Human Age Estimation Based on Locality and Ordinal Information

Changsheng Li, Qingshan Liu, *Senior Member, IEEE*, Weishan Dong, Xiaobin Zhu, Jing Liu, *Member, IEEE*, and Hanqing Lu, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel feature selection-based method for facial age estimation. The face aging is a typical temporal process, and facial images should have certain ordinal patterns in the aging feature space. From the geometrical perspective, a facial image can be usually seen as sampled from a low-dimensional manifold embedded in the original high-dimensional feature space. Thus, we first measure the energy of each feature in preserving the underlying local structure information and the ordinal information of the facial images, respectively, and then we intend to learn a low-dimensional aging representation that can maximally preserve both kinds of information. To further improve the performance, we try to eliminate the redundant local information and ordinal information as much as possible by minimizing nonlinear correlation and rank correlation among features. Finally, we formulate all these issues into a unified optimization problem, which is similar to linear discriminant analysis in format. Since it is expensive to collect the labeled facial aging images in practice, we extend the proposed supervised method to a semi-supervised learning mode including the semi-supervised feature selection method and the semi-supervised age prediction algorithm. Extensive experiments are conducted on the FACES dataset, the Images of Groups dataset, and the FG-NET aging dataset to show the power of the proposed algorithms, compared to the state-of-the-arts.**

***Index Terms*—Age estimation, feature selection, local manifold structure, ordinal pattern, semi-supervised learning.**

I. Introduction

HUMAN faces provide large amounts of information, such as identity, age, gender, expression, and emotion [1], [2]. As a result, many research topics based on facial images have been extensively studied including face recognition, face reconstruction, expression recognition, gender and race classification, etc. [3]–[7]. In recent years, automatic human age estimation attracted much attention due to its potential applications in soft-biometrics [8], human-computer interaction [9], security control [9], surveillance monitoring [10], and electronic customer relationship management [8].

The purpose of age estimation is to automatically label a facial image with exact age (year) or age group (year range). Generally, a facial age estimation system consists of two key modules: 1) how to represent a facial image and 2) how to estimate its age based on the representation. Some methods have been proposed for facial image representation. Kwon and Lobo [11] proposed an anthropometric model based on the cranio-facial development theory and facial skin wrinkle analysis. The changes of face shape and texture patterns related to growth are measured to classify a face into one of the age groups. This model is suitable to estimate ages for young people [8]. Lanitis *et al*. [12] extended the active appearance model [13], which combined shape and intensity variation in facial images. A facial image is then represented by a set of model parameters. In recent years, some studies have shown that subspace learning is an effective method for aging feature representation [9], [14]. These studies can be divided into two categories based on different subspace learning techniques. The first one is to generate the components of the compact representation from the original feature space by a feature transformation algorithm. Geng *et al*. [15], [16] proposed to define an image sequence of one subject as an aging pattern based on principal component analysis, and age estimation is performed by searching the proper position at aging patterns. Later, Geng *et al*. [17] extended it into a nonlinear aging subspace. Instead of learning a specific aging pattern for each individual, Fu and Huang [9] learned a common aging pattern or trend for many individuals via manifold learning. Guo and Mu [18] used the kernel partial least squares regression to simultaneously reduce feature dimensionality and learn an aging function for age estimation. Chen and Hsu [19] learned a set of ranking features by feature transformation, where a sequence of constrained optimization problems are solved. The second one is to directly select a feature subset from all the original features by a feature selection algorithm. Comparing with feature transformation, feature selection keeps the same space with the input data, thus it has better interpretability for age estimation. Ricanek *et al*. [20] proposed a generalized multiethnic age estimation technique, which makes use of the least angle regression (LAR) [21] to select a subset of aging features. Shan [22] adopted Adaboost to learn

Manuscript received July 28, 2014; revised November 17, 2014; accepted November 23, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61272223 and Grant 61402023 and in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK2012045. This paper was recommended by Associate Editor Q. Zhao. *(Corresponding author: Q. Liu.)*

C. Li and W. Dong are with IBM Research-China, Beijing 100094, China (e-mail: lcsheng@cn.ibm.com; dongweis@cn.ibm.com).

Q. Liu is with B-DAT Laboratory, School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: qsliu@nlpr.ia.ac.cn).

X. Zhu is with the School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China (e-mail: xbzhu@nlpr.ia.ac.cn).

J. Liu and H. Lu are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jliu@nlpr.ia.ac.cn; luhq@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2376517

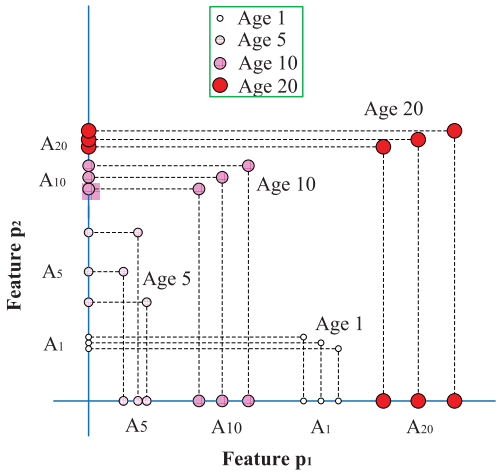


Fig. 1. Illustration of the idea of learning ordinal discriminative features for facial age estimation, in which *Age b* denotes the age label of the facial image, $b = 1$, 5, 10, and 20.

discriminative local aging features, and utilized support vector machine (SVM) as the classifier to predict the age. Recently, Yang *et al*. [14] employed the RankBoost algorithm [23] to conduct feature selection, which takes age estimation as a ranking problem.

After obtaining a feature representation, the next step is to estimate the age. Since the age is usually approximated by an integer, age estimation can be converted into a multiclass classification problem [1], [24] or a regression problem [5], [25], [26]. Due to the temporal property of the aging process, facial images should display certain ordinal pattern in the aging feature space. For example, the face of a five-year-old person is much more similar to the face of a ten-year-old one than the face of a 30-year-old one. Based on the ordinal characteristic of aging faces, some methods regard age estimation as a ranking problem [27]–[29]. Chang *et al*. [28] employed a parallel hyperplanes model to reduce the age estimation problem into a set of simpler binary questions, and combined the binary decisions to predict the age. Later, instead of building the parallel hyperplanes model, Chang *et al*. [27] converted the age estimation into a series of subproblems of binary classifications according to the ordinal relationship, and introduced a cost-sensitive property to further improve the performance.

In this paper, we propose a novel age estimation method. Our goal is to learn a low-dimensional aging feature representation by feature selection, which can depict the local manifold structure of facial images and preserve the ordinal relationship among aging faces, so that the compact representation can better describe a facial aging process from a baby, child, growing up, and to an old person [30]. Fig. 1 simply illustrates the motivation of the proposed idea. The facial images lie on a 2-D manifold with four different age labels. Although the feature $\mathbf{p}_1$ is the optimal feature for classification and clustering tasks because of its stronger ability of separation, it cannot keep the ordinal relationship of the data. We can see that $\mathbf{p}_1$ cannot discriminate $\mathbf{A}_1$ group accurately. To keep both the ordinal information and the local manifold structure, the feature $\mathbf{p}_2$ is more preferable obviously. Our goal is to find $\mathbf{p}_2$ for age estimation. In a high-dimensional aging feature space, good features not only need to preserve both the local manifold structure and the ordinal information among aging faces, but also should be independent as much as possible. In addition, many studies have shown that eliminating redundant features can bring with performance improvement [31], [32]. In light of these factors, we first measure the ability of each feature in preserving the local structure and the ordinal information, respectively. We also define nonlinear correlation and rank correlation to measure the redundant geometrical and ordinal information among the features respectively. Based on these definitions, we formulate the feature selection problem into finding a subset of features, which can maximally preserve both the locality and the ordinal information of the facial images. In format, the objective function is similar to linear discriminant analysis [33]. In practice, the cost of manually labeling facial aging images may be very high, and at the same time substantive unlabeled data is often easily accessible, thus we further extend this paper into a semi-supervised feature selection method for age estimation. Finally, we adopt a supervised age estimation technique, and also develop a semi-supervised age estimation algorithm to predict the age based on the low-dimensional aging features learned by our feature selection methods, respectively. This paper is different from [9] and [34]. The idea of [9] is to learn a feature transformation to preserve the manifold structure of facial images. Liu *et al.* [34] focused on learn a regression model to preserve the manifold structure and ordinal information of the data. The main contributions of this paper are summarized as follows.

1) We propose a novel supervised feature selection method for human age estimation, which aims to preserve both the ordinal information and the geometrical information as much as possible. The ordinal information can guarantee that the selected features have a good ordinal discriminative ability, and the geometrical information can assure that the data points represented by the selected features are smooth on the manifold.
2) A semi-supervised feature selection method is devised by exploiting the underlying structural information of the unlabeled data, which makes the proposed feature selection method fit in well with the real-world scenarios.
3) We further develop a semi-supervised age estimation algorithm by adding a manifold regularization, such that the decision function is smooth on the manifold. Therefore, the structure information of both the labeled data and the unlabeled data can be well exploited in the design of the age estimation algorithm.

