Learning Ordinal Discriminative Features for Age Estimation

Changsheng Li[1], Qingshan Liu[2], Jing Liu[1], Hanqing Lu[1]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190
[2]CICE, Nanjing University of Information Science and Technology, Nanjing, China, 210044
{csli,qsliu,jliu,luhq}@nlpr.ia.ac.cn

Abstract

In this paper, we present a new method for facial age estimation based on ordinal discriminative feature learning. Considering the temporally ordinal and continuous characteristic of aging process, the proposed method not only aims at preserving the local manifold structure of facial images, but also it wants to keep the ordinal information among aging faces. Moreover, we try to remove redundant information from both the locality information and ordinal information as much as possible by minimizing nonlinear correlation and rank correlation. Finally, we formulate these two issues into a unified optimization problem of feature selection and present an efficient solution. The experiments are conducted on the public available Images of Groups dataset and the FG-NET dataset, and the experimental results demonstrate the power of the proposed method against the state-of-the-art methods.

1. Introduction

In recent years, human age estimation attracted much attention in the communities of computer vision and pattern recognition due to its potential applications in soft-biometrics [8], human-computer interaction (HCI) [11], security control [8], surveillance monitoring [24], and electronic customer relationship management [8].

The purpose of age estimation is to label a face image automatically with the exact age (year) or the age group(year range). Generally, a facial age estimation system consists of two key modules: how to represent face image and how to estimate age based on facial feature. For face image representation, there are several popular methods including anthropometric models [19], active appearance model (AAM) [5], age subspace [11, 12] and manifold [9]. Given facial features, age estimation can be converted into a multi-class classification problem [11, 9] or a regression problem [14, 30, 28]. However, facial aging process is an ordinal procedure. For example, the face of a 5-year-old person is much more related to the face of a 10-year-old one than the face of a 30-year-old one. Motivated by ordinal characteristic of aging faces, some methods take age estimation as a ranking problem [4, 29, 20].

In this paper, we present a new age estimation method based on ordinal discriminative feature learning. We try to preserve the local manifold structure of facial images and the ordinal information among aging faces, which can better represent a facial aging process from a baby, child, growing up, to an old person as the years pass by. Figure 1 simply illustrates the motivation of the proposed idea. The facial images lie on a two-dimensional manifold with four different age labels. Although the feature $\mathbf{f}_1$ is the optimal feature for data representation, it cannot keep the ordinal information of data, which is of great importance to age estimation. We can see that $\mathbf{f}_1$ can not discriminate $\mathbf{A}_1$ group accurately. To keep both the ordinal information and the local manifold structure, the feature $\mathbf{f}_2$ is more preferable obviously. Our goal is to find $\mathbf{f}_2$ for age estimation.

In addition, good aging features not only need to preserve the local manifold structure and the ordinal information among aging faces, but also they should be independent as much as possible. Many studies have shown that eliminating redundant features can result in performance improvement [21, 2].

Thus, we first define the energy of preserving the locality and the energy of keeping the ordinal information for each feature, respectively. We also define the nonlinear correlation and the rank correlation to measure the redundant information between features. Based on these definitions, we formulate the feature selection problem into finding a subset of features, which can maximize preserving both the locality and the ordinal information and removing the redundancy among facial features for age estimation. In format, the objective function is similar to linear discriminant analysis [6], so we call it ordinal discriminant features learning. We conduct the experiments on two benchmarks: the Images of Groups dataset [10] and the FG-NET dataset [1], which are widely used for evaluating age estimation algorithms. The experimental results demonstrate the power of the proposed method compared to the state-of-the-arts.

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

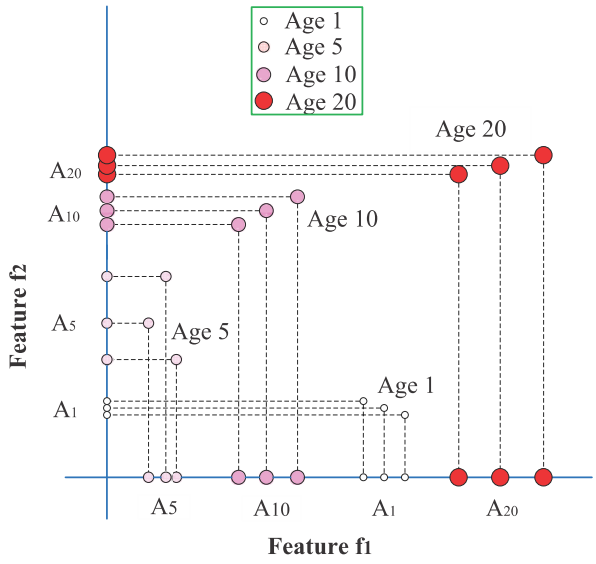


Figure 1: Illustration of the idea of learning ordinal discriminative features for facial age estimation, in which $Age\ b$ denotes the age label of the facial image, b=1, 5, 10, 20.

