corresponding age label. At the same time, it sorts the ages in ascending order and makes full use of the ordered information between age labels [15, 16] to integrate the estimation models of each age into a model matrix. It implements the age estimation model through the model matrix, instead of building an age estimation model by constructing multiple binary classifiers. In the model building process, it introduces trace norm [17] of the model matrix to control the complexity of its model algorithm and uses the matrix recovery theory to achieve the optimal solution of the model [18, 19]. After the age estimation model is acquired, a correlation vector is obtained through prediction of the face sample when performing the algorithm model test, all the elements in the vector are sorted in descending order, and the age label with the largest correlation is selected as the estimated age, while the basis for selection is that the larger sorting value indicates the higher relevance between the face image and the age label. This age estimation model shows a great advantage by making full use of the limited face age estimation dataset and successfully introducing a multi-label sorting technique, and at the same time, it is simple and effective to shorten the time of model training [20] in the operation of the matrix.

3.1 Based on multi-label face samples

In this paper, in order to adopt multi-label sorting to learn the age estimation model to let multiple age labels mark face images, first, $X = [x_1, x_2, \cdots, x_n] \in R^{d \times n}$ represents the input of the training sample set, $T = \{t_1, t_2, \cdots, t_m\}$ is labeled collection of all age labels, $t_1 \prec t_2 \prec \cdots \prec t_m$ represents the order relationship between the age label, and $Y = [y_1, y_2, \cdots, y_n] \in \{0, 1\}^{m \times n}$ is the age marker status corresponding to the training sample set. Wherein, if the sample x_i is marked with the age label t_i, the two are related, correspondingly to the elements $y_{ij} = 1$ and $y_{ij} = 0$ in the age label vector y_i; otherwise, they are not related. According to the label's representation method, a multiple age label is used instead of a single age label in the age dataset.

Convert a single-label sample to a multi-label sample, where each face image corresponds to a label vector. For

the traditional face age estimation problem, it is often converted into multiple category classification problem based on individual age labels, then the positive and negative samples are divided for each age value to construct binary classifier. After the use of multi-label representation, simple and effective matrix operations can not only achieve the age estimation algorithm for multi-label learning, but also learn the relationship between ages.

3.2 The establishment of age estimation model

After multi-label representation of face images, on the one hand, all age labels are integrated in ascending order. On the other hand, the traditional face age estimation problem based on multi-age classification is converted into the study of age matrix. In order to control the complexity of the age estimation model, the matrix norm of the model is introduced, and the matrix estimation theory is used to solve function of age estimation. The specific process of the model is shown in Fig. 1.

First, perform a mathematical description of the correlation function before the age estimation function is established. Assuming that $\ell(z)$ is a loss function, the prediction function $f_i(x)$ is used to estimate the age of test sample as the t_i confidence level, generally the confidence value is between 0 and 1. The learned model marks the age of the age tag and gives it a higher score. In order to measure whether the predicted age value predicted by age estimation function is accurate, c is used to calculate the sorting loss of face sample x_i marked as vector y_i between ages t_j and t_k. The specific calculation Eq. (2) is as follows:

$$\mathscr{E}_{j,k}(x_i, y_i) = I\left(y_{ij} \neq y_{ik}\right) \ell\left(\left(y_{ij} - y_{ik}\right)\left(f_j(x_i) - f_k(x_i)\right)\right) \quad (2)$$

In which the indicating function is $I(z)$, the confidence value is 1 when z is logical "true," and the confidence value is 0 when z is logical "false." The function shows that if the age label of face sample x_i is t_j instead of t_k, that is $y_{ij} = 1$ and $y_{ik} = 0$, the resulting correlation measurement should have $f_j(x_i) > f_k(x_i)$ according to the prediction function. The smaller the age difference between the age labels t_j and t_k, the closer the predicted
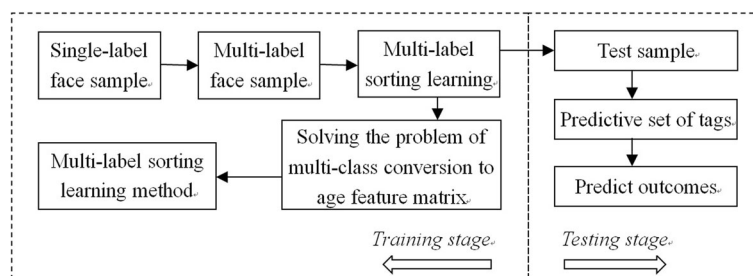

