2014 22nd International Conference on Pattern Recognition

Facial Age Estimation by Adaptive Label Distribution Learning

Xin Geng*, Qin Wang, and Yu Xia
School of Computer Science and Engineering
Key Lab of Computer Network and Information Integration, Ministry of Education
Southeast University, Nanjing 211189, China
Email: {xgeng, qinwang, xiayu}@seu.edu.cn

***Abstract*—Lack of sufficient and complete training data is one of the most prominent challenges in the problem of facial age estimation. Due to appearance similarity of the faces at close ages, the face images at the neighboring ages may be utilized while learning a particular age. As a result, the training images for each age are boosted without actually increase the total number of training images. This is achieved by assigning a label distribution instead of a single label of the chronological age to each face image. The label distribution should accord with the tendency of facial aging, which might be significantly different at different ages, e.g., the facial appearance during childhood and senior age generally changes faster than that during middle age. In this paper, two adaptive label distribution learning (ALDL) algorithms, IIS-ALDL and BFGS-ALDL, are proposed to automatically learn the label distributions adapted to different ages. Experimental results show that the ALDL algorithms perform remarkably better than the compared state-of-the-art algorithms.**

I. Introduction

Facial age estimation has recently become an active research topic in pattern recognition. Inspired by the vast potential applications in age-specific human-computer interaction, electronic customer relationship management, security control, surveillance monitoring, etc [7], many algorithms have been proposed for facial age estimation in recent years. For example, Lanitis et al. [16], [15] represented the aging pattern by a quadratic function and proposed the Weighted Appearance Specific (WAS) method and the Appearance and Age Specific (AAS) method. Geng et al. [9], [8] proposed the AGES algorithm, which learns a subspace from the aging pattern vectors. Fu et al. [6], [5] applied multiple linear regression to the discriminative aging manifold. Guo et al. [12] proposed a locally adjusted robust regressor for age estimation. Yan et al. [24] used semidefinite programming to solve a regression problem with nonnegative label intervals for age estimation. Ni et al. [18] proposed a robust multi-instance regression method to model the age estimator. Guo et al. applied the Biologically Inspired Features (BIF) [14] and the Kernel Partial Least Squares (KPLS) [13] regression to age estimation. Chang et al. [3] transformed an age estimation task into multiple cost-sensitive binary classification subproblems, and solved the problem with the Ordinal Hyperplane Ranking (OHRank) algorithm. Chen et al. [4] used a cumulative attribute to learn a regressor for age estimation.

Although a lot of algorithms have been successfully applied to facial age estimation, many challenges still remain in this problem. Perhaps one of the most prominent challenges is the lack of sufficient and complete training data due to difficulties in collecting data over a large span of time. Fortunately, the faces at close ages look quite similar since aging is a slow and gradual process. This prompts us to utilize the face images at neighboring ages when modeling a particular age. As described in our previous work [10], [11], the idea can be implemented by assigning a *label distribution* (LD) rather than a single label of the chronological age to each face image. The label distribution covers a certain number of neighboring ages, representing the degree that each age describes the corresponding face image. As a result, each face image can contribute to the learning of not only its chronological age, but also the neighboring ages. In such way, the training images for each age are boosted without actually increase the total number of training images.

However, the label distribution for each face image is not available in the original training data and must be generated with proper assumption and training. It has been observed that the facial aging process could be significantly different at different aging stages. From birth to adulthood, the main change is the craniofacial growth [21], during which the face size gets larger and the head shape changes gradually. From adulthood to senior age, the main change is the skin aging [1], during which the skin becomes darker and wrinklier. The speed of facial aging is also variable. Generally speaking, the change of facial appearance during the childhood and senior age is more apparent than that during the middle age. As an example, Fig. 1 shows the face images of Albert Einstein in three age ranges, i.e., 0-20, 20-50, and 50-76, which roughly correspond to three aging stages including childhood, middle age and senior age, respectively. As can be seen, Einstein's face changes more apparently in 0-20 and 50-76 than in 30-50, which indicates that the facial aging speed during his childhood and senior age is faster than that during his middle age. How to generate label distributions that accord with such tendency of facial aging is crucial to achieve good performance. This requires the algorithm to be able to learn the label distributions adapted to different ages. For this purpose, two adaptive label distribution learning (ALDL) algorithms, IIS-ALDL and BFGS-ALDL, are proposed in this paper to automatically learn a proper label distribution for each age, and estimate the facial age based on the adapted label distributions.