2. Related Work

In the past years, a lot of techniques have been proposed to extract discriminative aging features for age estimation, and they can be categorized into two classes. The first one is to learn a new low-dimensional feature space to represent facial images by feature transformation. Geng et al. [11, 12] proposed to define an image sequence of one subject as an aging pattern based on PCA model, and age estimation is performed by searching the proper position at age patterns. Instead of learning a specific aging pattern for each individual, Fu et al. [9] learned a common aging pattern or trend for many individuals at each age via manifold learning. Guo et al. [13] used the kernel partial least squares (KPLS) regression to reduce feature dimensionality and learn aging function simultaneously for age estimation.

The second one is directly to select a feature subset from original features, i.e., feature selection. Ricanek et al. [22] proposed a generalized multi-ethnic age estimation techniques, making use of the Least Angle Regression (LAR) [7] to select a subset of aging features to form a robust regression model. Shan [23] adopted Adaboost to learn discriminative local aging features. All these methods treat age estimation as either a classification or a regression problem. Recently, Yang et al. [29] employed the RankBoost algorithm to conduct feature selection for each individual. This approach extracts discriminative aging features for each individual, while they do not consider preserving structure information of data.

We propose a new approach to select the ordinal discriminative aging features for age estimation. It belongs to the categorization of feature selection. In machine learning community, there also have some popular feature selection methods. For example, Duda et al. [6] proposed a feature selection algorithm called Fisher score. Fisher score assigns the highest score to the feature, on which the data points of different classes are far from each other and the data points of the same class are required to be close to each other, and then it selects the top-d ranked features with high scores. In [15], a different performance criterion was developed for feature selection, in which the importance of a feature is evaluated by its power of locality preserving. Cai et al. [3] proposed a multi-cluster feature selection approach, which selected those features such that the multi-cluster structure of the data can be best preserved. Although these feature selection algorithms achieved much success in some tasks of clustering and classification, they are not suitable for age estimation, because they did not take account of ordinal information of aging processing at all.

3. The Proposed Method

Suppose that $\mathcal{Q} = \{(\mathbf{x}_i, l_i)\}_{i=1}^{N}$ is a training set of facial images, where $\mathbf{x}_i \in \mathbb{R}^M$ and l_i are the facial representation and the age label of the i-th person, respectively. M is the feature dimensionality of the training set. Let $\mathcal{F} = \{\mathbf{f}_u\}_{u=1}^{M}$ be the whole feature set, where $\mathbf{f}_u$ is the u-th feature. Our goal is to find a d-dimensional feature subset from $\mathcal{F}$, which contains the most informative features. In other words, the data $\mathbf{x}'_i, i = 1, 2, \cdots N$, represented in the d-dimensional space $\mathbb{R}^d$, can well preserve the local manifold structure and the ordinal information as the data represented in the original space $\mathbb{R}^M$.

3.1. Preserving Locality and Ordinal Information

In order to preserve the local manifold structure of the data and the ordinal information among data groups of different age labels, we formulate the objective function as:

$$\begin{aligned} \max J_1(y_1, y_2, \cdots, y_M) &= \sum_{u=1}^{M} w_u^L y_u + \alpha \sum_{u=1}^{M} w_u^R y_u \\ s.t.\ \ y_u &\in \{0, 1\}, u = 1, 2, \cdots M \\ \sum_{u=1}^{M} y_u &= d \end{aligned} \tag{1}$$

where w_u^L is the importance of the feature $\mathbf{f}_u$ in preserving the local manifold structure of the data, and w_u^R is the importance of the feature $\mathbf{f}_u$ in keeping the ordinal information among the data. $y_u = 1$ (or 0) indicates the feature $\mathbf{f}_u$ is selected (or not). M is the dimensionality of the original feature set, and d is the dimensionality of the selected feature

subset. α is a parameter to balance the importance of the local manifold structure and that of the ordinal information.

Clearly, the first term of the objective function in (1) intends to preserve the local manifold structure information of the data, while the second term aims to keep the ordinal information of the observations. By maximizing these two terms jointly, the selected features can preserve well both kinds of information.

3.1.1 The Importance of Preserving Locality

Manifold learning uncovers the nonlinear structure by integrating the descriptions of a set of local patches using the neighborhood graph [16]. Note that different features have different degrees which they maintains the graph structure to. A "good" feature should guarantee two data points close to each other only if the two points are two neighborhood points in original space. In order to evaluate whether a feature is "good" or not, a reasonable criterion is introduced as in [15]:

$$L_u = \frac{\sum_{i,j}(f_{u,i} - f_{u,j})^2 \mathbf{A}_{i,j}}{Var(\mathbf{f}_u)} \tag{2}$$