Fig. 1 Algorithm flow chart

correlation measurements are, then the smaller the sorting loss between them; on the contrary, the larger the gap of age label, the greater the similarity difference calculated, then the more likely to cause larger sorting loss. Therefore, it is in line with the definition that the correct age label is ranked in front of the incorrect age label in the face age estimation process. When the age label of face sample x_i is neither t_j nor t_k, then there is no sorting loss between the age labels t_j and t_k. From this, it can be seen that the sorting loss is caused by a piece of face image sample on all age labels as follows (3):

$$\varepsilon(x_i, y_i) = \sum\nolimits_{j,k=1}^{m} \varepsilon_{j,k}(x_i, y_i) \tag{3}$$

The sorting loss over the entire training data set is calculated based on a single face sample image and then obtained according to $\sum_{i=1}^{n} \mathcal{E}(x_i, y_i)$. To simplify the calculation, strictly limit the prediction function to a linear function $f_i(x) = w_i^T x$. Combine the prediction functions corresponding to all age labels into a parameter matrix, and define $W = [w_1, w_2, \cdots, w_m] \in R^{d \times m}$ as the matrix parameter to be learned. Assuming that $f(W)$ is the sorting loss over the entire training data set, define $f(W)$ according to the previous multi-label sorting theory as shown in Eq. (4):

$$
\begin{aligned}
f(W) &= \frac{1}{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j,k=1}^{m} \varepsilon_{j,k}(x_i, y_i) \\
&= \frac{1}{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j,k=1}^{m} I\left(y_{ij} \neq y_{ik}\right) \ell\left(\left(y_{ij} - y_{ik}\right)\left(w_j^T x_i - w_k^T x_i\right)\right)
\end{aligned}
\tag{4}
$$

In the case of serious shortage of training samples, this practice is easy to produce over fitting phenomenon by directly searching the value of W to minimize the sorting loss $f(W)$. Base on this, a long-term age estimation study found that the face age estimation process is slowly changing and there is an orderly correlation between the age labels. To take full advantage of this correlation, we believe that the prediction functions about W are linearly dependent, so W will also cause the matrix W to be low rank, and its final optimization problem can be expressed by Eq. (5):

$$\min_{W \in \psi} f(W) \ \ s.t. \psi = \left\{W \in R^{d \times m}, \ \text{rank}(W) \leq r, ||W||_2 \leq s\right\} \tag{5}$$

Among them, $,||\cdot||_2$ denotes the matrix spectral norm and ψ denotes the range of matrix W; the range consists of the complexity of the control prediction model and the low rank matrix characteristics. Because it is non-convex, the amount of calculation to directly solve Eq. (5) is very large. For convenience calculations, introduce the inequality Eq. (6):

$$||W||_* \leq \text{rank}(W)||W||_2 \tag{6}$$

Among them, $||\cdot||_*$ denotes the matrix trace norm. Use the inequality (2–5) to replace the non-convex item with the Eq. (7):

$$\psi' = \left\{W \in R^{d \times m} : ||W||_* \leq sr\right\} \tag{7}$$

So the problem of original optimization becomes a solution to $\min_{W \in \psi} f(W)$.

In order to further simplify the model, the final age estimation objective function is established by the value range of matrix as regular terms and the optimization problem as Eq. (8) shows:

$$\min_{W \in R^{d \times m}} F(W) := f(W) + \lambda||W||_* \tag{8}$$

Among them, the regular item parameter takes the value $\lambda > 0$. The parameter is used to balance the loss of the regular term and the training sample set and prevents the over fitting of the objective function.

3.3 Optimal solution
After establishing a multi-label ordered objective function, a gradient descent algorithm is used to prove that the objective function is a convex function. In order to simplify the calculation, the logistic function [21] $\ell(z) = \log(1 + e^{-z})$ is used in this paper as the formula in Eq. (2). At the time of the t iteration, under the premise that the solution of Eq. (8) is W_t, firstly calculate the gradient of the objective function $F(W)$ at $W = W_t$, and if the gradient is set to $\nabla F(W_t)$, then get the updated solution of the objective function as shown in Eq. (9):

$$W_{t+1} = W_t - \eta_t \nabla F(W_t) \tag{9}$$

Among them, η_t represents the step length that is updated at the tth iteration, which is generally set to a value greater than 0. Since $U_t \Sigma_t$ is gradient of W_* in $W = W_t$, $W_t = U_t \Sigma_t V_t^T$ is SVD (singular value decomposition) decomposition of W_t, then:

$$\nabla F(W_t) = \lambda U_t V_t^T + \frac{1}{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j,k=1}^{m} \alpha_{jk}^i x_i \left(e_j^m - e_k^m\right)^T \tag{10}$$

where, in the formula, $\alpha_{jk}^i = I(y_{ij} \neq y_{ik}) \ell'((y_{ij} - y_{ik}) x_i^T (w_j - w_k))$ and e_j^m a is an m vector, in which only the jth element is 1 and the other positions are 0.