The rest of the paper is organized as follows. In Section II, the concept of adaptive label distribution learning is introduced and two ALDL algorithms for facial age estimation is proposed. After that, the experimental results are reported

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.764

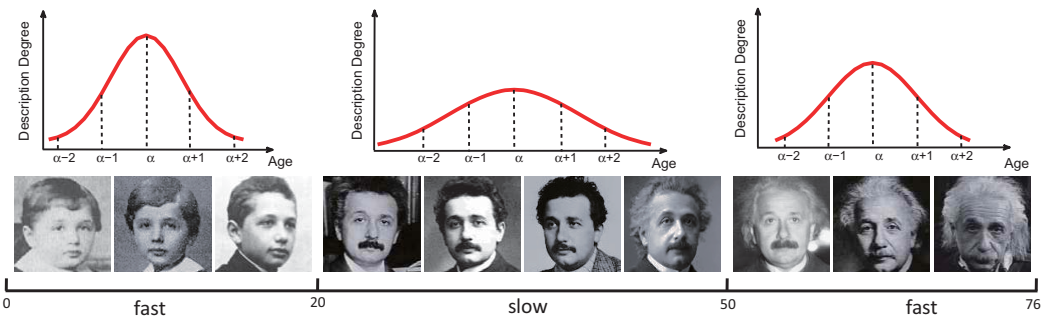


Fig. 1: Facial Appearance Aging of Albert Einstein in Different Age Ranges

and analyzed in Section III. Finally, conclusions are drawn in Section IV.

II. Adaptive Label Distribution Learning

A. Problem Formulation

For a face image $\boldsymbol{x}$, its label distribution is defined as a vector containing the description degrees of a certain number of neighboring ages. The description degree of the age y is a real number $d_{y,\boldsymbol{x}} \in [0,1]$ representing the degree to which y describes $\boldsymbol{x}$. For a face image $\boldsymbol{x}$, the description degrees of all ages involved in the label distribution sum up to 1, i.e., $\sum_y d_{y,\boldsymbol{x}} = 1$, which means that using all ages can always fully describe the face image.

The label distribution of a face image at the chronological age α should satisfy the following two properties. First, the description degree of α is the highest in the label distribution. Second, the description degrees of the neighboring ages decrease with the increase of the distance away from α. Among many possible choices satisfying the above two properties, the label distribution following a discretized Gaussian distribution centered at the chronological age α might be a suitable one, i.e.,

$$d_{y,\boldsymbol{x}} = \frac{1}{\sigma\sqrt{2\pi}Z}\exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right), \tag{1}$$

where σ is the standard deviation of the Gaussian distribution, and Z is a normalization factor that makes sure $\sum_y d_{y,\boldsymbol{x}} = 1$, i.e.,

$$Z = \frac{1}{\sigma\sqrt{2\pi}}\sum_y \exp\left(-\frac{(y-\alpha)^2}{2\sigma^2}\right). \tag{2}$$

Fig. 1 also shows some examples of the label distributions generated by Eq. (1) above the face images. The standard deviation σ controls the shape of the label distribution at each age. On the one hand, for those ages where facial appearance changes faster, e.g., the ages within 0-20 or 50-76 in Fig. 1, the neighboring ages can contribute less to the learning of the chronological age. Thus, σ should be smaller, resulting in sharper label distributions. On the other hand, for those ages where facial appearance changes slower, e.g., the ages within 20-50 in Fig. 1, the neighboring ages can contribute more to the learning of the chronological age. Thus, σ should be larger, resulting in smoother label distributions. In this sense, label distribution adaptation means finding proper σ for each age.