where $f_{u,i}$ and $f_{u,j}$ denote the u-th feature of the i-th sample and the j-th sample, respectively. $Var(\mathbf{f}_u)$ is the variance of the feature $\mathbf{f}_u$ in the data manifold $\mathcal{M}$. $Var(\mathbf{f}_u) = \int_{\mathcal{M}}(\mathbf{f}_u - \vec{\mathbf{f}_u})dP(\mathbf{f}_u)$, where $P(\mathbf{f}_u)$ is the probability measure and $\vec{\mathbf{f}_u}$ is the expected value of $\mathbf{f}_u$. $\mathbf{A}$ is an $N \times N$ adjacency matrix, and it can be constructed by a neighborhood graph. The adjacency matrix is defined as:

$$\mathbf{A}_{i,j} = \begin{cases} \exp(-\frac{d(x_i,x_j)^2}{2\sigma}) & \text{if } j \in \mathcal{N}_i \text{ and } i \in \mathcal{N}_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathcal{N}_i$ denotes the index set of the K-nearest neighbors of $\mathbf{x}_i$, and σ is empirically set by $\sigma = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{x}_{i_K})^2/N$ where $\mathbf{x}_{i_K}$ is the K-th nearest neighbor of $\mathbf{x}_i$.

After calculating L_u by Eq.(2), we prefer those features with the smaller values of L_u. The smaller L_u is, the stronger the ability of the feature $\mathbf{f}_u$ in locality preservation is. Based on the results of L_u, we define the importance of the feature $\mathbf{f}_u$ in locality preservation as:

$$w_u^L = \frac{1}{L_u}, u = 1, 2, \cdots, M \tag{4}$$

3.1.2 The Importance of Keeping Ordinal Information

As mentioned above, besides preserving the locality, we also try to keep the ordinal information, because aging process is an ordinal dynamic processing in temporal domain. Thus, we hope to select those features with strong ability in keeping the ordinal information.

We first uniformly split the training set $\mathcal{Q}$ into two parts: the training subset $\mathcal{Q}_1 = \{(\mathbf{x}_i, l_i)\}(i = 1, 2, \cdots, \widetilde{N})$ used for training ranking models, and the evaluation subset $\mathcal{Q}_2 = \{(\mathbf{x}_i, l_i)\}(i = \widetilde{N}+1, \widetilde{N}+2, \cdots, N)$ used for evaluating the ranking models. We then use each feature $\mathbf{f}_u(u = 1, 2, \cdots, M)$ to represent the training subset $\mathcal{Q}_1$, and train a Ranking SVM model [17] for each feature $\mathbf{f}_u$. Totally, we train M ranking models, and obtain M prediction score lists $\{\hat{\mathcal{L}}_u\}_{u=1}^{M}$ for the evaluation subset $\mathcal{Q}_2$ correspondingly. Finally, for each prediction list, we use an evaluation measure to calculate the similarity between the predicted list and the age label list of the ground truth $\mathcal{L} = \{l_{\widetilde{N}+1}, \cdots, l_N\}$, and we take the similarity score as the importance in keeping the ordinal information. The similarity between two lists is measured by Kendall's τ [18], which is demonstrated to be a good similarity measurement for ranking. The Kendall's τ value between two lists can be calculated as:

$$\tau(\hat{\mathcal{L}}_u, \mathcal{L}) = \frac{\sum_{i,j=\widetilde{N}+1, i\neq j}^{N} [\![(\hat{l}_{u,i} - \hat{l}_{u,j})(l_i - l_j)]\!]}{(N - \widetilde{N})(N - \widetilde{N} - 1)} \tag{5}$$

where $[\![\bullet]\!]$ is 1 if the inner condition is positive, and 0 otherwise. $\hat{l}_{u,i}$ is the prediction score of the i-th sample in $\hat{\mathcal{L}}_u$, and l_i is the age label of the i-th sample.

Clearly, the more similar the prediction list and the ground-truth age label list are, the stronger the ability of the corresponding feature in keeping the ordinal information is. Then we define the importance in keeping the ordinal information as:

$$w_u^R = \tau(\hat{\mathcal{L}}_u, \mathcal{L}), u = 1, 2, \cdots M \tag{6}$$

3.2. Removing Redundancy among Features

In order to eliminate the redundant features, we present another objective function as follows:

$$\begin{aligned} \min J_2(y_1, y_2, \cdots, y_M) &= \sum_{\substack{u,v=1 \\ u\neq v}}^{M} (s_{u,v}^L + \beta s_{u,v}^R) y_u y_v \\ s.t.\ & y_u \in \{0, 1\}, u = 1, 2, \cdots M \\ & \sum_{u=1}^{M} y_u = d \end{aligned} \tag{7}$$

where β is a weighting factor. $s_{u,v}^L$ denotes the redundant information between the feature $\mathbf{f}_u$ and the feature $\mathbf{f}_v$ in preserving the locality information. $s_{u,v}^R$ denotes the redundant information between $\mathbf{f}_u$ and $\mathbf{f}_v$ in keeping the ordinal information. By minimizing these two terms jointly, the selected features will contain minimal redundant information.