Because of the complexity of the SVD decomposition calculation, it is easy to cause huge computational cost when solving the gradient Eq. (10). In the calculation of the gradient of the smooth objective function, many researchers find that its convergence rate can reach $O(T^{-2})$. In recent years, they have found a similar pattern, which is that the objective function can be

solved in an accelerated optimization manner if it contains a smooth term and a trace norm regular term. However, in this paper, accelerated proximal gradient (APG) algorithm [22] is used to solve the optimization problem of the objective Eq. (9). The available update values are shown in Eq. (11) according to the APG algorithm:

$$W_t = \arg \min_w \frac{1}{2\eta_t} ||\mathrm{W} - W_t'||_F 2 + \lambda ||W||_* \qquad (11)$$

Among them, $W_t' = W_{t-1} - \eta_t \nabla f(W_{t-1})$ is the optimal solution of the objective function according to the SVD decomposition algorithm, as shown in Eq. (12):

$$W_t = U \Sigma_{\lambda \eta t} V^T \qquad (12)$$

In Eq. (12), $\Sigma_{\lambda \eta t}$ is the diagonal matrix, and $(\Sigma_{\lambda \eta t})_{ij} = \max\{0, \Sigma_{ij} - \lambda \eta_t\}$ is also.

After the optimal solution is solved, we need to determine the step value η_t of each iteration, which plays an important role in the acceleration algorithm. In this paper, a simple linear search is used to find the most suitable η_t. Assume that $P_\eta(W_{t-1})$ is the optimal solution of Eq. (11) and $Q_\eta(P_\eta(W_{t-1}), W_{k-1})$ is the optimal value calculated by Eq. (11). First, assume a step value and then search for the optimal step value for each iteration based on $F(P_\eta(W_{t-1})) > Q_\eta(P_\eta(W_{t-1}), W_{k-1})$. The objective function can get the optimal matrix after the optimization calculation according to Table 1.

4 Age estimation prediction and model evaluation criteria

4.1 Age estimation prediction

Facial age recognition algorithm based on multi-label sorting makes full use of the ordered information between age labels in training samples. Through the multi-label sorting function established in the previous section and the optimization solution, the age characteristic matrix is finally obtained. Assuming that the obtained age characteristic matrix is a, the prediction

Table 1 Multi-label sorting calculation algorithm

Algorithm: Multi-label sorting calculation algorithm

Enter: Training sample set $X = \{x_i \in R^d\}_{i=1}^n$, corresponding to the age label set $Y = \{y_i \in \{0, 1\}^m\}_{i=1}^n$, and parameter λ initialized: $\eta_0 = 1$, $\gamma = 2$, $a_1 = 1$, $W_0 = Z_0 = Z_1 = R^{d \times n}$.

Cycling when conditions do not converge:
1. Set up $\overline{\eta} = \eta_{k-1}$
2. When $F(P_{\overline{\eta}}(Z_{k-1})) > Q_{\overline{\eta}}(P_{\overline{\eta}}(Z_{k-1}), Z_{k-1})$, so that $\overline{\eta} := \frac{\overline{\eta}}{\gamma}$
3. Set $\eta_k = \overline{\eta}$, and update $\eta_k = \overline{\eta}$:

$W_k = P_{\eta k}(Z_k)$
$a_{k+1} = \frac{1 + \sqrt{1 + 4a_k^2}}{2}$
$Z_{k+1} = W_k + (\frac{a_k - 1}{a_{k+1}})(W_k - W_{k-1})$

End Cycle

function of age estimation for this algorithm constructed thereby is as shown in Eq. (13).

$$y_t = W_*^T x_t \qquad (13)$$

Among them, x_t is the facial feature vector of the test face sample and y_t is the age label relevance vector calculated from the prediction Eq. (13). The size of each element represents the correlation degree between the tested face sample and the corresponding age label in this vector. So all elements are sorted in a descending order according to the correlation degree in the vector, which gives the result that it is most likely to approach the real age when the correlation between the top-ranked age value and the tested face sample is maximized. Therefore, the age of facial image can be estimated, thereby completing the design and implementation of the entire face age estimation algorithm based on multi-tag sorting.

4.2 Estimation model evaluation criteria

The face age estimation algorithm mainly uses mean absolute error (MAE) and cumulative score (CS) as the standard to measure the accuracy of age estimation [23].