II. Learning Ordinal Discriminant Features

Since the face aging process is a typical ordinal procedure in temporal dimension, and the high-dimensional facial images usually lie on or close to a low-dimensional local manifold, we present to learn ordinal discriminant features for age estimation, in which both the ordinal information and the geometrical information of the facial aging images are considered to be preserved. The ordinal information tries to guarantee that the selected aging features have a

good ordinal discriminative ability, while the geometrical information assures that the data points represented by the selected features are smooth on the manifold.

Suppose that $\mathcal{Q} = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}$ is a training set of facial images, where $\mathbf{x}_i \in \mathbb{R}^M$ and $l_i \in \mathbb{N}$ are the facial feature representation and the age label of the ith person, respectively. Let $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ denote the whole feature set, where $\mathbf{p}_u$ is the uth feature. Our goal is to find a feature subset with d features from $\mathcal{P}$, which can preserve the ordinal aging pattern, as well as the intrinsic geometric structure of facial images.

A. Preserving Locality and Ordinal Information (PLO)

In order to preserve both the local manifold structure of the facial images and the ordinal information among the images groups having different age labels, we formulate the objective function as

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

where w_u^L is the importance of feature $\mathbf{p}_u$ in preserving the local manifold of the images, and w_u^R is the importance of feature $\mathbf{p}_u$ in keeping the ordinal information among the facial images. $y_u = 1$ (or 0) indicates that the feature $\mathbf{p}_u$ is selected (or not). M is the number of the original features, and d is the number of the selected features. η is a parameter to balance the importance of the local manifold structure and that of the ordinal pattern. The first term of the objective function in (1) intends to preserve the intrinsic geometric structure information, and 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.

Next, we will show how to derive the importance of each feature in preserving the above two kinds of information, respectively.

1) Importance of Preserving Locality: Since facial images lie on or close to a low-dimensional local manifold, it is necessary to preserve the local manifold structure of the data in the selected feature space. In the feature space of the images, a "good" feature should guarantee two facial images are as close as possible only if the two images are two neighborhood points in the original space. In order to evaluate whether a feature is good or not, a reasonable criterion is introduced as in [35]

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

where $p_{u,i}$ and $p_{u,j}$ denote the uth feature observation values of the ith sample and the jth sample, respectively. $\mathrm{Var}(\mathbf{p}_u)$ is the variance of feature $\mathbf{p}_u$ in the data manifold $\mathcal{M}$. $\mathbf{A}$ is an $N \times N$ adjacency matrix, and it can be constructed by a neighborhood graph. The adjacency matrix is defined as

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

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 $\sigma = \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{x}_{i_K})/N$, where $\mathbf{x}_{i_K}$ is the Kth nearest neighbor of $\mathbf{x}_i$.

The criterion of constructing the adjacency matrix $\mathbf{A}$ adopts the "and" hypothesis, which means two facial images are connected if and only if they are neighbors to each other. A similar idea has been presented by [34] and [36]: they constructed a symmetry-favored graph by putting more weight on the connection agreed by both data points, and such a setting is more reliable than the K graph adopting the "or" hypothesis.

After calculating L_u by (2), we prefer those features with a smaller L_u. The smaller L_u is, the stronger the ability of feature $\mathbf{p}_u$ in locality preservation is. We thus define the importance of feature $\mathbf{p}_u$ in preserving the local information as

$$w_u^L = \frac{1}{L_u}, u = 1, \ldots, M. \tag{3}$$

2) Importance of Keeping Ordinal Information: Since the aging process is a typical ordinal dynamic processing in temporal domain, we thus hope to select such features that maximally keep the ordinal information in addition to the local information. To this end, we first uniformly split the training set $\mathcal{Q}$ into two parts: 1) the subset $\mathcal{Q}_1 = \{(\mathbf{x}_i, l_i)\}, i = 1, \ldots, \widetilde{N}$ and 2) the subset $\mathcal{Q}_2 = \{(\mathbf{x}_i, l_i)\}, i = \widetilde{N} + 1, \ldots, N$. Based on $\mathcal{Q}_1$, we train a ranking SVM model [37] for each feature. Thus there are M ranking models in total. We then evaluate each model on $\mathcal{Q}_2$, and obtain M corresponding prediction score lists, $\{\hat{\mathcal{L}}_1, \ldots, \hat{\mathcal{L}}_M\}$. Finally, we utilize an evaluation measure to calculate the similarity between each predicted list and the real age label list, and take the similarity score as the importance of the feature in keeping the ordinal information. The similarity between two lists is measured by Kendall τ [38], which is demonstrated to be a good similarity measurement for ranking. The Kendall τ value between two lists can be calculated as

$$\tau(\hat{\mathcal{L}}_u, \mathcal{L}) = \frac{\sum_{i,j=\widetilde{N}+1, i \neq j}^{N} [\![(\hat{l}_{u,i} - \hat{l}_{u,j})(l_i - l_j)]\!]}{(N - \widetilde{N})(N - \widetilde{N} - 1)} \tag{4}$$

where $[\![\bullet]\!]$ is 1 if the inner condition is positive, and 0 otherwise. $\hat{l}_{u,i}$ is the prediction score of the ith sample in $\hat{\mathcal{L}}_u$. $\mathcal{L} = \{l_{\widetilde{N}+1}, \ldots, l_N\}$ is the real age list of $\mathcal{Q}_2$, and l_i is the real age label of the ith sample.

Clearly, the more similar the prediction list and the real age label list is, the stronger the ability of the feature in keeping the ordinal information is. We thus define the ordinal information preservation importance of the feature $\mathbf{p}_u$ as

$$w_u^R = \tau(\hat{\mathcal{L}}_u, \mathcal{L}), u = 1, \ldots, M. \tag{5}$$

B. Removing Redundancy Among Features

The feature subset learned based on the objective function (1) might have substantial redundancy including locality redundancy and ordinal information redundancy. Consequently, to eliminate such redundancy among features,

we propose a new objective function as

$$\min J_2(y_1,\ldots,y_M) = \sum_{\substack{u,v=1\\u\neq v}}^{M} \left(s^L_{u,v} + \lambda s^R_{u,v}\right) y_u y_v$$
$$\text{s.t.} \quad \sum_{u=1}^{M} y_u = d, \quad y_u \in \{0,1\}, u = 1,\ldots,M \tag{6}$$

where λ is a weighting factor, $s^L_{u,v}$ denotes the redundant local information between $\mathbf{p}_u$ and $\mathbf{p}_v$, and $s^R_{u,v}$ denotes the redundant ordinal information between $\mathbf{p}_u$ and $\mathbf{p}_v$. By minimizing these two terms jointly, the selected features will contain minimal redundant information.

Since local manifold is a kind of nonlinear structure, we take advantage of nonlinear correlation to measure the redundant local information between two features. Meanwhile, we employ ranking correlation to measure the redundant ordinal information. In the following, we will discuss how to calculate nonlinear correlation and ranking correlation in detail, respectively.

1) Nonlinear Correlation: We take advantage of the nonlinear correlation coefficient (NCC) [39] to evaluate the redundant geometrical information between two features. Without loss of generality, we will take features $\mathbf{p}_u = \{p_{u,i}\}_{i=1}^N$ and $\mathbf{p}_v = \{p_{v,i}\}_{i=1}^N$ as an example, where N is the number of the training data. Assume that the values of each feature are sorted in ascending order. For each feature, we 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 $\{(p_{u,i}, p_{v,i})\}_{i=1}^N$ can be placed into the $b \times b$ 2-D rank grids by comparing the sample pairs to the rank sequences of $\mathbf{p}_u$ and $\mathbf{p}_v$. Then the NCC is defined as

$$\text{NCC}(\mathbf{p}_u, \mathbf{p}_v) = -\sum_{i=1}^{b}\left(\frac{n_{u,i}}{N}\log_b\frac{n_{u,i}}{N} + \frac{n_{v,i}}{N}\log_b\frac{n_{v,i}}{N}\right) + \sum_{i=1}^{b}\sum_{j=1}^{b}\frac{n_{i,j}}{N}\log_b\frac{n_{i,j}}{N} \tag{7}$$

where $n_{u,i}$ and $n_{v,i}$ are the number of the samples in the ith ranks of $\mathbf{p}_u$ and $\mathbf{p}_v$, respectively, and $n_{i,j}$ is the number of the samples distributed in the ijth rank grid. Note that the first term and the second term in (7) are the revised entropies of $\mathbf{p}_u$ and $\mathbf{p}_v$, respectively. The last term is the revised joint entropy of $\mathbf{p}_u$ and $\mathbf{p}_v$. Thus, the NCC can be deemed as the revised mutual information.