Since manifold is a kind of nonlinear geometry structure, we use nonlinear correlation to measure the redundancy local structure information between features. Meanwhile, we

utilize ranking correlation to measure the redundancy ordinal information between them.

3.2.1 Nonlinear Correlation

The nonlinear correlation coefficient (NCC) [25] is a good way to measure the correlation between nonlinear features. We use it to evaluate the redundant local information between two features. Considering two features $\mathbf{f}_u = \{f_{u,i}\}_{i=1}^N$ and $\mathbf{f}_v = \{f_{v,i}\}_{i=1}^N$, where N is the number of the training data. Assume that the values of each feature are sorted in ascending order. For each feature, put the sorted values into b ranks, i.e., put the first N/b samples into the first rank and the second N/b samples into the second rank, and so on. Then all the samples pairs $\{f_{u,i}, f_{v,i}\}_{i=1}^N$ can be placed into the $b \times b$ two dimensional rank grids by comparing the sample pairs to the rank sequences of $\mathbf{f}_u$ and $\mathbf{f}_v$, and the nonlinear correlation coefficient is defined as:

$$NCC(\mathbf{f}_u, \mathbf{f}_v) = -\sum_{i=1}^{b} \frac{n_{u,i}}{N} \log_b \frac{n_{u,i}}{N} - \sum_{i=1}^{b} \frac{n_{v,i}}{N} \log_b \frac{n_{v,i}}{N} + \sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{i,j}}{N} \log_b \frac{n_{i,j}}{N} \quad (8)$$

where $n_{u,i}$ and $n_{v,i}$ are the number of the samples distributed in the i-th ranks of about features $\mathbf{f}_u$ and $\mathbf{f}_v$, respectively , and $n_{i,j}$ is the number of the samples distributed in the ij-th rank grid.

Note that the first term and the second term in (8) are the revised entropies of the features $\mathbf{f}_u$ and $\mathbf{f}_v$ repectively, and the last term is the revised joint entropy of the features $\mathbf{f}_u$ and $\mathbf{f}_v$. Thus, the nonlinear correlation coefficient can be deemed as the revised mutual information.

After obtaining the nonlinear correlation coefficient, the redundant local structure information between the feature $\mathbf{f}_u$ and the feature $\mathbf{f}_v$ is defined as:

$$s_{u,v}^L = NCC(\mathbf{f}_u, \mathbf{f}_v), u, v = 1, 2, \cdots M, u \neq v \quad (9)$$

3.2.2 Ranking Correlation

Ranking correlation reflects the redundant ordinal information between the features. According to subsection 3.1.2, we have M prediction lists $\{\hat{\mathcal{L}}_u\}, u = 1, 2, \cdots M$, based on M features. Then we define the ranking correlation as:

$$RC(\mathbf{f}_u, \mathbf{f}_v) = \frac{\sum_{i,j=\widetilde{N}+1, i\neq j}^{N} [\![(\hat{l}_{u,i} - \hat{l}_{u,j})(\hat{l}_{v,i} - \hat{l}_{v,j})]\!]}{(N - \widetilde{N})(N - \widetilde{N} - 1)} \quad (10)$$

Based on the above definition, the redundant ordinal information can be represented by:

$$s_{u,v}^R = RC(\mathbf{f}_u, \mathbf{f}_v), u, v = 1, 2, \cdots M, u \neq v \quad (11)$$

3.3. Optimization Problem

To preserve both the local manifold structure and the ordinal information and remove the corresponding redundant information, we combine the objective function (1) with (7), and formulate them into a unified optimization problem as:

$$\begin{aligned} \max J(y_1, y_2, \cdots y_M) &= \frac{J_1(y_1, y_2, \cdots y_M)}{J_2(y_1, y_2, \cdots y_M)} \\ &= \frac{\sum_{u=1}^{M} (w_u^L + \alpha w_u^R) y_u}{\sum_{\substack{u,v=1 \\ u\neq v}}^{M} (s_{u,v}^L + \beta s_{u,v}^R) y_u y_v} \\ s.t.\ \ & y_u \in \{0, 1\}, u = 1, 2, \cdots M \\ & \sum_{u=1}^{M} y_u = d \end{aligned} \quad (12)$$

From (12), we can see that maximizing J is equivalent to maximizing J_1 and minimizing J_2, so we can obtain the desired feature subset. In format, it is similar to linear discriminant analysis [6]. The optimization in (12) is a typical 0-1 integer programming problem. When the original feature dimensionality M is high, finding its optimal solution by exhaustive search is very difficult because of the huge computation cost, $O(C_M^d)$. We utilize a greedy heuristics algorithm for solving the problem as: use an iterative process to select features, and assume that the feature subset in the next iteration contains the feature subset in the current iteration, which is widely used in many additive models, such as Boosting. We choose a feature from the unselected feature candidate set each time, and it thus needs d iterations for selecting d features. The proposed algorithm can be summarized as in algorithm 1.