Label distribution shares the same properties with probability distribution, i.e., $d_{y,\boldsymbol{x}} \in [0,1]$ and $\sum_y d_{y,\boldsymbol{x}} = 1$, thus, many theories and methods from statistics can be borrowed to deal with label distributions. For example, the description degree $d_{y,\boldsymbol{x}}$ can be represented by the form of conditional probability mass function (p.m.f.), i.e., $d_{y,\boldsymbol{x}} = p(y|\boldsymbol{x})$. Then, the problem of adaptive label distribution learning for facial age estimation can be formulated as:

> Let $\mathcal{X} = \mathbb{R}^q$ denote the image feature space and $\mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$ denote the finite set of possible ages. Given a training set $S = \{(\boldsymbol{x}_1, \alpha_1), (\boldsymbol{x}_2, \alpha_2), \cdots, (\boldsymbol{x}_n, \alpha_n)\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the feature vector of a face image, $\alpha_i \in \mathcal{Y}$ is the chronological age of $\boldsymbol{x}_i$. The goal of adaptive label distribution learning is to learn a conditional probability mass function $p(y|\boldsymbol{x})$ ($\boldsymbol{x} \in \mathcal{X}$, $y \in \mathcal{Y}$) from S, which can generate a label distribution adapted to the chronological age of the given face image $\boldsymbol{x}$.

Suppose $p(y|\boldsymbol{x})$ is a parametric model $p(y|\boldsymbol{x};\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ as the parameter vector, and the label distributions of the training face images are defined by Eq. (1). Given the training set S, the goal of ALDL is to find the $\boldsymbol{\theta}$ that can generate label distributions similar to those created by Eq. (1), where the standard deviation is adapted to the chronological age of each training image.

B. ALDL algorithms

According to the formulation of ALDL given in Section II-A, there are two targets to be optimized. One is the parameter vector $\boldsymbol{\theta}$ in the conditional probability mass function. The other is the standard deviation σ in Eq. (1). These two targets can be optimized alternatively, as the flowchart shown in Fig. 2. The original training set is first initialized by Eq. (1) with the same standard deviation at every age, i.e., $\sigma_\alpha^0 = \sigma^0, \forall\alpha \in \mathcal{Y}$, where σ^0 is a predefined initial standard deviation, say, $\sigma^0 = 3$. After initialization, each training image $\boldsymbol{x}_i$ is associated with an initial label distribution $\boldsymbol{D}_i^0 = [d_{y_1,\boldsymbol{x}_i}^0, d_{y_2,\boldsymbol{x}_i}^0, \cdots, d_{y_c,\boldsymbol{x}_i}^0]$, which is calculated by

$$d_{y_j,\boldsymbol{x}_i}^0 = \frac{1}{\sigma\sqrt{2\pi}Z}\exp\left(-\frac{(y_j-\alpha_i)^2}{2(\sigma_{\alpha_i}^0)^2}\right), j = 1, 2, \ldots, c. \tag{3}$$

Next, in each outer iteration k, i.e., the LD adaptation iteration, the first step is to find a $\boldsymbol{\theta}$ that can generate label distributions most similar to $\boldsymbol{D}_i^{k-1}$. If the Kullback-Leibler divergence is used to measure the similarity between