After obtaining the NCC, we measure redundant local information between $\mathbf{p}_u$ and $\mathbf{p}_v$ by

$$s^L_{u,v} = \text{NCC}(\mathbf{p}_u, \mathbf{p}_v), u, v = 1,\ldots,M, u \neq v. \tag{8}$$

2) Ranking Correlation: Similarly, we also try to eliminate redundant ordinal information among features. Ranking correlation can be used to describe the kind of redundancy between two features. According to Section II-A2, we have acquired M prediction lists, $\{\hat{\mathcal{L}}_u\}, u = 1,\ldots,M$. We then define the ranking correlation based on the M prediction lists as

$$\text{RC}(\mathbf{p}_u, \mathbf{p}_v) = \frac{\sum_{i,j=\widetilde{N}+1, i\neq j}^{N}[\![(\hat{l}_{u,i} - \hat{l}_{u,j})(\hat{l}_{v,i} - \hat{l}_{v,j})]\!]}{(N-\widetilde{N})(N-\widetilde{N}-1)}. \tag{9}$$

Based on the above definition, the redundant ordinal information can be calculated as

$$s^R_{u,v} = \text{RC}(\mathbf{p}_u, \mathbf{p}_v), u, v = 1,\ldots,M, u \neq v. \tag{10}$$

C. Optimization Procedure

To preserve the local manifold structure and the ordinal information, and remove the corresponding redundant information, we combine the objective function (1) with (6), and formulate them into a unified optimization problem as

$$\max J = \frac{J_1}{J_2} = \frac{\sum_{u=1}^{M}\left(w^L_u + \eta w^R_u\right) y_u}{\sum_{\substack{u,v=1\\u\neq v}}^{M}\left(s^L_{u,v} + \lambda s^R_{u,v}\right) y_u y_v}$$
$$\text{s.t.} \quad \sum_{u=1}^{M} y_u = d, \quad y_u \in \{0,1\}, u = 1,\ldots,M. \tag{11}$$

From (11), 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 [33].

The optimization in (11) is a typical 0–1 integer programming problem. When the original feature dimension M is high, it is difficult to find its optimal solution by exhaustive search, due to huge computation cost, $O(C^d_M)$. Next, we adopt a sequential optimization scheme to select the most informative feature subset. Suppose that $k(k \geq 0)$ features have been selected into the candidate set Θ_k

$$\Theta_k = \{g(1),\ldots,g(k)\}$$

where $g(i)$ denotes the index of the ith selected feature, and it has the following conditions:

$$\begin{cases} g(i) \in \{1,\ldots,M\}, 1 \leq i \leq k \\ g(i) \neq g(j), 1 \leq i \neq j \leq k. \end{cases}$$

Then, the $(k+1)$th feature can be selected by solving the following problem:

$$g(k+1) = \arg\max_{h\notin\Theta_k} U(h)/V(h)$$

where

$$U(h) = \sum_{i=1}^{k}\left(w^L_{g(i)} + \eta w^R_{g(i)}\right) + \left(w^L_h + \eta w^R_h\right)$$
$$V(h) = \sum_{i,j=1, i\neq j}^{k}\left(s^L_{g(i),g(j)} + \lambda s^R_{g(i),g(j)}\right) + \sum_{i=1}^{k}\left(s^L_{g(i),h} + \lambda s^R_{g(i),h}\right).$$

Initially

$$g(1) = \arg\max_{h\in\{1,\ldots,M\}}\left(w^L_h + \eta w^R_h\right).$$

Once $g(k+1)$ is obtained, the candidate set Θ_{k+1} can be updated as

$$\Theta_{k+1} = \Theta_k \cup g(k+1) = \{g(1),\ldots,g(k),g(k+1)\}.$$

The proposed supervised feature selection algorithm is summarized as in Table I.

TABLE I
PROPOSED LEARNING ORDINAL DISCRIMINANT FEATURES ALGORITHM

Input: Data set $\mathcal{Q} = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_N, l_N)\}$;
The feature dimension d; The parameters η and λ.
Output: The desired feature subset Θ_d containing d features.
Initialize $\Theta_0 = \phi, \Omega_0 = \{1, \ldots, M\}$.
Method
for u=1:M **do**
Compute w_u^L and w_u^R based on (2),(3), and (4), (5), respectively;
end for
for u=1:$M-1$ **do**
for v=u+1:M **do**
Compute $s_{u,v}^L$ based on (7) and (8); $s_{v,u}^L \leftarrow s_{u,v}^L$;
Compute $s_{u,v}^R$ based on (9) and (10); $s_{v,u}^R \leftarrow s_{u,v}^R$;
end for
end for
$g(1) \leftarrow \arg\max_{h \in \Omega_0}(w_h^L + \eta w_h^R)$;
$\Theta_1 \leftarrow \Theta_0 \cup \{g(1)\}$; $\Omega_1 \leftarrow \Omega_0 \setminus \{g(1)\}$;
for $k = 1, \ldots, d-1$ **do**
$g(k+1) \leftarrow \arg\max_{h \in \Omega_k} J(\Theta_k \cup \{h\})$
$\Theta_{k+1} \leftarrow \Theta_k \cup \{g(k+1)\}$; $\Omega_{k+1} \leftarrow \Omega_k \setminus \{g(k+1)\}$
end for
end Method

D. Extension to Semi-Supervised Feature Selection Method

In age estimation, it is often expensive to collect the labeled facial images. Meanwhile, the unlabeled images are relatively easy to get, and might be helpful for discovering the intrinsic geometric distribution of the facial images. Thus, we further extend our supervised feature selection method to a semi-supervised learning method.

Before explaining how to extend our method, we first analyze the nature of (11). When the parameters $\eta = \lambda = 0$, the objective function J aims at preserving the local manifold structure as much as possible. Then our method becomes an unsupervised learning method, because it does not use any age label information. When $\eta = \lambda = +\infty$, J aims to maximally keep the ordinal pattern of the facial images, and it is a supervised one because of incorporating age label information into the process of the learning. When $0 < \eta, \lambda < +\infty$, both the local manifold structure and the ordinal information are leveraged. Then it is still a supervised learning method. Based on the above analysis, we can easily extend our model into a semi-supervised one.

Let $\Xi = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^M$ denote the training set, which consists of N data points in an M-dimensional space. We assume that the first N_l images within Ξ are labeled by $\mathcal{L}_{N_l} = \{l_1, \ldots, l_{N_l}\}$, and the rest are unlabeled. For convenience, we denote the collection of the labeled images by $\Xi_{N_l} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_l}\}$, and the unlabeled images by $\Xi_{N_u} = \{\mathbf{x}_{N_{l+1}}, \ldots, \mathbf{x}_N\}$, such that $\Xi = \{\Xi_{N_l}, \Xi_{N_u}\}$. In order to learn ordinal discriminant features based on both the labeled and unlabeled facial images, we propose a similar objective function with (11) as

$$\max \widehat{J}(y_1, \ldots, y_M) = \frac{\sum_{u=1}^{M} \left(\widehat{w}_u^L + \eta \widehat{w}_u^R\right) y_u}{\sum_{\substack{u,v=1 \\ u \neq v}}^{M} \left(\widehat{s}_{u,v}^L + \lambda \widehat{s}_{u,v}^R\right) y_u y_v}$$

$$\text{s.t.} \quad \sum_{u=1}^{M} y_u = d, \;\; y_u \in \{0, 1\}, u = 1, \ldots, M \tag{12}$$

where $\widehat{w}_u^L$ and $\widehat{s}_{u,v}^L$, respectively, measure the locality preservation importance of $\mathbf{p}_u$ and the redundant local information between $\mathbf{p}_u$ and $\mathbf{p}_v$ on the entire training data Ξ. $\widehat{w}_u^R$ and $\widehat{s}_{u,v}^R$, respectively, measure the importance of $\mathbf{p}_u$ in preserving ordinal information and the redundant ordinal information between $\mathbf{p}_u$ and $\mathbf{p}_v$ on the labeled image set Ξ_{N_l}. Thus, maximizing $\widehat{J}$ can preserve both the intrinsic geometric structure of the whole training set Ξ and the ordinal aging pattern of the labeled image set Ξ_{N_l} well. Comparing (12) with (11), we can see that the only difference lies in the number of the images that are used for measuring the locality preservation importance of the feature and the redundant local information between features. Thus, our supervised learning method can be easily realized in a semi-supervised learning mode. The solution to this optimization problem (12) is the same as (11).

III. RANKING ON ORDINAL FEATURE REPRESENTATION

After finding the new low-dimensional aging feature representation $\{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_N\}, \widehat{\mathbf{x}}_i \in \mathbb{R}^d$, we take advantage of a ranking model to predict the age. In this paper, we adopt the ordinal hyperplanes ranker (OHRank) [27] model as the age estimator. Based on OHRank, we further propose a semi-supervised OHRank model.