4. Experiments

To evaluate the performance of the proposed method, we conduct the experiments on two public available datasets: Images of Groups dataset [10] and the FG-NET aging dataset [1]. We compare the proposed method with some related feature selection algorithms, such as Fisher Score [6], RankBoost [29], Laplacian Score [15], and Least Angle Regression (LAR) [22, 7]. For simplicity, we name the proposed method PLO. The number of the nearest neighbors (K) for constructing the neighborhood graph is set to 10 in the experiments. Following [9], we set 150 as the upper limit for the dimension of each learned feature subset.

All the above methods are about feature selection, so based on the selected features, we adopt the ordinal hyperplanes ranker (OHRank) model as the age estimator, which has been demonstrated to be a good age estimator [4]. OHRank treats age estimation as a ranking problem, and it employs relative ordinal information between ages and converts into a series of P subproblems of binary classifications according to the ordinal property. We test the proposed

Algorithm 1 The proposed method

Input: Data set $\mathcal{Q} = \{\mathbf{x}_i, l_i\}_{i=1}^{N}$
Original feature set $\mathcal{F} = \{\mathbf{f}_u\}_{u=1}^{M}$
Dimensionality d of the desired feature subset
The parameters α and β
Output: Feature subset Θ_d containing d features
Initialize $\Theta_0 = \phi, \Omega_0 = \mathcal{F}$
Method
for u=1:M **do**
Compute w_u^L according to Eq. (2) and (4)
Compute w_u^R according to Eq. (5) and (6)
end for
for u=1:$M-1$ **do**
for v=u+1:M **do**
Compute $s_{u,v}^L$ according to Eq. (8) and (9)
$s_{v,u}^L \leftarrow s_{u,v}^L$
Compute $s_{u,v}^R$ according to Eq. (10) and (11)
$s_{v,u}^R \leftarrow s_{u,v}^R$
end for
end for
$\mathbf{f} \leftarrow \arg\max_{\mathbf{f}_u \in \Omega_0}(w_u^L + \alpha w_u^R) y_u$
$\Theta_1 \leftarrow \Theta_0 \cup \{\mathbf{f}\}$
$\Omega_1 \leftarrow \Omega_0 \setminus \{\mathbf{f}\}$
for $t = 1, 2, \cdots, d-1$ **do**
$\mathbf{f} \leftarrow \arg\max_{\mathbf{f}_u \in \Omega_t} J(\Theta_t \cup \{\mathbf{f}_u\})$
$\Theta_{t+1} \leftarrow \Theta_t \cup \{\mathbf{f}\}$
$\Omega_{t+1} \leftarrow \Omega_t \setminus \{\mathbf{f}\}$
end for
end Method

method by two ways: age range estimation and exact age estimation.

4.1. Age Range Estimation

4.1.1 Experimental Settings

We conduct the experiments over the Images of Groups dataset [10] for age range estimation. This dataset consists of 28,231 faces from 5,080 Flickr images. Each face is labeled with age category, and seven age categories are considered: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+, roughly corresponding to different life stages. As in [10], face images are normalized to 61x49 pixels based on eye centers. Some typical aging face images in this dataset are shown in Figure 2. As in [20], the Gabor features are extracted to represent each facial image, and the corresponding dimensionality of feature vector is 1868.

Besides comparing with related feature selection algorithms, we also compare with the method in [10], which utilizes contextual features for age range estimation. For fair comparison with [10], we also randomly choose 3500/1050 faces from the whole database as the training/test sets.

Figure 2: Examples of faces in the Images of Group dataset.

Table 1: Age Range estimation on Images of Groups dataset [10].

Method	AEM	AEO
Apprearance+Context [10]	42.9%	78.1%
All Original Feature	27.8%	68.9%
Laplacian Score (150 dims) [15]	35.5%	74.5%
LAR (110 dims) [22, 7]	44.8%	84.9%
Fisher Score (150 dims) [6]	42.8%	83.7%
RankBoost (150 dims) [29]	44.8%	84.5%
PLO (150 dims)(Ours)	**48.5%**	**88.0%**

The evaluation measures are the accuracy of an exact match (AEM) and the accuracy of allowing an error of one age category (AEO) (e.g. an 8-12 year old predicted as 13-19 year old). They are defined as follows [10]

$$AEM = \frac{\hat{N}_m}{N'} \times 100\% \tag{13}$$

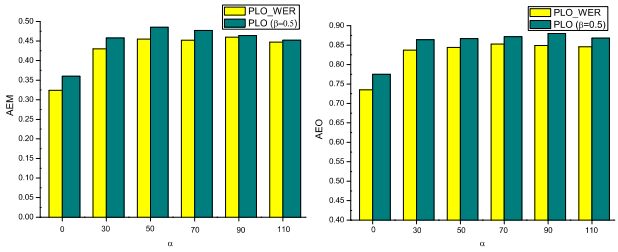
$$AEO = \frac{\hat{N}_o}{N'} \times 100\% \tag{14}$$

where $\hat{N}_m$ is the number of an exact match with N' test images, and $\hat{N}_o$ is the number of correct prediction when allows an error of one age category.