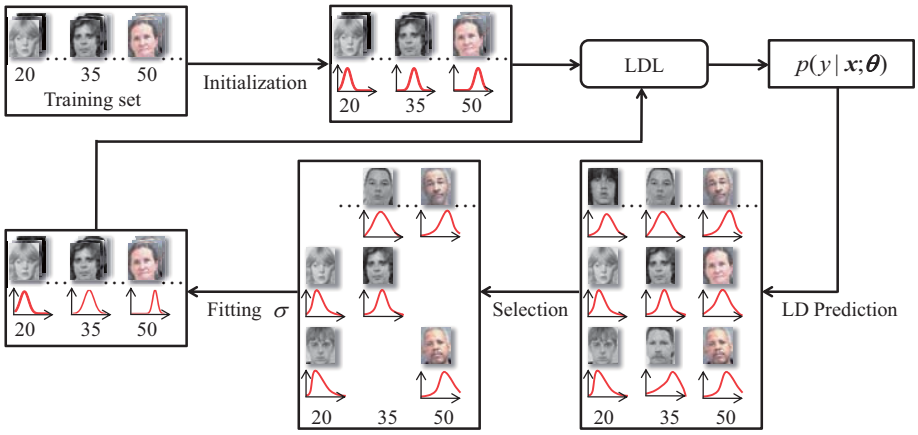


Fig. 2: The flowchart of the ALDL algorithms.

distributions, then the best parameter vector in iteration k is determined by

$$\begin{aligned}\boldsymbol{\theta}^k &= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i,j} d_{y_j,\boldsymbol{x}_i}^{k-1} \ln \frac{d_{y_j,\boldsymbol{x}_i}^{k-1}}{p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta})} \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i,j} d_{y_j,\boldsymbol{x}_i}^{k-1} \ln p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta}). \end{aligned} \tag{4}$$

Assume $p(y|\boldsymbol{x};\boldsymbol{\theta})$ to be a maximum entropy model [2], i.e.,

$$p(y_j|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\Lambda} \exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r) \tag{5}$$

where $\Lambda = \sum_j \exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r)$ is the normalization factor, $\boldsymbol{x}_i^r$ is the r-th feature of $\boldsymbol{x}_i$, and $\boldsymbol{\theta}_{y_j,r}$ is an element in $\boldsymbol{\theta}$ corresponding to the age y_j and the r-th feature. Substituting Eq. (5) into Eq. (4) yields

$$\begin{aligned}\boldsymbol{\theta}^k &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i,j} d_{y_j,\boldsymbol{x}_i}^{k-1} \sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r - \\ &\quad \sum_i \ln \sum_j \exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r). \end{aligned} \tag{6}$$

The maximization problem of Eq. (6) can be solved by a strategy similar to Improved Iterative scaling (IIS) [20]. The $\boldsymbol{\theta}$ is initialized by the result of the last outer iteration $k-1$, i.e., $\boldsymbol{\theta}^{k,0} = \boldsymbol{\theta}^{k-1}$, and $\boldsymbol{\theta}^0$ is a predefined initial parameter vector, say, a zero vector. Then, in each inner iteration l of the optimization, it updates the current estimate of parameters $\boldsymbol{\theta}^{k,l-1}$ to $\boldsymbol{\theta}^{k,l} = \boldsymbol{\theta}^{k,l-1} + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ maximizes a lower bound to the change of the target function. As proved in [11], the element of $\boldsymbol{\Delta}$, $\delta_{y_j,r}$, can be obtained by solving the equation

$$\begin{aligned}\sum_i p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta}^{k,l-1}) \boldsymbol{x}_i^r \exp(\delta_{y_j,r} s(\boldsymbol{x}_i^r) \#(\boldsymbol{x}_i)) \\ - \sum_i d_{y_j,\boldsymbol{x}_i}^{k-1} \boldsymbol{x}_i^r = 0, \end{aligned} \tag{7}$$

where $\#(\boldsymbol{x}_i) = \sum_r |\boldsymbol{x}_i^r|$ and $s(\boldsymbol{x}_i^r)$ is the sign of $\boldsymbol{x}_i^r$. Eq. (7) can be solved by nonlinear equation solvers, such as the Gauss-Newton method. This process is called label distribution learning (LDL) in our previous work [10], [11].