A. OHRank

OHRank employs relative ordinal information among ages, and converts age estimation into a series of subproblems of binary classifications according to the ordinal property. In OHRank, age label l_i of each facial image is used as a rank order, $l_i \in \{1, \ldots, \mathcal{K}\}$, where $\mathcal{K}$ is the number of labels. Given a query: "Is the age of the face older than age k," the age estimation task becomes a binary classification problem to determine which face is older. The dataset is then separated into two subset, D_k^+ and D_k^-

$$D_k = \begin{cases} D_k^+ = \{(\widehat{\mathbf{x}}_i, +1) | l_i > k\} \\ D_k^- = \{(\widehat{\mathbf{x}}_i, -1) | l_i \leq k\}. \end{cases}$$

Based on the new data set D_k, OHRank uses a reweighted SVM to solve the kth bi-class classification

$$\min_{w_k, b_k, \xi} \quad \frac{1}{2}\|w_k\|_2^2 + C \sum_i c_k(i)\xi_i$$
$$\text{s.t.} \quad y_k(i)\left(w_k^T \phi_k(\widehat{\mathbf{x}}_i) + b_k\right) \geq 1 - \xi_i, \xi_i \geq 0, \;\; \forall_i \tag{13}$$

where $c_k(i)$ is the cost function of misclassifying $\widehat{\mathbf{x}}_i$ in the kth subproblem. We use an absolute cost function to represent $c_k(i)$, due to its better performance (please refer to [27] for the details of the cost function). $y_k(i) = +1$ if $\widehat{\mathbf{x}}_i \in D_k^+$ and $y_k(i) = -1$ if $\widehat{\mathbf{x}}_i \in D_k^-$. ϕ_k is an implicit mapping in the Hilbert space. w_k and b_k are the hyperplane parameters in the implicit feature space defined by ϕ_k.

Then the decision function $f_k(\widehat{\mathbf{x}})$, used to classify whether $\widehat{\mathbf{x}}$ is larger than age k, is defined as

$$f_k(\widehat{\mathbf{x}}) = w_k^T \phi_k(\widehat{\mathbf{x}}) + b_k. \tag{14}$$

After training a classifier for each age label k, the age l of any data point $\widehat{\mathbf{x}}$ is determined as follows:

$$l(\widehat{\mathbf{x}}) = 1 + \sum_{k=1}^{\mathcal{K}-1} [\![f_k(\widehat{\mathbf{x}}) > 0]\!]. \tag{15}$$

B. Semi-Supervised OHRank

In order to incorporate the information embedded in the unlabel images into the learning process of OHRank, we develop a semi-supervised OHRank. We name it semi-OHRank for short. In the semi-OHRank, we assume that the first N_l images in Ξ are labeled, and the rest N_u images are unlabeled. Following the framework of manifold regularization [40], the semi-supervised OHRank can be obtained by penalizing a regularization term for the kth classifier

$$\begin{aligned} \min_{w_k, b_k, \xi} \quad & \frac{1}{2}\|w_k\|_2^2 + C\sum_{i=1}^{N_l} c_k(i)\xi_i + \rho\|\mathbf{f}_k\|_{\mathcal{M}}^2 \\ \text{s.t.} \quad & y_k(i)(w_k^T\phi_k(\widehat{\mathbf{x}}_i) + b_k) \geq 1 - \xi_i, i = 1, \ldots, N_l \\ & \xi_i \geq 0, i = 1, \ldots, N_l \end{aligned} \tag{16}$$

where ρ is a trade-off parameter satisfying $\rho \geq 0$. $\|\mathbf{f}_k\|_{\mathcal{M}}^2$ is a manifold regularization term, and is defined as

$$\|\mathbf{f}_k\|_{\mathcal{M}}^2 = \sum_{i,j=1}^{N_l+N_u} W_{ij}\left(f_k(\widehat{\mathbf{x}}_i) - f_k(\widehat{\mathbf{x}}_j)\right)^2 = \mathbf{f}_k^T\mathbf{L}\mathbf{f}_k$$

where W_{ij} is the edge weight defined on a pair of original input data nodes $(\mathbf{x}_i, \mathbf{x}_j)$ of the adjacency graph. $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian, and $\mathbf{D}$ is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{N_l+N_u} W_{ij}$. $\mathbf{f}_k = [f_k(\widehat{\mathbf{x}}_1), \ldots, f_k(\widehat{\mathbf{x}}_{N_l+N_u})]^T$ denotes the decision function values for age label k over both the labeled and the unlabeled images, and here $f_k(\widehat{\mathbf{x}}) = w_k^T\phi_k(\widehat{\mathbf{x}}) + b_k$. In (16), minimizing the regularization term $\|w_k\|_2^2$ aims to maximize the margin of the kth bi-class classification. Different from $\|w_k\|_2^2$, minimizing the term $\|\mathbf{f}_k\|_{\mathcal{M}}^2$ plays the role of making the decision function $f_k(\mathbf{x})$ smooth on the manifold. In other words, minimizing $\|\mathbf{f}_k\|_{\mathcal{M}}^2$ is to make $f_k(\mathbf{x}_i)$ and $f_k(\mathbf{x}_j)$ close in the decision space if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the input space.

According to the representer theorem [41], w_k can be expressed as the expansion over Ξ

$$w_k = \sum_{i=1}^{N_l+N_u} \alpha_i^k \phi_k(\widehat{\mathbf{x}}_i) = \Phi_k\boldsymbol{\alpha}^k \tag{17}$$

where $\Phi_k = [\phi_k(\widehat{\mathbf{x}}_1), \ldots, \phi_k(\widehat{\mathbf{x}}_{N_l+N_u})]$, and $\boldsymbol{\alpha}^k = [\alpha_1^k, \ldots, \alpha_{N_l+N_u}^k]^T$. Based on (17), the decision function $f_k(\widehat{\mathbf{x}})$ can be rewritten as

$$f_k(\widehat{\mathbf{x}}) = \sum_{i=1}^{N_l+N_u} \alpha_i^k K_k(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}) + b_k \tag{18}$$

where $\mathbf{K}_k$ is the kernel matrix formed by kernel functions $K_k(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j) = \langle\phi_k(\widehat{\mathbf{x}}_i), \phi_k(\widehat{\mathbf{x}}_j)\rangle$.

Based on (17) and (18), we can rewrite the manifold regularization term $\|\mathbf{f}_k\|_{\mathcal{M}}^2$

$$\|\mathbf{f}_k\|_{\mathcal{M}}^2 = \mathbf{f}_k^T\mathbf{L}\mathbf{f}_k = (\boldsymbol{\alpha}^k)^T\mathbf{K}_k\mathbf{L}\mathbf{K}_k\boldsymbol{\alpha}^k. \tag{19}$$

TABLE II
SUMMARY OF EXPERIMENTAL DATASETS. FEATURE, #SIZE, #DIM, AND #AGE CLASS DENOTE THE EXTRACTED FEATURE TYPE, THE NUMBER OF SAMPLES, THE NUMBER OF FEATURES, AND THE NUMBER OF AGE CATEGORIES, RESPECTIVELY

Dataset	Feature	#Size	#Dim	# Age Class
FACES	BIF [45]	1,026	3,208	37
FG-NET	BIF [25]	1,002	3,208	11
Images of Group	Gist [46]	9,000	600	7

Substituting (17) and (19) into (16), the objective function (16) becomes

$$\begin{aligned} \min_{\boldsymbol{\alpha}^k, \xi} \quad & \frac{1}{2}(\boldsymbol{\alpha}^k)^T\mathbf{K}_k\boldsymbol{\alpha}^k + C\sum_{i=1}^{N_l} c_k(i)\xi_i + \rho(\boldsymbol{\alpha}^k)^T\mathbf{K}_k\mathbf{L}\mathbf{K}_k\boldsymbol{\alpha}^k \\ \text{s.t.} \quad & y_k(i)\left(\sum_{j=1}^{N_l+N_u} \alpha_j^k K_k(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j) + b_k\right) \geq 1 - \xi_i \\ & i = 1, \ldots, N_l \xi_i \geq 0, i = 1, \ldots, N_l. \end{aligned} \tag{20}$$

In order to efficiently solve the optimization problem (20), we introduce the following lemma.

Lemma 1: The dual problem of (20) can be written as

$$\begin{aligned} \max_{\boldsymbol{\beta}^k} \quad & \sum_{i=1}^{N_l} \beta_i^k - \frac{1}{2}\sum_{i,j=1}^{N_l} y_k(i)y_k(j)\beta_i^k\beta_j^k K_k^i\mathbf{A}_k^{-1}(K_k^j)^T \\ \text{s.t.} \quad & 0 \leq \beta_i^k \leq C \cdot c_k(i), i = 1, \ldots, N_l \\ & \sum_{i=1}^{N_l} \beta_i^k y_k(i) = 0 \end{aligned} \tag{21}$$

where $\mathbf{A}_k = (\mathbf{K}_k + 2\rho\mathbf{K}_k\mathbf{L}\mathbf{K}_k)$, and $\boldsymbol{\beta}^k = [\beta_1^k, \ldots, \beta_{N_l}^k]^T$. K_k^i is the ith row of the kernel matrix $\mathbf{K}_k$.