4.1.2 Experimental Results

The experimental results are reported in Table 1. PLO achieves the highest accuracies in terms of both AEM and AEO than all the other methods. For AEM, the improvements of PLO is more than 5% compared with the method in [10], even up to 10% in terms of AEO, which is significant and promising. In addition, the performance of using all the original features is the worst among these methods. The reason is that facial images come from Flickr and the quality of the images is very low, thus the extracted features contain much irrelevant and redundant information, which makes the performance drop.

Because this dataset is a real-word dataset, performing the experiments to deeply analyze the proposed method is very significant, including evaluating the performance of eliminating redundant information, the importance of respectively preserving the locality, the ordinal information and



(a) AEM (b) AEO

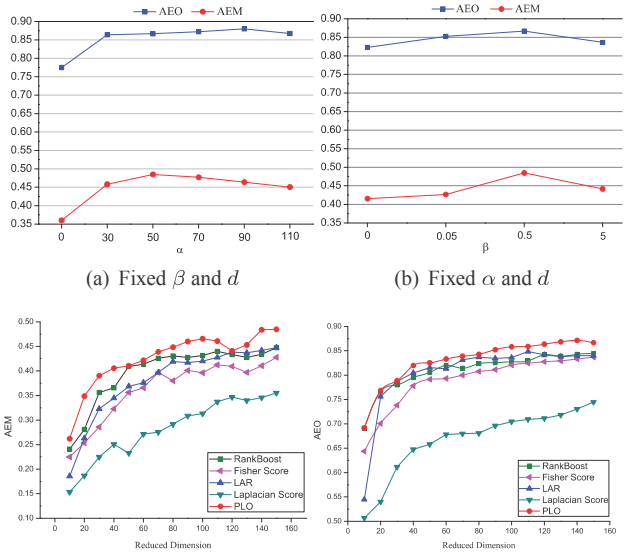
Figure 3: The results of two methods under different parameter α. (a) AEM. (b) AEO

removing the redundancy, and investigating the influence of different parameter settings. Only maximizing the numerator term in objective function (12) means that we do not take account of removing redundant information in PLO.

The results with or without eliminating redundant information are shown in Figure 3, where β is setting 0.5 in PLO and "PLO_WER" denotes PLO without eliminating the redundancy. It turns out that eliminating the redundancy performs better than that of without eliminating the redundant information under all the α values. In addition, we also give some analysis on the effectiveness of only preserving the locality, only keeping the ordinal information, and only removing the redundancy for improving the performance. The best AEM values of PLO_L (i.e., only preserving the locality information), PLO_O (i.e., only keeping the ordinal information), PLO_R (i.e., only removing the redundancy) and PLO (i.e., the combination of the three terms) are 35.5%, 44.9%, 39.1% and 48.5%, respectively, and the best AEO values of PLO_L, PLO_O, PLO_R and PLO are 74.5%, 85.2%, 78.2 and 88.0%, respectively. Through the results, we find keeping the ordinal information is more important than preserving the locality information and removing the redundancy, and the combination of the three terms gives the best performance.

The critical parameters in PLO are α, β, and the feature dimensionality d. We first fix $\beta = 0.5$, $d = 150$, and vary α, then test the performance of PLO. The results are shown in Figure 4(a). The performance is stable in most cases when α changes. Both the curves of the accuracies in terms of AEM and AEO fall after rising with α increasing. We take AEM as an example. When $\alpha \leq 50$, AEM will increase when α increases, which shows keeping the ordinal information is important for age estimation. When α is greater than 50, AEM will decrease with α increasing, which shows preserving the local structure information of facial images is important for age estimation. The value of α in [50,90] obtains the better performance with both AEM and AEO measurement, respectively.

We then fix $\alpha = 50$ and $d = 150$ to evaluate the performance with different β. Figure 4(b) shows the results. The



(a) Fixed β and d (b) Fixed α and d

(c) Fixed α and β (d) Fixed α and β

Figure 4: The effects of α, β and dimensionality d in PLO algorithm.

accuracy curves sill fall after rising with β increasing, which shows that removing the redundant local information and ordinal information is important for age estimation. Since both turning points appear when $\beta = 0.5$, we set $\beta = 0.5$ in the experiments.

Furthermore, we fix $\alpha = 50$ and $\beta = 0.5$ to investigate the influence of the selected feature dimensionality d. The results are shown in Figure 4(c) and 4(d). PLO outperforms the other algorithms under most of dimensions.