4.2.1 Mean absolute error (MAE)

The mean absolute error is the average value of the absolute error between the predicted age and the true age through the age estimation of all tested face images. The specific formula is shown in Eq. (14):

$$\mathrm{MAE} = \frac{1}{\widehat{N}} \sum_{i=1}^{\widehat{N}} |a_i - \widehat{a}_i| \qquad (14)$$

Among them, $\widehat{N}$ is the number of tested face samples and a_i and $\widehat{a}_i$ are the true age of the ith tested face sample and the predicted age obtained by the age estimation algorithm, respectively. MAE visually describes the accuracy of sample set estimation through the age estimation algorithm. The smaller the MAE value, the higher the accuracy of the age estimation.

4.2.2 Cumulative score (CS)

The cumulative score indicates that the tested face image is predicted by the age estimation process, while the difference between the age estimated by the algorithm and the true age of the face image is less than or equal to the ratio between the number of test samples with predetermined threshold and the total number of test samples. The specific formula is shown in Eq. (15):

$$\mathrm{CS}(e) = \frac{1}{\widehat{N}} \sum_{i=1}^{\widehat{N}} g(|a_i - \widehat{a}_i| - e) \qquad (15)$$

Among them, $g(\cdot)$ is a Boolean function, and if $x \leq 0$, then $g(x) = 1$; conversely, $g(x) = 0$; e is the fault tolerance

rate, which is the set threshold; $\hat{N}$ is the number of test samples; a_i and $\hat{a}_i$ are the true age of the ith tested face sample and predicted age, respectively. In the case of determining the value of threshold e, the larger the value of CS, the more samples satisfy the condition and the better the effect of age estimation. This threshold is generally set to 10 years old because it has reached the upper limit of the maximum age estimation error.

The performance of the age estimation method can be evaluated from different perspectives based on two evaluation criteria. The MAE values reflect the error level of the age estimation algorithm as a whole; the CS reflects the accuracy of the age estimation method through the error statistic curve within each age error range. These two evaluation methods are complementary and coordinated.

5 Experimental result and discussions

To verify the accuracy of the proposed age estimation algorithm, a series of test experiments were conducted on the two authoritative age datasets FG-NET and Refined-MORPH in this paper. At the same time, to test the performance of the algorithm, the experimental comparison of the algorithm introduced in this paper will work with multiple mainstream algorithms with higher accuracy of age estimation.

5.1 Experiment setup

In order to test the performance of multi-label sorting algorithm, this paper selects two public age datasets FG-NET and Refined-MORP; according to the characteristics of the respective datasets, different facial features were extracted and the experiments were organized and tested.

(1) The FG-NET dataset collects images of 1002 face images of 82 different individuals scanned by old photographs. Each of these images has an average of 6 to 18 face images of different ages, and the age distribution is 0–69 years old. The AAM (active appearance model feature) [24] is extracted from face features using am_tools; it also incorporates face shape and texture information at the same time, thus fully embodying the change characteristics of the facial skull and slack skin during human growth. In the process of extracting the AAM feature of the face image, first, calibrate 68 key feature points of face, then extract face facial features with reference to the AAM feature section, and finally pick out the 95% AAM feature used by 95% of tests. The AAM feature is selected from the extracted features using a PCA (principal component analysis) dimension reduction algorithm, and its data dimension reaches 200 dimensions after selected. According to the characteristics of the age dataset itself, LOPO (leave-one-person-out) processing is used to divide training sets and test sets

when conducting the experimental design. That is all face images of an object are selected as test set, and all remaining face images are used as training set. After 82 experiments, the average of all results was used as the final age estimate. Since the FG-NET dataset has smaller number of pictures and is aimed at object acquisition, the use of the LOPO processing method is more precise and scientific for experimental design.

(2) In order to avoid the influence of gender and ethnicity on face age estimation studies, this paper uses the Refined-MORPH dataset. The dataset is carefully selected out of 21,060 face data samples from the MORPH-II dataset. These samples include 2570 white female face photos (white female, WF), 7960 white male face photos (white male, WM), 2570 black female face photos (black female, BF), and 7960 black male face photos (black male, BM). The number distribution of the dataset is balanced on the white and black pictures. To verify the influence of gender and ethnicity for face age estimation, a total of 16 experiments were set up for the following four aspects: ① no difference between gender and race, ② cross-race, ③ trans-sex, and ④ cross-race and gender. Before conducting the test experiment, the 4096-dimensional BIF feature was extracted from this dataset. It has been favored by many researchers in recent years in the field of face age estimation.