Lemma 1 can be easily verified using the Lagrange theory. We give the proof of Lemma 1 in the Appendix. The objective function (21) is a convex quadratic programming problem with linear constraints. Some standard algorithms, such as the conjugate gradient method, can be employed to solve this problem. Then we can obtain the decision function $f_k(\widehat{\mathbf{x}})$ by substituting $\boldsymbol{\alpha}^k$ into (18) (refer to [42] for the method of calculating b_k). For a new test image, we can easily determine its age by (15).

IV. EXPERIMENTAL RESULTS

In this section, we investigate the performance of the proposed method on three public available datasets: 1) the FACES dataset [43]; 2) the Images of Groups dataset [44]; and 3) the FG-NET aging dataset.[1] The details of the datasets used in the experiments are summarized in Table II.

To further evaluate the performance of the proposed feature selection method for age estimation, we compare it with six related feature selection approaches.

1) *FS-ED [47]:* It is designed for ranking, which aims to select a subset of highly ordinal discriminating features over relevance categories based on the expected divergence.

[1] http://www.fgnet.rsunit.com/

Fig. 2. Examples of facial aging images with different expressions. Each individual faces have significant expression changes (see each row) but with the same ground truth age (30 and 70 years old, respectively).

2) *LAR [20]:* It is a regression algorithm, which aims at finding some effective aging features for age estimation.
3) *RankBoost [14]:* Its goal is to select aging features to keep the ordinal information for each individual.
4) *Laplacian Score [35]:* It aims at preserving the underlying manifold structure in the selected feature space.
5) *Fisher Score [33]:* It selects those features to make the distances between data points of different classes are as large as possible, while the distances between data points of the same class are minimized.
6) *Adaboost [48]:* It learns a small number of weak classifiers, and boosts them iteratively into a strong classifier of higher accuracy.

We name our algorithm as PLO. We also verify the effectiveness of our semi-supervised feature selection method combined with the semi-supervised OHRank, which we call semi-PLO.

Following [30], we vary the number of the selected features from 10 to 150 with an incremental step of 10. In OHRank and semi-supervised OHRank, the RBF kernel is used with the default value of gamma, and the regularization parameter C is chosen by cross validation. The parameters η, λ, and ρ in our method are also chosen by cross validation. The number of the nearest neighbors K for constructing the neighborhood graph is set to 10 in the experiments. The performance of age estimation is measured by the mean absolute error (MAE), which can be calculated by: $\text{MAE} = \sum_{i=1}^{N_{\text{test}}} \mid \widehat{l}_i - l_i \mid /N_{\text{test}}$, where l_i is the ground truth age for the ith test image, and $\widehat{l}_i$ is the estimated age. N_{test} is the number of test images.

A. Experiments on the FACES Dataset

This dataset provides both age and facial expressions with ground truth labels, which is first introduced in [45] for studying the human age estimation under different facial expressions. It has two sets, and each set contains 171 individuals with six expressions (anger, happiness, disgust, neutrality, sadness, and fear) for each person in frontal view. Because the two sets are almost the same, we only use one set like [45]. Fig. 2 shows some examples taken from this dataset. As in [45], biologically-inspired features (BIF) [49] is extracted for representing facial images. We adopt fivefold cross validation test strategy, and report the average result.

Table III shows the best experimental results of different algorithms, as well as the corresponding optimal numbers of the selected features that are listed in the brackets of the first column. PLO achieves the best performance among all the feature selection methods. For example, our method brings about 3% relative deduction of MAEs over Rankboost that obtains the second best result in Table III. FS-ED and RankBoost are two ranking algorithms. Although they can preserve the ordinal information of the facial images, they ignore the local structure of the data. LAR is a regression algorithm. It can preserve the ordinal relationship of the data to some extent, but it does not consider preserving the local structure, either. Laplacian score is an unsupervised feature selection method, which aims to capture the intrinsic local structure of the data, but it does not take account of the ordinal information of the data. Fisher score and Adaboost are two classification methods. The proposed method attempts to simultaneously preserve the ordinal information and the local manifold structure of the facial images in the selected feature space, which is of great importance for age estimation.

TABLE III
MAEs COMPARISON OF DIFFERENT FEATURE SELECTION ALGORITHMS ON THE FACES DATASET

Method	MAE
Adaboost (150 dims) [48]	12.85
Laplacian Score (150dims) [35]	10.73
Fisher Score (100 dims) [33]	10.76
RankBoost (140 dims) [14]	8.38
LAR (140 dims) [20]	8.53
FS-ED (150 dims) [47]	10.61
PLO (150 dims)(Ours)	**8.16**

TABLE IV
MAEs COMPARISON OF PLO AND [45] ON "HAPPY," "NEUTRAL," AND FROM "HAPPY" TO "NEUTRAL"

Training	Testing	[45]	PLO	MAE Reduction Rate
Neutral	Neutral	8.14	**5.16**	31.1%
Happy	Happy	10.32	**6.31**	38.9%
Happy	Neutral	8.11	**5.49**	32.3%

We also make a comparison with [45] that studies age estimation under different facial expressions on this dataset. Guo and Wang [45] selected two different kinds of facial expressions in each group of experiments. Based on the experimental results, they arrive at a conclusion that there exists significant influence of expressions on age estimation. In [45], the best result is 8.11, which is obtained by using the "happy" expression data as the training data and the "neutral" expression data as the testing data. Following the experimental setting of [45], we applied PLO to only neutral, happy, and from happy to neutral, respectively. The results are listed in Table IV. From Table IV, we can see PLO significantly outperforms [45] under all the cases, which indicates that our method can effectively estimate facial ages under both the same expression and different expressions.

In order to investigate how the performance is influenced by the number of the selected features, we plot the curves of MAE versus the number of the selected features in Fig. 3. PLO outperforms the other methods under all the dimensions except Rankboost. PLO achieves better results than Rankboost under most of the dimensions, especially when the number of the selected number is large. We note that when the feature

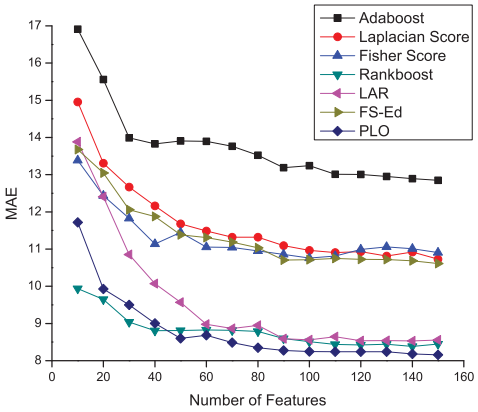


Fig. 3. MAE versus the number of the selected features on the FACES dataset.

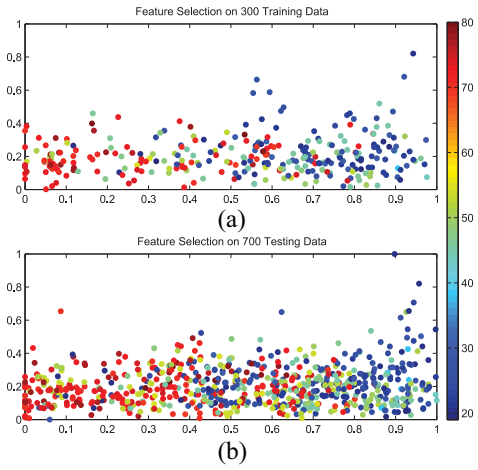


Fig. 4. 2-D visualization of (a) and (b) selecting two most important features by our feature selection method. The color of the point represents age with blue being the youngest and red being the oldest.

dimension is less than or equal to 40, PLO performs worse than Rankboost. This is because that the facial images lie on a higher dimensional manifold, while the lower dimensional feature subset cannot represent the manifold structure sufficiently, which degrades the performance of PLO. It is feasible to mitigate this effect by adjusting η to lower its weight in the objective function (11).

In addition, we perform a 2-D visualization to demonstrate the effectiveness of PLO. We first randomly select 300 images as the training data to find the first two features based on (12). Then all the training images are represented by the two features, and are plotted in Fig. 4(a). Similarly, we use the selected two features to represent the rest testing data, and plot them in Fig. 4(b). In Fig. 4, each point represents an individual. The color of the point represents age with blue being the youngest and red being the oldest. We can see that the red points and the blue points are mainly distributed in the left and right parts of Fig. 4(a) and (b), respectively. And most of the yellow points representing middle-aged individuals are scattered in the middle part of Fig. 4(a) and (b). Therefore, a rough age pattern or trend can be discovered just relying on two features extracted by our method.