4.2. Exact Age Estimation

4.2.1 Experimental Settings

The FG-NET aging dataset [1] contains 1002 face images with large variations in pose, expression and lighting, which has widely been used for exact age estimation. There are 82 subjects in total with the age ranges from 0 to age 69 years. Some typical aging face images in this dataset are shown in Figure 5. We also use the Gabor texture features in the experiment and construct a 2720-dimensional feature vector to represent the facial image.

As in [11, 4, 26, 27, 12], the algorithms are tested by the Leave-One-Person-Out (LOPO) mode. The parameters α and β in the experiments are determined based on the analysis results obtained on the first experiment. The performance of age estimation is measured by the mean absolute error (MAE) and the cumulative score (CS) [12].

The MAE is defined as the average of the absolute errors

Figure 5: Typical aging face sequences from younger to older of one subject in the FG-NET aging dataset.

between the estimated ages and the ground truth ages:

$$MAE = \sum_{i=1}^{N'} | \widetilde{l}_i - l_i | /N' \quad (15)$$

where l_i is the ground truth age for the test image i, $\widetilde{l}_i$ is the estimated age, and N' is the total number of test images.

The cumulative score is defined as:

$$CS(L) = (N_{e \leq L}/N') \times 100\% \quad (16)$$

where $N_{e \leq L}$ is the number of test images on which the age estimation makes an absolute error no higher than L years.

4.2.2 Experimental Results

Table 2 and Figure 6 show the MAE results and CS results with the optimal reduced dimensionality derived on the FG-NET dataset respectively. Since our method is different from [13, 14], we do not compare with them. The results demonstrate that PLO outperforms the other feature selection algorithms. From Table 2, we can see that with the optimal dimensionality, the MAE of PLO is 4.82, while the MAE of using all the original features is 5.56, which shows that it is beneficial to conduct feature selection for age estimation. In addition, RankBoost and LAR have better performance than Laplacian Score and Fisher Score. The reason may be that the ranking or regression based algorithm has the advantage of employing the ordinal relationship among aging faces.

Table 2: MAEs comparison of different algorithms on the FG-NET aging dataset.

Method	MAE
All Original Features	5.56
Laplacian Score (150dims) [15]	6.95
LAR (140 dims) [22, 7]	5.64
Fisher Score (120 dims) [6]	7.65
RankBoost (150 dims) [29]	6.02
PLO (90 dims)(Ours)	**4.82**

Figure 7 reports the influence of different reduced dimensions, and it shows that our method outperforms the other methods under most of dimensions. Meanwhile, PLO achieves the best performance for subjects with four to five years of MAE.

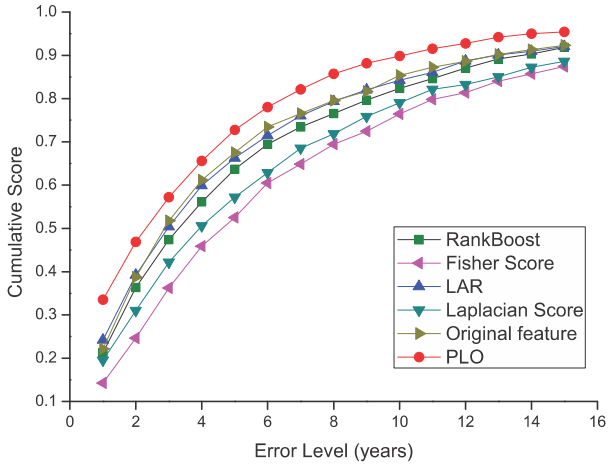


Figure 6: CS curves of the error levels from 1 to 15 years of different feature selection algorithms with the optimal feature subset.

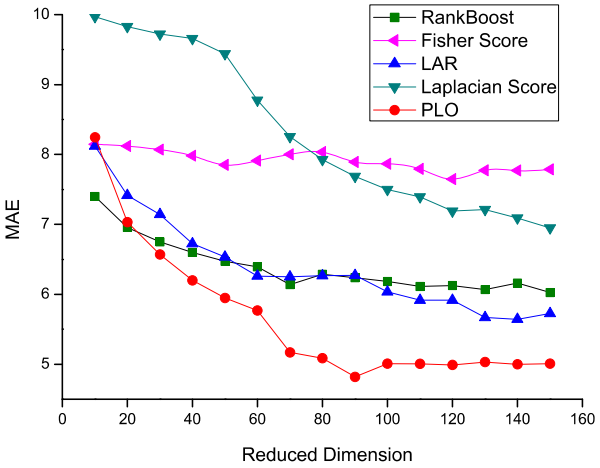


Figure 7: Exact age estimation results in terms of MAE on FG-NET aging dataset vs. the dimensionality of the selected features.

We also investigate the performance of PLO in estimating the relative order of two aging faces. Specifically, after obtaining the estimated ages of the test faces, we compare the correct rate of different methods on aging ranking of pairs of faces. In 82 folds, there are in total 5794 pairs of faces to be evaluated. The results are shown in Figure 8. PLO is better than the other methods too.