After the LDL step, the conditional p.m.f. $p(y|\boldsymbol{x};\boldsymbol{\theta}^k)$ is obtained. Thus, the label distribution of each training image $\boldsymbol{x}_i$ can be predicted as $p(y|\boldsymbol{x}_i;\boldsymbol{\theta}^k)$. According to the predicted label distribution, the age of $\boldsymbol{x}_i$ can be estimated as $\hat{\alpha}_i = \operatorname{argmax}_y p(y|\boldsymbol{x}_i;\boldsymbol{\theta}^k)$. The absolute error of each age estimate is then calculated by $e_i = |\alpha_i - \hat{\alpha}_i|$, and those predicted label distributions with an absolute age estimation error lower than the MAE (mean absolute error) of the whole training set, i.e., $MAE = \frac{1}{n}\sum_i e_i$, are selected as the training set for fitting the standard deviation σ to each age.

The selected label distribution predictions are divided into c subsets by their chronological ages, i.e., the examples in each subset have the same chronological age. Suppose the image index set corresponding to the subset of the age α is denoted by I_α, and the label distributions of the images indexed by I_α are all generated with the same standard deviation σ_α, i.e.,

$$d_{y_j,\boldsymbol{x}_m} = \frac{1}{\sigma_\alpha \sqrt{2\pi} Z} \exp\left(-\frac{(y_j-\alpha)^2}{2\sigma_\alpha^2}\right), \tag{8}$$

where $m \in I_\alpha$ and $j = 1, 2, \ldots, c$. Then the best σ_α should be the one that can generate a label distribution most similar to the predicted label distributions in the subset. If the Kullback-Leibler divergence is again used to measure the similarity between two distributions, then

$$\begin{aligned}\sigma_\alpha^k &= \operatorname*{argmin}_{\sigma_\alpha} \sum_{m \in I_\alpha} \sum_j d_{y_j,\boldsymbol{x}_m} \ln \frac{d_{y_j,\boldsymbol{x}_m}}{p(y_j|\boldsymbol{x}_m;\boldsymbol{\theta}^k)}, \\ &\quad s.t.\ \ \sigma_\alpha > 0. \end{aligned} \tag{9}$$

Substituting Eq.(8) into Eq.(9) yields a nonlinear programming problem, which can be effectively solved by the log barrier interior-point method [23].

After σ_α^k is determined for every age in $\mathcal{Y}$, the label distributions of all training images are updated with the new standard deviation σ_α^k to get $\boldsymbol{D}_i^k$. Finally, the training set with the updated label distributions $\boldsymbol{D}_i^k$ is sent into the LDL step again to start the next outer iteration $k+1$. The whole process repeats until the MAE difference between two adjacent iterations is less than a predefined threshold ε. This algorithm is denoted by IIS-ALDL.

In the iterations of IIS-ALDL, the LDL step based on the optimization similar to IIS is most time-consuming. It has been reported in the literature [17] that IIS often performs less efficiently than several other optimization algorithms such as conjugate gradient and quasi-Newton methods. Here we follow the idea of a quasi-Newton method BFGS [19] to further improve IIS-ALDL. In detail, denote the negative target function of the optimization in Eq. (6) as

$$T(\boldsymbol{\theta}) = \sum_i \ln \sum_j \exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r) - \sum_{i,j} d_{y_j,\boldsymbol{x}_i}^{k-1} \sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r. \quad (10)$$

In the inner iteration l, consider the second-order Taylor series of $T(\boldsymbol{\theta})$ at the current estimate of the parameter vector $\boldsymbol{\theta}^{k,l-1}$:

$$T(\boldsymbol{\theta}) \approx T(\boldsymbol{\theta}^{k,l-1}) + (\nabla T(\boldsymbol{\theta}^{k,l-1}))' \boldsymbol{\Delta} + \frac{1}{2} \boldsymbol{\Delta}' \boldsymbol{H}(\boldsymbol{\theta}^{k,l-1}) \boldsymbol{\Delta}, \quad (11)$$

where $\Delta = \boldsymbol{\theta} - \boldsymbol{\theta}^{k,l-1}$, $\nabla T(\boldsymbol{\theta}^{k,l-1})$ and $\boldsymbol{H}(\boldsymbol{\theta}^{k,l-1})$ are the gradient and Hessian matrix of $T(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{k,l-1}$, respectively. The minimizer of Eq.(11) is

$$\boldsymbol{\Delta}^l = -\boldsymbol{H}^{-1}(\boldsymbol{\theta}^{k,l-1}) \nabla T(\boldsymbol{\theta}^{k,l-1}) \quad (12)$$