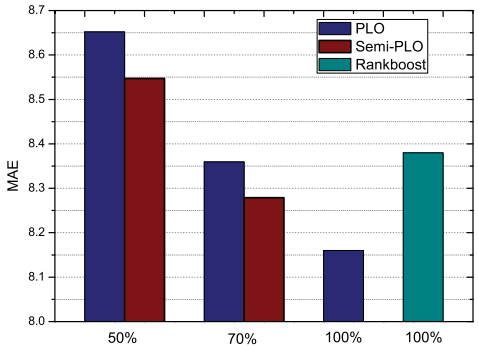


Fig. 5. MAEs of semi-PLO under different sizes of the labeled images.

Fig. 6. Examples of facial aging images in the Images of Groups dataset.

Finally, we evaluate the effectiveness of the proposed semi-PLO method. In this experiment, semi-PLO respectively randomly selects 50% and 70% images from the original training data mentioned above as the labeled images, and uses the rest training data as the unlabeled images. The testing data is kept unchanged. In order to verify its effectiveness, we also test the performance of PLO that uses the same labeled images with semi-PLO as its new training data. We set the number of the selected features to 150 which is the optimal number in PLO on this dataset. Fig. 5 shows the experimental result. By taking advantage of the underlying structure information of the unlabeled images, semi-PLO consistently outperforms PLO under different sizes of the labeled data. In addition, we note that only using 70% of the original training data as the labeled images, the result of semi-PLO is better than that of Rankboost using all the original training data, which further demonstrates the effectiveness of semi-PLO.

B. Experiments on the Images of Groups Dataset

Images of Groups dataset consists of 28 231 faces from 5080 Flickr images, which has been widely used for age range estimation [41]. Seven age categories are considered: 1) 0–2; 2) 3–7; 3) 8–12; 4) 13–19; 5) 20–36; 6) 37–65; and 7) 66+, roughly corresponding to different life stages, which are respectively labeled as $1, 2, \ldots, 7$. Each facial image is normalized to 61×49 pixels based on eye centers. Some typical aging face images in this dataset are shown in Fig. 6. Since the facial images are downloaded from web, the quality of many of them is extremely low. We thus pick out 9000 high quality images used for the experiments. We randomly select 7000/2000 images as the training/testing data. In the training data, 6000 images are randomly chosen as the labeled images, and the rest 1000 images are used as the unlabeled images. The 600-D gist features that the database itself provides are used to represent the facial images.[2]

[2]http://chenlab.ece.cornell.edu/people/Andy/ImagesOfGroups.html

TABLE V
MAEs COMPARISON OF DIFFERENT FEATURE SELECTION ALGORITHMS ON THE IMAGES OF GROUPS DATASET

Method	MAE
Adaboost+SVM (140 dims) [22]	0.997
Adaboost (130 dims)	0.949
Laplacian Score (150 dims)	0.926
Fisher Score (120 dims)	0.924
RankBoost (150 dims)	0.911
LAR (70 dims)	0.885
FS-ED(140 dims)	0.907
PLO (150 dims)(Ours)	**0.864**

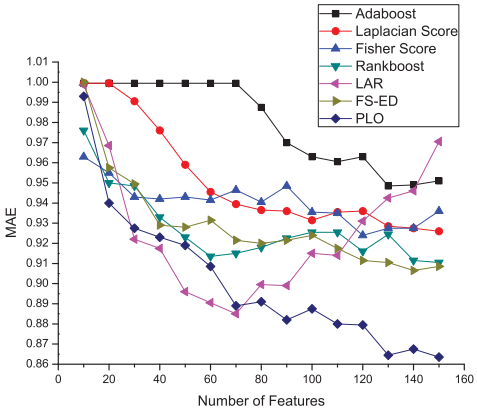


Fig. 7. MAE versus different feature numbers on the Images of Group dataset.

We first report the best result of each algorithm listed in Table V. PLO obtains better performance than all the other feature selection methods. We also compare our PLO with [22] that utilizes Adaboost combined with SVM to predict age on this dataset. Comparing with [22], PLO achieves 13.3% relative improvement. In addition, we also observe that PLO, RankBoost, LAR, and FS-ED have better performance than Laplacian Score, Fisher Score, and Adaboost. The reason is that the ranking or regression based algorithms take advantage of the ordinal relationship among aging faces. This shows that preserving ordinal information is important for age estimation.

We also investigate the influence of different numbers of the selected features on the performance of different algorithms for age estimation. The result is shown in Fig. 7. PLO outperforms all the other methods under almost all the dimensions except LAR. LAR reaches the best result quickly, which just needs 70 dimensions. Our method has the lowest estimation error, and outperforms LAR in most of the cases.

The images of this dataset are from web images, so it is meaningful to perform deep experimental analysis to the proposed method. In objective function (12), there are four components.

1) Only maximizing the first term of the numerator means preserving the local manifold structure.
2) Only maximizing the second term of the numerator intends to keep the ordinal information.
3) Only minimizing the first term of the denominator indicates that we select features by eliminating redundant local information.

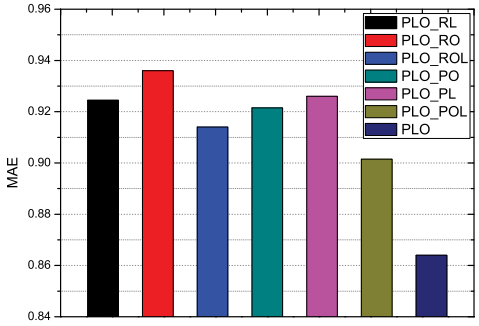


Fig. 8. Verify the effectiveness of each component in PLO on the Images of Group dataset.

4) Only minimizing the second term of the denominator aims to remove the features with redundant ordinal information.

We test the effectiveness of only preserving the locality information, only keeping the ordinal information, preserving both the two kinds of information with no consideration of removing the redundancy, only removing local information redundancy, only removing ordinal information redundancy, and only removing both kinds of information redundancy, which are respectively named PLO_PL, PLO_PO, PLO_POL, PLO_RL, PLO_RO, and PLO_ROL. The number of the selected features is set to 150 in the experiment.

The results are shown in Fig. 8. PLO_PO is superior to PLO_PL. It implies that keeping ordinal information is more important than preserving locality information for age estimation. Moreover, PLO_POL achieves better result than PLO_PL and PLO_PO. It means that simultaneously preserving both kinds of information is good for age estimation. PLO_RL obtains better performance than PLO_RO. It shows that eliminating local information redundancy is more important than eliminating ordinal information redundancy. Meanwhile, PLO_ROL outperforms PLO_RL and PLO_RO, which indicates that eliminating both kinds of information redundancy is beneficial for age estimation. In addition, PLO_POL outperforming PLO_ROL demonstrates that preserving both kinds of information is more important than eliminating redundant information. Finally, the combination of PLO_POL and PLO_ROL, i.e., PLO, significantly improves the performance, which shows that the combination is effective. In other words, the mixture of the four components causes a good chemical reaction.

We verify the effectiveness of the proposed semi-supervised feature selection method and semi-supervised age estimation algorithm semi-OHRank on this bigger dataset, respectively. The number of the selected features is set to 150. We perform two groups of experiments in this dataset: our supervised feature selection method and semi-OHRank are jointly used to estimate the age as the first experiment. We name it PLO_SO; we use the proposed semi-supervised feature selection method combined with semi-OHRank, i.e., semi-PLO, to predict the age as the second experiment. The result is plotted in Fig. 9. By integrating a manifold regularization into the OHRank, the result of PLO_SO is superior to that of PLO, which indicates that the proposed

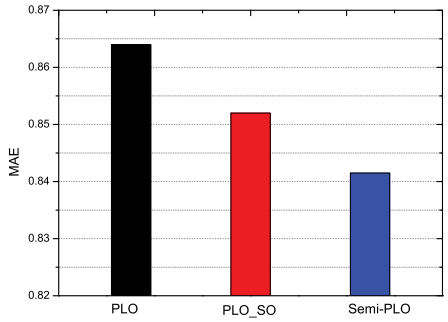


Fig. 9. Verify the effectiveness of both the proposed semi-supervised feature selection method and the semi-supervised age estimation algorithm semi-OHRank on the Images of Group dataset.

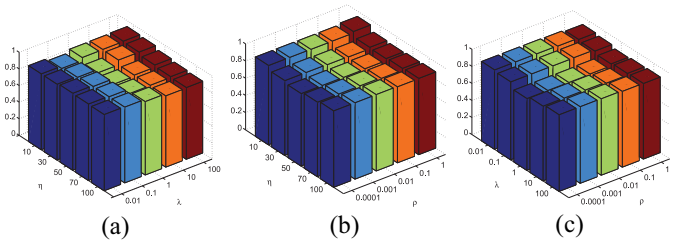


Fig. 10. MAEs of semi-PLO with different parameters on the Images of Groups dataset. (a) Vary η, λ, and fix ρ. (b) Vary η, ρ, and fix λ. (c) Vary λ, ρ, and fix η.

semi-OHRank is effective for age estimation. By incorporating the local structure information embedded in the unlabeled images into the process of feature selection, semi-PLO further improves the performance over PLO_SO, which shows that the proposed semi-supervised feature selection method is also effective.