In the process of experimental comparison, in order to ensure the scientificity and consistency of the experiment and eliminate some influences, the age estimation algorithm proposed in this paper and other comparison algorithms are performed under the same conditions including the dataset and face recognition. To verify the performance of multi-label order learning on age estimation, this chapter selects several classic age estimation algorithms as comparison objects, including commonly used classification or regression algorithms to solve the problem of age estimation. In the experiment, UBSVM tool software was used to train SVM and SVR of face age estimation algorithm model. In the kNN algorithm, the k value is set to 30 according to general experience. The experiments of AGES, OHRank, and LD were performed according to the algorithm and parameters designed by the author. But beyond that, the two datasets also verify the convergence of the algorithm.

Meanwhile, in order to further verify the validity of the proposed multi-label learning in this paper and evaluate the impact of the ordering loss (R) and the norm (T) on age estimation in the objective function, compare it with the original classification loss (C) and the F norm (F) and compare it with the four groups of target loss functions: C&F, C&T, R&F, and R&T.

5.2 Analysis of results

According to the above experimental setup, this paper compares the proposed age estimation algorithm with the

Table 2 MAE results of FG-NET dataset

Methods	FG-NET
AGES	6.77
SVM	7.25
SVR	5.91
kNN	8.24
OHRank	4.48
IIS-LLD	5.76
CPNN	4.76
Multi-label	4.35

corresponding algorithm on the FG-NET and Refined-MORPH datasets and conducts detailed analysis and evaluation. In order to make a scientific evaluation of the algorithm, two evaluation indexes the MAE and the CS were used in this paper.

5.2.1 FG-NET dataset
Calculate in accordance with MAE and CS curves on the FG-NET dataset. The calculated MAE values are shown in Table 2 for all algorithms.

Table 2 shows the MAE values of the AGES, SVM, SVR, kNN, OHRank, IIS-LLD, and CPNN algorithms in the FG-NET dataset. It is thus clear that although the new proposed OHRank and CPNN algorithms have been reduced in age estimation error in recent years, the multi-label sorting algorithm proposed in this paper has reduced the estimated error rate by 3% and 9% respectively compared to the two, and it has achieved the best effect in all comparison algorithms. The CS curve is calculated on the FG-NET dataset according to the CS evaluation index calculation method as shown in Fig. 2,

and the threshold setting here accepts an age error value of 10 years.

It can be seen from Fig. 3 that although the FG-NET dataset is small, the multi-label sorting learning algorithm proposed in this paper has almost achieved the maximum CS value at each error value. It shows that the algorithm can estimate the number of accurate samples more and more as the acceptable error increases on the whole.

5.2.2 Refined-MORPH dataset
Table 3 shows the MAE values calculated by using the different algorithms in the Refined-MORPH dataset based on the face experiments of different genders and races.

Table 3 shows the MAE values of the AGRES, SVM, SVR, kNN, OHRank, IIS-LLD, and CPNN algorithms in the Refined-MORPH dataset. In this table, the newly proposed OHRank and label distribution algorithm still belongs to the previously proposed algorithm in general, but it is not as good as the multi-label sorting learning algorithm proposed in this paper on the whole. In addition, from this table, it is also found that the smallest age estimation error and the next smallest MAE value are between same-sex ethnic groups, while the largest MAE value is regardless of gender and ethnicity. This finding shows that the problem of age estimation of face is vulnerable to gender and ethnicity. From the data in Table 3, it can be concluded that the algorithm proposed in this paper is superior to other algorithms in the experiments of face age recognition of different sexes and races.

5.2.3 Convergence detection
Figures 3 and 4 show the convergence of the algorithm's objective function on the FG-NET and Refined-MORPH
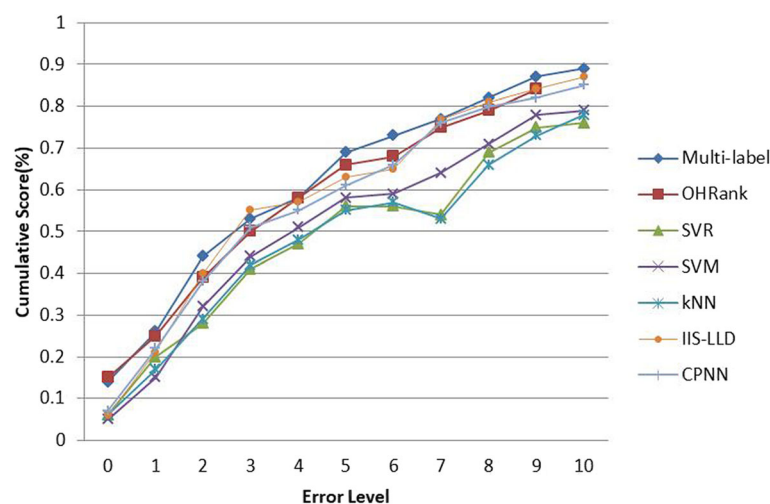
Fig. 2 CS curves with different error levels on the FG-NET dataset

datasets, respectively. Convergence has achieved a faster rate on both datasets; it illustrates the problem of translating the age estimation into matrix model through multi-label sorting learning, which not only simplifies the steps of the age estimation model, but also shortens the construction time of the model.