The line search Newton method uses $\boldsymbol{\Delta}^l$ as the search direction $\boldsymbol{p}^l = \boldsymbol{\Delta}^l$ and updates the parameter vector by

$$\boldsymbol{\theta}^{k,l} = \boldsymbol{\theta}^{k,l-1} + \alpha^l \boldsymbol{p}^l \quad (13)$$

where the step length α^l is obtained from a line search procedure to satisfy the *strong Wolfe conditions* [19]:

$$T(\boldsymbol{\theta}^{k,l-1} + \alpha^l \boldsymbol{p}^l) \leq T(\boldsymbol{\theta}^{k,l-1}) + c_1 \alpha^l (\nabla T(\boldsymbol{\theta}^{k,l-1}))' \boldsymbol{p}^l, \quad (14)$$

$$|(\nabla T(\boldsymbol{\theta}^{k,l-1} + \alpha^l \boldsymbol{p}^l))' \boldsymbol{p}^l| \leq c_2 |(\nabla T(\boldsymbol{\theta}^{k,l-1}))' \boldsymbol{p}^l|. \quad (15)$$

where $0 < c_1 < c_2 < 1$. The idea of BFGS is to avoid explicit calculation of $\boldsymbol{H}^{-1}(\boldsymbol{\theta}^{k,l-1})$ by approximating it with an iteratively updated matrix $\boldsymbol{B}$, i.e.,

$$\boldsymbol{B}^l = (\boldsymbol{I} - \boldsymbol{\rho}^l \boldsymbol{s}^l (\boldsymbol{u}^l)') \boldsymbol{B}^{l-1} (\boldsymbol{I} - \boldsymbol{\rho}^l \boldsymbol{u}^l (\boldsymbol{s}^l)') + \boldsymbol{\rho}^l \boldsymbol{s}^l (\boldsymbol{s}^l)', \quad (16)$$

where

$$\boldsymbol{s}^l = \boldsymbol{\theta}^{k,l} - \boldsymbol{\theta}^{k,l-1}, \quad (17)$$

$$\boldsymbol{u}^l = \nabla T(\boldsymbol{\theta}^{k,l}) - \nabla T(\boldsymbol{\theta}^{k,l-1}), \quad (18)$$

$$\boldsymbol{\rho}^l = \frac{1}{\boldsymbol{s}^l \boldsymbol{u}^l}. \quad (19)$$

Thus, the computation of BFGS is mainly related to the first-order gradient, which can be obtained through

$$\frac{\partial T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{y_j,r}} = \sum_i \frac{\exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r) \boldsymbol{x}_i^r}{\sum_j \exp(\sum_r \boldsymbol{\theta}_{y_j,r} \boldsymbol{x}_i^r)} - \sum_i d_{y_j,\boldsymbol{x}_i}^{k-1} \boldsymbol{x}_i^r \quad (20)$$

Based on previous studies [17], the BFGS-based ALDL algorithm stands a good chance of outperforming the IIS-based algorithm IIS-ALDL. This improved algorithm is denoted by BFGS-ALDL.