In our method, there are three main parameters: η, λ, and ρ. We study the sensitiveness of these parameters in our algorithm. Fig. 10 shows the results. We can see that our method is not sensitive to η, λ, and ρ with wide ranges.

C. Experiments on the FG-NET Aging Dataset

The FG-NET aging dataset contains 1002 face images with large variations in pose, expression and lighting, which is a popular dataset for studying age estimation [16], [27], [30]. There are 82 subjects in total with the age ranges from 0 to 69 years old. Since the facial images with high ages in the collection are scarce, it is extremely challenging to build an accurate age estimator for predicting the exact age of old people. Therefore, we conduct age range estimation in this dataset. We divide the dataset into 11 age groups, and show the details in Table VI. We use BIF as in [25] to represent the facial images, and adopt fivefold cross validation to evaluate the performance in the experiment.

Table VII shows the best MAE results of different algorithms with the optimal numbers of the selected features. The optimal numbers of the selected features are listed in the brackets of the first column in Table VII. We can see that PLO obtains the lowest MAEs score. By leveraging the ordinal information with the locality information, PLO achieves

TABLE VI
AGE GROUP DISTRIBUTION IN THE FG-NET AGING DATASET

Age Group (year)	Image Number	Label
0-4	193	1
5-9	178	2
10-14	174	3
15-19	165	4
20-24	84	5
25-29	60	6
30-34	46	7
35-39	33	8
40-44	28	9
45-49	18	10
>= 50	23	11

TABLE VII
MAEs OF DIFFERENT ALGORITHMS ON THE FG-NET DATASET

Method	MAE
Adaboost (150 dims)	1.652
Laplacian Score (150dims)	1.654
Fisher Score (130 dims)	1.497
RankBoost (150 dims)	1.400
LAR (150 dims)	1.429
FS-ED(100 dims)	1.527
PLO (150 dims)(Ours)	**1.306**

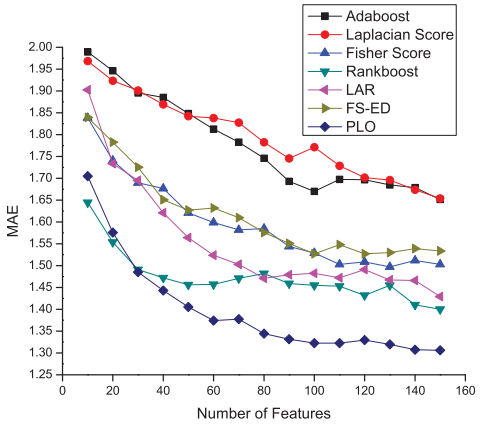


Fig. 11. MAE versus the number of the selected features on the FG-NET aging dataset.

6.7% relative error reduction over Rankboost that obtains the second best results in Table VII.

Fig. 11 reports the influence of the different numbers of the selected features on the age estimation. We can come to the same conclusion with Section IV-A.

In this dataset, we verify the improvement by fusing different kinds of facial features for age estimation on the FG-NET aging dataset. In the experiments, we extracted two kinds of features from each facial image: 3208-D BIF and 2048-D gist features. After that, we apply PLO to the two kinds of features, respectively. The reduced dimension is set to 150. At the stage of testing, we first estimate the ages l_{bif} and l_{gist} for each testing image $\mathbf{x}$ based on BIF and gist, respectively. Then we adopt a simple fusing strategy to calculate the final estimated age $l = \theta \cdot l_{\text{bif}} + (1-\theta) \cdot l_{\text{gist}}$, where θ is a trade-off parameter, and is set to 0.7 in the experiment.

TABLE VIII
RESULTS OF FEATURE FUSING

Feature	RankBoost	PLO	MAE Reduction Rate
BIF	1.400	**1.306**	6.7%
Gist	1.483	**1.418**	4.4%
BIF+Gist	1.335	**1.268**	5.0%

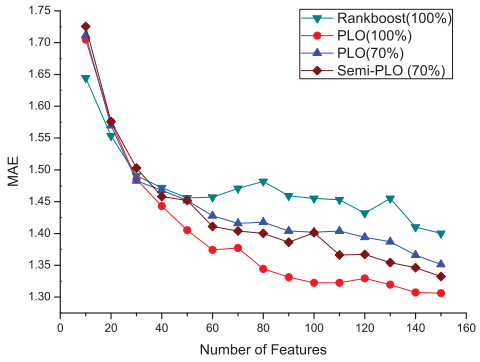


Fig. 12. MAEs of semi-PLO under different numbers of the selected features.

We also compare PLO with RankBoost to verify whether our method still outperforms other algorithms under different features or feature fusing. The experimental results are reported in Table VIII. From the second column of Table VIII, we can see that the performance of PLO is improved by fusing the two kinds of features. In addition, we can also see that PLO outperforms RankBoost under different facial features. After feature fusing, PLO still achieves better performance than RankBoost.

Finally, we test the effectiveness of semi-PLO under different feature numbers. Semi-PLO randomly selects 70% of the original training data as the labeled images, and uses the rest as the unlabeled images. The testing set is kept unchanged. Fig. 12 shows the result. By incorporating the structure information of the unlabeled images and adding a manifold regularization term into OHRank, semi-PLO outperforms PLO under almost all the dimensions, where PLO uses the same labeled images with semi-PLO as the new training data. In addition, when uses 70% of the original training data, PLO achieves better result than Rankboost using the original training data, which also demonstrates the power of PLO.

V. CONCLUSION

This paper presented two age estimation algorithms, called PLO and semi-PLO, from learning ordinal discriminate features perspective. Based on the observation that facial aging images lie on a local manifold and they are ordinal in temporal domain, PLO aims at preserving both the locality and the ordinal information simultaneously. Likewise, semi-PLO aims to select some powerful features that can preserve the locality information of the whole training data and the ordinal information of the labeled image set. In addition, by adding a manifold regularization into OHRank, the decision function acquired in semi-PLO is smooth on the manifold. The proposed methods were tested on three public datasets. Extensive experimental results showed that PLO outperformed the other feature selection methods for both exact age estimation and age range estimation. Semi-PLO could well deal with the practical cases where only a few labeled facial images were available.

APPENDIX

Proof of Lemma 1

Proof: In order to obtain Lemma 1, we first define the following Lagrangian equation:

$$\begin{aligned} & L\left(\boldsymbol{\alpha}^k, b^k, \boldsymbol{\xi}^k, \boldsymbol{\beta}^k, \boldsymbol{\delta}^k\right) \\ &= \frac{1}{2}\left(\boldsymbol{\alpha}^k\right)^T \mathbf{K}_k \boldsymbol{\alpha}^k + C\sum_{i=1}^{N_l} c_k(i)\xi_i^k + \rho\left(\boldsymbol{\alpha}^k\right)^T \mathbf{K}_k \mathbf{L} \mathbf{K}_k \boldsymbol{\alpha}^k \\ &\quad - \sum_{i=1}^{N_l} \beta_i^k \left(y_k(i)\left(\sum_{j=1}^{N_l+N_u} \alpha_j^k K_k\left(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j\right) + b^k \right) + \xi_i^k - 1 \right) \\ &\quad - \sum_{i=1}^{N_l} \delta_i^k \xi_i^k \end{aligned} \tag{22}$$

where $\boldsymbol{\beta}^k = [\beta_1^k, \ldots, \beta_{N_l}^k]$, and $\boldsymbol{\delta}^k = [\delta_1^k, \ldots, \delta_{N_l}^k]$. β_i^k and δ_i^k are the Lagrange multipliers satisfying $\beta_i^k \geq 0, \delta_i^k \geq 0$.

To derive the optimal solution, we conduct the following differentiations, and set them to zero:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{\alpha}^k} = \mathbf{K}_k \boldsymbol{\alpha}^k + 2\rho \mathbf{K}_k \mathbf{L} \mathbf{K}_k \boldsymbol{\alpha}^k - \sum_{i=1}^{N_l} \beta_i^k y_k(i) \left(K_k^i\right)^T = 0 \\ \dfrac{\partial L}{\partial b^k} = \sum_{i=1}^{N_l} \beta_i^k y_k(i) = 0 \\ \dfrac{\partial L}{\partial \xi_i^k} = C c_k(i) - \beta_i^k - \delta_i^k = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \boldsymbol{\alpha}^k = (\mathbf{K}_k + 2\rho \mathbf{K}_k \mathbf{L} \mathbf{K}_k)^{-1} \sum_{i=1}^{N_l} \beta_i^k y_k(i) (K_k^i)^T \\ \sum_{i=1}^{N_l} \beta_i^k y_k(i) = 0 \\ \delta_i^k = C c_k(i) - \beta_i^k \end{cases} \tag{23}$$

where K_k^i is the ith row of the kernel matrix $\mathbf{K}$.