5.2.4 Comparison between different loss functions

To verify the effect of RL(R) and the norm (T) on age estimation in the objective loss function. In this paper, classification loss (C) and F norm (F) are used as benchmarks for comparison to form C&F, C&T, R&F, R&T total four age-estimated loss functions. Table 4 shows the MAE values obtained on the FG-NET and Refined-MORGH datasets. From the table, we can see that R&T achieves a smaller MAE value than other loss functions, which proves the effectiveness of the algorithm.

Based on the above experimental analysis, the age estimation algorithm of multi-label sorting learning proposed by this paper has achieved good results whether it is from the MAE value of evaluating indicator, the CS curve or the convergence speed, which is mainly attributed to the effectiveness of multi-label sorting learning. First, using multi-age labels to represent face samples in the case of limited training samples, it not only enriches the representation of the dataset to certain extent, but also transforms the traditional multi-category age estimation method into the solution of age matrix, thereby shortening the training time of the age estimation model. At the same time, in order to learn the orderly information between the age tags, the algorithm uses the sorting loss function and introduces the matrix norm, which not only successfully reduces the age estimation error, but also verifies the effectiveness of the proposed target loss function in this paper through the fourth set of experiments.

6 Conclusions

Although multi-label learning is widely used in the fields of text analysis, bioinformatics analysis, and so on and

Table 3 MAE results of Refined-MORPH dataset

Train	Test	Methods						
		SVM	SVR	kNN	OHRank	IIS-LLD	CPNN	Multi-label
WF	WF	7.77	7.75	6.79	5.10	4.86	5.32	4.56
	BF	8.27	7.79	9.38	6.56	7.14	7.22	6.65
	WM	9.41	9.46	9.51	6.93	6.34	7.48	6.36
	BM	9.03	8.96	9.49	8.13	8.26	8.51	8.21
WM	WF	9.69	6.66	10.74	8.39	7.29	8.43	5.86
	BF	9.04	7.32	10.25	8.51	8.11	8.71	6.63
	WM	5.05	5.14	4.90	4.16	3.59	4.15	3.43
	BM	8.58	8.33	9.68	6.24	6.26	6.45	6.67
BF	WF	8.37	8.06	10.66	7.47	8.47	7.29	6.79
	BF	7.57	7.12	7.85	5.95	5.49	6.35	5.25
	WM	9.41	8.46	10.72	8.34	9.38	8.90	8.12
	BM	9.23	9.33	10.30	7.61	7.67	8.19	7.26
BM	WF	13.33	7.47	11.33	11.47	10.32	7.29	8.45
	BF	11.47	6.99	11.16	9.72	9.06	6.35	6.56
	WM	7.53	6.95	9.38	6.13	6.95	8.90	6.01
	BM	6.33	5.77	7.39	4.97	4.24	8.19	3.87