5. Conclusion

In this paper, we propose a novel facial age estimation method based on learning the ordinal discriminative aging features. The proposed method is based on two reasonable assumptions that facial aging images lie on a local mani-

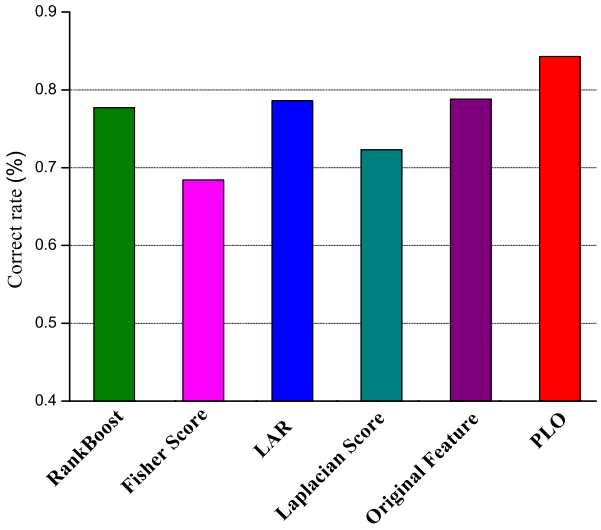


Figure 8: Correct rate (%) of different methods on aging ranking of pairs of faces.

fold and they are ordinal in temporal domain, so we aim at preserving both the locality and the ordinal information simultaneously. Furthermore, we try to remove the redundant local and ordinal information within the feature representation for facial images. Extensive experiments on two benchmarks demonstrate the power of the proposed method compared to several related methods.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 60833006, 60835002 and 60903146) and the funding of PAPD.

References

[1] The FG-NET aging Database, availabel at http://sting.cycollege.ac.cy/alanitis/fgnetaging/index.htm.

[2] A. Appice, M. Ceci, S. Rawles, and P. Flach. Redundant feature elimination for multi-class problems. In *Proc. ICML*, pages 507–514, 2004.

[3] D. Cai, C. Zhang, , and X. He. Unsupervised feature selection for multi-cluster data. In *Proc. KDD*, 2010.

[4] K.-Y. Chang, C. S. Chen, and Y. P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proc. CVPR*, pages 585–592, 2011.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. ECCV*, pages 484–498, 1998.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, Hoboken, 2000.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annal of Statistics*, 2004.

[8] Y. Fu, G. Guo, and T. Huang. Age synthesis and estimation via faces: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(11):1955–1976, 2010.

[9] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Trans. on Multimedia*, 10(4):578–584, 2008.

[10] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, pages 256–263, 2009.

[11] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 129(12), 2007.

[12] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *Proc. ACM Multimedia*, pages 307–316, 2006.

[13] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proc. CVPR*, 2011.

[14] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *Proc. CVPR*, 2009.

[15] X. He, D. Cai, , and P. Niyogi. Laplacian score for feature selection. In *Proc. NIPS*, pages 507–514, 2005.

[16] X. He and P. Niyogi. Locality preserving projections. In *Proc. NIPS*, 2003.

[17] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. KDD*, pages 133–142, 2002.

[18] M. Kendall. *Rank correlation methods*. Oxford University Press, 1990.

[19] Y. Kwon and N. Lobo. Age classification from facial images. *Computer Vision and Image Understanding.*, 74(1):1–21, 1999.

[20] Y. Ma, T. Xiong, Y. Zou, and K. Wang. Person-specific age estimation under ranking framework. In *Proc. ICMR*, 2011.

[21] H. Peng, F. Long, , and Q. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[22] K. Ricanek, Y. Wang, C. Chen, , and S. Simmons. Generalized multi-ethnic face age-estimation. In *IEEE conf. on BTAS*, 2009.

[23] C. Shan. Learning local features for age estimation on real-life faces. In *In ACM MM Workshops on MPVA*, 2010.

[24] Z. Song, B. Ni, D. Guo, T. Sim, and S. Yan. Learning universal multi-view age estimator by video contexts. In *Proc. ICCV*, 2011.

[25] Q. Wang, Y. Shen, and J. Q. Zhang. A nonlinear correlation measure for multivariable data set. In *Physica D*, 2005.

[26] S. Yan, H. Wang, T. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *Proc. ICME*, pages 96–99, 2007.

[27] S. Yan, H. Wang, X. Tang, and T. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *Proc. ICCV*, pages 1–8, 2007.

[28] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang. Regression from patch-kernel. In *Proc. CVPR*, pages 1–8, 2008.

[29] P. Yang, L. Zhong, and D. Metaxas. Ranking model for facial age estimation. In *Proc. ICPR*, 2010.

[30] Y. Zhang and D. Yeung. Multi-task warped gaussian process for personalized age estimation. In *Proc. CVPR*, 2010.