Substituting (23) into (22), the dual function of (20) can be obtained as

$$\max_{\boldsymbol{\beta}^k} \sum_{i=1}^{N_l} \beta_i^k - \frac{1}{2} \sum_{i,j=1}^{N_l} y_k(i) y_k(j) \beta_i^k \beta_j^k K_k^i \mathbf{A}_k^{-1} (K_k^j)^T$$

where $\mathbf{A}_k = \mathbf{K}_k + 2\rho \mathbf{K}_k \mathbf{L} \mathbf{K}_k$.

Based on the conditions $\beta_i^k \geq 0$, $\delta_i^k \geq 0$, and $\delta_i^k = C c_k(i) - \beta_i^k$, we can obtain the following inequalities:

$$\begin{cases} C c_k(i) - \beta_i^k \geq 0 \\ \beta_i^k \geq 0 \end{cases} \Rightarrow 0 \leq \beta_i^k \leq C c_k(i).$$

Therefore, the final dual function can be written as

$$\max_{\boldsymbol{\beta}^k} \sum_{i=1}^{N_l} \beta_i^k - \frac{1}{2} \sum_{i,j=1}^{N_l} y_k(i) y_k(j) \beta_i^k \beta_j^k K_k^i \mathbf{A}_k^{-1} (K_k^j)^T$$

$$\text{s.t.} \quad 0 \leq \beta_i^k \leq C \cdot c_k(i), i = 1, \ldots, N_l; \ \sum_{i=1}^{N_l} \beta_i^k y_k(i) = 0.$$

The proof is complete. ■

REFERENCES

[1] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

[2] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 161–174, Feb. 2014.

[3] Y. Xu, X. Li, J. Yang, Z. Lai, and D. Zhang, "Integrating conventional and inverse representation for face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1738–1746, Oct. 2014.

[4] Y. Xu *et al.*, "Data uncertainty in face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1950–1961, Oct. 2014.

[5] Y. Zhang and D. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2622–2629.

[6] F. Dornaika and A. Bosaghzadeh, "Exponential local discriminant embedding and its application to face recognition," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 921–934, Mar. 2013.

[7] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 315–328, Mar. 2014.

[8] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[9] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[10] Z. Song, B. Ni, D. Guo, T. Sim, and S. Yan, "Learning universal multi-view age estimator using video contexts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 241–248.

[11] Y. Kwon and N. Lobo, "Age classification from facial images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 1994, pp. 762–767.

[12] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[14] P. Yang, L. Zhong, and D. Metaxas, "Ranking model for facial age estimation," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, 2010, pp. 3404–3407.

[15] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, USA, 2006, pp. 307–316.

[16] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[17] X. Geng, K. Smith-Miles, and Z.-H. Zhou, "Facial age estimation by nonlinear aging pattern subspace," in *Proc. ACM Int. Conf. Multimedia*, Vancouver, BC, Canada, 2008, pp. 721–724.

[18] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 657–664.

[19] Y. L. Chen and C. T. Hsu, "Subspace learning for facial age estimation via pairwise age ranking," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2164–2176, Dec. 2013.

[20] K. Ricanek, Y. Wang, C. Chen, and S. Simmons, "Generalized multi-ethnic face age-estimation," in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst.*, Washington, DC, USA, 2009, pp. 1–6.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–451, 2004.

[22] C. Shan, "Learning local features for age estimation on real-life faces," in *Proc. ACM Workshops Multimodal Pervasive Video Anal.*, Florence, Italy, 2010, pp. 23–28.

[23] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preference," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, Dec. 2003.

[24] K. Ueki, T. Hayashida, and T. Kobayashi, "Subspace-based age-group classification using facial images under various lighting conditions," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Southampton, U.K., Apr. 2006, pp. 43–48.

[25] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 112–119.

[26] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[27] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 585–592.

[28] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human age estimation based on face images," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 3396–3399.

[29] Y. Ma, T. Xiong, Y. Zou, and K. Wang, "Person-specific age estimation under ranking framework," in *Proc. ACM Int. Conf. Multimedia Retrieval*, Trento, Italy, 2011, Art. ID 38.

[30] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2570–2577.

[31] A. Appice, M. Ceci, S. Rawles, and P. Flach, "Redundant feature elimination for multi-class problems," in *Proc. Int. Conf. Mach. Learn.*, Banff, AB, Canada, Jul. 2004, pp. 33–40.

[32] 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.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

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

[34] Y. Liu, Y. Liu, and K. C. Chan, "Ordinal regression via manifold learning," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2011, pp. 398–403.

[35] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, Whistler, BC V0N, Canada, 2005, pp. 507–514.

[36] W. Liu and S.-F. Chang, "Robust multi-class transductive learning with graphs," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 381–388.

[37] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Edmonton, AB, Canada, 2002, pp. 133–142.

[38] M. Kendall, *Rank Correlation Methods*. New York, NY, USA: Oxford Univ. Press, 1990.

[39] Q. Wang, Y. Shen, and J. Q. Zhang, "A nonlinear correlation measure for multivariable data set," *Physica D*, vol. 200, nos. 3–4, pp. 287–295, 2005.

[40] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. 11, pp. 2399–2434, 2006.

[41] B. Scholkopf, R. Herbrich, A. J. Smola, and R. Williamson, "A generalized representer theorem," in *Proc. 14th Annu. Conf. Comput. Learn. Theory*, Amsterdam, The Netherlands, 2001, pp. 416–426.

[42] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *J. Data Min. Knowl. Disc.*, vol. 2, no. 2, pp. 121–167, 1998.

[43] N. Ebner, M. Riediger, and U. Lindenberger, "Faces—A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behav. Res. Methods*, vol. 42, no. 1, pp. 351–362, 2010.

[44] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 256–263.

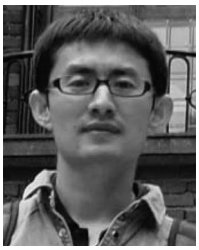
[45] G. Guo and X. Wang, "A study on human age estimation under facial expression changes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2547–2553.

[46] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[47] P. Gupta and P. Rosso, "Expected divergence based feature selection for learning to rank," in *Proc. Int. Conf. Comput. Linguist. (COLING)*, Mumbai, India, 2012, pp. 431–439.

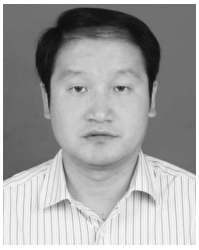
[48] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, 1999.

[49] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.



Changsheng Li received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008 and 2013, respectively.

He was a Research Assistant with Hong Kong Polytechnic University, Hong Kong, from 2009 to 2010. He joined IBM Research-China, Beijing, in 2013, where he is a Staff Researcher. His current research interests include machine learning and data mining.



Qingshan Liu (M'05–SM'07) received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, Beijing, China, and the M.S. degree from the Department of Auto Control, Southeast University, Nanjing, in 2003 and 2000, respectively.

He was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Science, and an Associate Researcher with the Multimedia Laboratory, Chinese University of Hong Kong, Hong Kong, from 2004 and 2005. He was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, Rutgers, the State University of New Jersey, New Brunswick, NJ, USA, from 2010 to 2011. He is a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing, China. His current research interests include image and vision analysis and machine learning.

Dr. Liu was the recipient of the President Scholarship of the Chinese Academy of Sciences in 2003.

Weishan Dong received the B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He also studied as a joint Ph.D. student in the School of Computer Science, The University of Birmingham, UK, from 2008 to 2009.

He joined IBM Research-China, Beijing, in 2009, where he is a Research Staff Member. He is a research team leader of big data processing, spatiotemporal analytics, and mobile computing. His current research interests include data mining, especially mining big spatiotemporal data (e.g., location data and mobility data) with addressing large scale and low latency, evolutionary computation, and computer vision topics. He also focuses on real-world applications of these technologies in IBM Smarter City solutions and products, including scalable connected vehicle information platform, crime analytics solution, asset management system, and business intelligence software. He has over 30 refereed publications in international journals and conferences and over 20 inventions/patent applications.

Xiaobin Zhu received the M.E. degree from Beijing Normal University, Beijing, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2006 and 2013, respectively.

He is currently with the Beijing Technology and Business University, Beijing. His current research interests include machine learning, video analysis, and object tracking.

Jing Liu (M'08) received the B.E. degree and the M.E. degree from Shandong University, Jinan, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2001, 2004, and 2008, respectively.

She is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include machine learning, image content analysis and classification, and multimedia information indexing and retrieval.

Hanqing Lu (SM'06) received the B.E. degree and the M.E. degree from Harbin Institute of Technology, Harbin, China, and the Ph.D. degree from the Huazhong University of Sciences and Technology, Wuhan, China, in 1982, 1985, and 1992, respectively.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include computer vision, object tracking, recognition, and image retrieval. He has published over 300 papers in the above areas.