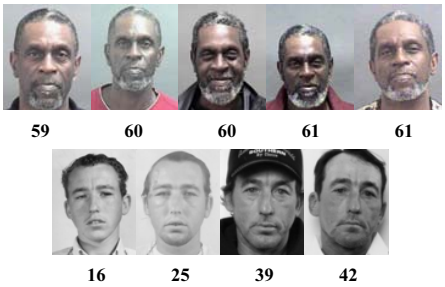


Fig. 3: Typical Aging Faces in the MORPH Database

III. Experiments

A. Methodology

The data set used in the experiments is the MORPH database [22]. It is one of the largest aging face databases publicly available, which contains $55,132$ face images from more than $13,000$ subjects. In average, there are 4 face images per subject. The ages of the face images range from 16 to 77 with a median age of 33. MORPH is a multi-ethnic database, where about 77% are African faces, 19% are European faces, and the remaining 4% includes Hispanic, Asian, Indian and other races. Some typical aging faces in this database are shown in Fig. 3. The features extracted from the face images are the Biologically Inspired Features (BIF) [14]. The dimensionality of the BIF vectors is further reduced to 200 by Marginal Fisher Analysis (MFA) [25].

For the ALDL algorithms (BFGS-ALDL and IIS-ALDL), the label distribution of each face image is initialized as a discretized Gaussian distribution (Eq. (1)) with its mean at the chronological age and standard deviation of 3. The convergence threshold ε for both BFGS-ALDL and IIS-ALDL is set to 0.02. After the optimal model parameter vector $\boldsymbol{\theta}^*$ is learned from the training set, the predicted age for a test image $\boldsymbol{x}'$ is determined by $y^* = \operatorname{argmax}_y p(y|\boldsymbol{x}'; \boldsymbol{\theta}^*)$. In order to show the effectiveness of label distribution adaptation, the non-adaptive version of BFGS-ALDL and IIS-ALDL, denoted by BFGS-LDL and IIS-LDL, are also tested, which directly apply the LDL step to the initialized label distribution without further adaptation.

Several existing algorithms for facial age estimation are compared as the baseline methods, which include KPLS [13], OHRank [3], AGES [8], WAS [16], and AAS [15]. Some conventional general-purpose classification methods are also compared, which include kNN (k-Nearest Neighbors), BP (Backpropagation neural network), C4.5 (C4.5 decision tree), and SVM (Support Vector Machine). For all the compared algorithms, several parameter configurations are tested and the best results are reported. KPLS uses the RBF kernel with the inverse width of 1. For OHRank, the absolute cost function and the RBF kernel are used. For AGES, the aging pattern subspace dimensionality is set to 20. In AAS, the error threshold in the appearance cluster training step is set to 3. For kNN, k is set to 30 and Euclidean distance is used to find the neighbors. The BP neural network has a hidden layer of 100 neurons with sigmoid activation functions. The parameters of C4.5 are set to the default values of the J4.8 implementation. SVM is implemented as the 'C-SVC' type in LIBSVM using the RBF

TABLE I: MAE of the Algorithms

Method	MAE
BFGS-ALDL	4.34 ± 0.10
IIS-ALDL	4.43 ± 0.05
BFGS-LDL	4.78 ± 0.08
IIS-LDL	5.67 ± 0.15
KPLS	6.15 ± 0.12
OHRank	6.28 ± 0.18
AGES	6.61 ± 0.11
WAS	9.21 ± 0.16
AAS	10.10 ± 0.26
kNN	9.64 ± 0.24
BP	12.59 ± 1.38
C4.5	7.48 ± 0.12
SVM	7.34 ± 0.17