presents good generalization ability in dealing with complex things, it is still unknown to improve the accuracy of face age estimation. For the insufficiency of the age dataset, this paper first transforms the single age label of the face image sample into the multi-label vector according to the correlation degree and then integrates a matrix of age characteristics following sequence of age, which changes the traditional method of multiple binary classification. The use and study of matrix operation of age simplifies the tedious steps of age estimation problem and shortens the model training time on age estimation model. At the same time, in order to take full advantage of the age-label ordering information to make up for the lack of training samples, the multi-label learning method uses ranking loss function to learn the
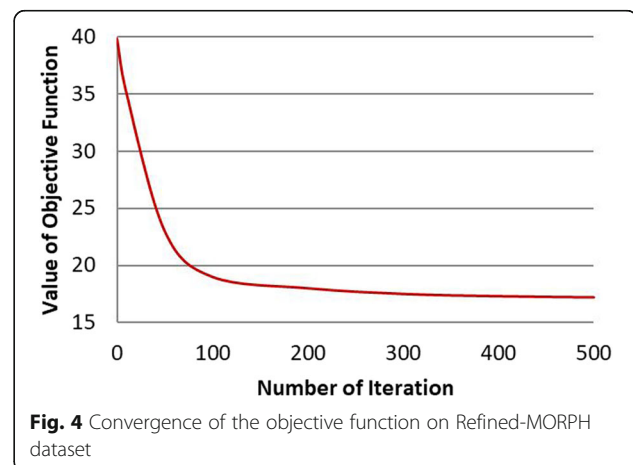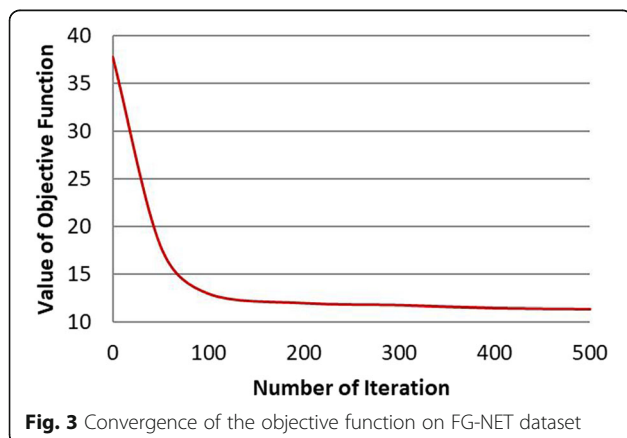

Fig. 3 Convergence of the objective function on FG-NET dataset

Fig. 4 Convergence of the objective function on Refined-MORPH dataset

Zhu *et al. EURASIP Journal on Image and Video Processing*   (2018) 2018:114

Page 9 of 10

Table 4 MAE results of FG-NET dataset

Methods	Image dataset	
	FG-NET	Refined-MORGH
C&F	6.91	10.78
C&T	6.46	10.17
R&F	5.43	9.79
R&T	4.35	8.12

sequence information between all age tags and introduces a matrix trace norm to control the complexity of an age estimation model. The optimization of an age estimation model is achieved through the APG algorithm during the solution process of the model. For the proposed multi-label learning algorithm, this paper conducted a series of experiments on two age datasets and verified its efficiency and accuracy compared to some classical algorithms of age estimation.

Abbreviations
AAM: Active appearance model feature; AGES: Aging Patterns Subspace; CPN: Conditional Probability Neural Network; CS: Cumulative score; IIS-LLD: Improved Iterative Scaling-Learning from Label Distribution; kNN: k-nearest neighbor sorting; LD: Label distribution; LOPO: Leave-one-person-out; MAE: Mean absolute error; OHRank: Ordinal hyperplanes ranker; PCA: Principal component analysis; RL: Ranking loss; SDV: Singular value decomposition; SVM: Support vector machine; SVR: Support vector regression; WM: White male

Availability of data and materials
We can provide the data.

About the authors
Zijiang Zhu received a master's degree in software engineering from Wuhan University in 2009 and a senior engineer title in 2008. From 2012 to 2016, he was the seventh batch of school level training objects of "thousand, 100 and ten" projects in Guangdong higher education institutions. He is a senior member of the China Computer Federation. He is currently an associate professor, dean of the School of Information Science and Technology, and deputy director of the Institute for intelligent information processing, South China Business College of Guangdong University of Foreign Studies, Guangzhou, China. His current research areas include image processing, machine learning, and big data technology.
Hang Chen received a master's degree in software engineering from Guangdong University of Technology in 2012 and a senior engineer title in 2007. From 2012 to 2016, he was the seventh batch of school level training objects of "thousand, 100 and ten" projects in Guangdong higher education institutions. He is currently an associate professor and associate dean of the School of Computer Science and Engineering, Tianhe College of Guangdong Polytechnic Normal University, Guangzhou, China. His current research areas include cloud computing technology, image recognition, data mining, and personalized recommendation technology.
Yi Hu received a master's degree in software engineering from South China University of Technology in 2018 and a senior engineer title for Information System Project Management in 2015. He is currently a lecturer and dean assistant of the School of Information Science and Technology, South China

Business College of Guangdong University of Foreign Studies, Guangzhou, China. His current research areas include image processing, machine learning, and big data technology.
Junshan Li was promoted to professor in 1999. He obtained a doctorate in computer system structure in 2001 and was elected as a provincial and ministerial expert in 2002. He is the head of the national boutique resource sharing course and the head of the national boutique course, the director of the China Computer Society, and the director of the Chinese Society of Image and Graphics. He is currently the director of the Institute of Intelligent Information Processing of South China Business School of Guangdong University of Foreign Studies. His current research interests include image processing and image understanding, intelligent computing, and intelligent systems.

Authors' contributions
All authors take part in the discussion of the work described in this paper. The author ZZ wrote the first version of the paper and did part of the experiments of the paper. HC, JL, and YH revised the paper in different versions of the paper. All authors read and approved he final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]School of Information Science and Technology, South China Business College, Guangdong University of Foreign Studies, No. 181 Liangtian Zhong Road, Baiyun District, Guangzhou 510545, China. [2]School of Computer Science and Engineering, Tianhe College of Guangdong Polytechnic Normal University, No. 628 Taihe Xingtai San Road, Baiyun District, Guangzhou 510540, China. [3]Institute of Intelligent Information Processing, South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, China.

References
1. X. Geng, Z.H. Zhou, K. Smithmiles, Correction to "automatic age estimation based on facial aging patterns". IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 368–381 (2008)
2. G. Guo, G. Mu, Y. Fu, Human age estimation using bio-inspired features. Comput. Vis. Pattern Recognit. **2009**. CVPR 2009. IEEE, 112–119 (2009)
3. G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. Comput. Vis. Pattern Recognit. IEEE, **42**(7), 657–664 (2011).
4. G. Guo, Y. Fu, T.S. Huang, Locally adjusted robust regression for human age estimation. IEEE Trans. Pattern Anal. Mach. Intell. **76**(6), 331–346 (2014)
5. C. Li, Q. Liu, Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Process. **64**(1), 1176–1188 (2016)
6. R. Hong, Z. Hu, L. Liu, Understanding blooming human groups in social networks. IEEE Trans. Multimedia **17**(11), 1–15 (2016)
7. C. Li, Q. Liu, W. Dong, Human age estimation based on locality and ordinal information. IEEE Trans. Cybern **45**(11), 2522–2534 (2017)
8. X. Geng, Z.H. Zhou, K.S. Miles, *Facial Age Estimation by Learning from Lable Distributions* (Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, 2010), pp. 451–456
9. C. Yin, X. Geng, Facial age estimation by conditional probability neural network. Pattern Recognit. Springer, Berlin Heidelberg **15**(2), 243–250 (2012)
10. Q. Zhao, X. Geng, Selection of objective functions in market-distributive learning. Com. Sci. Explor. **11**(5), 708–719 (2017)

11. Liao H B, Yan Y C, Dai W H and Fan P: Age Estimation of Face Images Based on CNN and Divide-And-Rule Strategy, Mathematical Problems in Engineering, 2018
12. R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**(3), 297–336 (1999)
13. G. Liu, Z. Lin, S. Yan, Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2010)
14. Z.B. Ren, L.L. Wang, Z.L. Fu, Multi-label classification integration learning algorithm based on ranking loss. Comput. Appl. **33**((S1)), 40–42 (2013) 68
15. C.W.L. Chao, J.Z. Liu, J.J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recogn. **46**(3), 628–641 (2013)
16. K. Chen, S. Gong, T. Xiang, Cumulative attribute space for age and crowd density estimation. IEEE Conf. Comput. Vis. Pattern Recognit. IEEE Comput Soc. **9**(4), 2467–2474 (2013)
17. K. Yu, X.J. Wu, Semi-supervised community discovery of latent mapping based on KL divergence matrix traces. Comput. Eng. **12**, 296–302 (2017)
18. K.-C. Toh, Yun, An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Pacif. J. Optim. **6**(3), 615–640 (2010)
19. S. Ji, J. Ye, *An Accelerated Gradient Method for Trace Norm Minimization* (International Conference on Machine Learning, ICML 2009, Montreal, 2009), pp. 457–464
20. S.J. Huang, Demilitarization of the Use of Marker Relationships in Multi-Marker Learning. Journal of Nanjing University (Natural Science Edition). **56**(8), 882-890 (2015)
21. Y.F. Guo, F.Y. Ning, H.H. Chao, A socialized matrix decomposition recommendation algorithm based on logistic function. J. Beijing Inst. Technol. **36**(1), 70–74 (2016)
22. Y.M. Wang, J.P. Zhai, Y. Mo, 3D reconstruction of human body based on orthogonal matching tracking and accelerating proximal gradient. Chin. J. Biomed. Eng. **36**(4), 385–393 (2017)
23. Q. Wang, Face Age Estimation Based on Adaptive Marker Distribution Learning. Journal of Southeast University (Natural Science Edition). **3**, 475–479 (2017)
24. L.F. Xu, J.Y. Wang, J.N. Cui, Dynamic expression recognition based on dynamic time warping and active appearance model. Chinese. J. Electron. Inf. (EIS) **40**(2), 338–345 (2018)