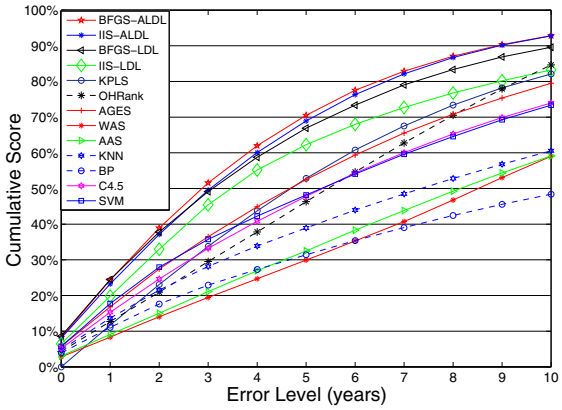


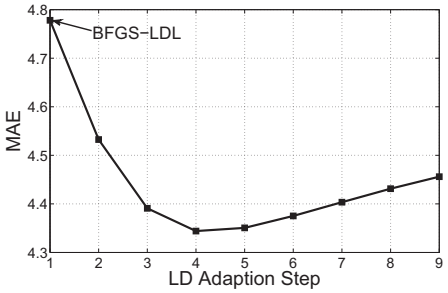
Fig. 4: CS of the Algorithms at the Error Levels 0-10

kernel with the inverse width of 1.

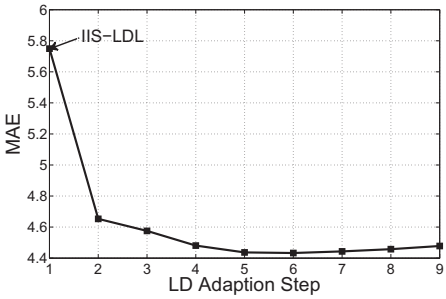
As suggested in [8], the performance of the age estimation algorithms is measured by MAE (Mean Absolute Error), i.e., the average absolute difference between the estimated age and the chronological age, as well as CS (Cumulative Score) at the error levels 0-10, i.e., the accuracy rate when "correct estimation" is defined as the estimation with an absolute error no higher than the error level. The algorithms are tested through the 10-fold cross validation. The average performance and the standard deviation over the 10 folds are recorded as the experimental results.

B. Results

Table I shows the MAE of all compared algorithms, and Fig. 4 shows the CS curves of the algorithms at the error levels from 0 to 10. As can be seen, the ALDL algorithms (BFGS-ALDL and IIS-ALDL) perform significantly better than all other compared algorithms. Specially, BFGS-ALDL and IIS-ALDL perform remarkably better than their non-adaptive versions BFGS-LDL and IIS-LDL, respectively. In detail, the average MAE of BFGS-ALDL is 9.2% lower than that of BFGS-LDL, and the average MAE of IIS-ALDL is 21.9% lower than that of IIS-LDL. The CS of BFGS-ALDL is constantly higher than that of BFGS-LDL. Their performance difference is even greater at the higher error levels. Similar result can be observed for the CS curves of



(a) BFGS-ALDL



(b) IIS-ALDL

Fig. 5: MAE Variation of the ALDL Algorithms w.r.t. the LD Adaptation Steps

IIS-ALDL vs. IIS-LDL, and the superiority of IIS-ALDL over IIS-LDL is even greater. This illustrates the effectiveness of the label distribution adaptation process in the ALDL algorithms. Also, the comparison between the BFGS-based algorithms (BFGS-ALDL and BFGS-LDL) and the IIS-based algorithms (IIS-ALDL and IIS-LDL) shows clear advantage of the former. In addition, the computational complexity of the BFGS-based algorithms is much less than that of the IIS-based algorithms, e.g., the time consumption of BFGS-ALDL is around $1/9$ of that of IIS-ALDL in the experiment. It is also worthwhile to mention that in either the MAE or CS results, there is a noticeable performance gap between the four label-distribution-based methods (BFGS-ALDL, IIS-ALDL, BFGS-LDL, and IIS-LDL) and the rest single-label-based methods. This illustrates the effectiveness of using label distributions to alleviate the problem of insufficient and incomplete training data.

To reveal the effects of the LD adaptation step (i.e., the outer iteration k in the ALDL algorithms) on the performance of the ALDL algorithms, Fig. 5 shows the MAE variation of the ALDL algorithms with respect to the LD adaptation steps. Note that the starting point for BFGS-ALDL is BFGS-LDL, while that for IIS-ALDL is IIS-LDL. The MAE variation curves of BFGS-ALDL and IIS-ALDL appear to be similar. Both of them drop rapidly within the first several LD adaptation steps. After that, the performance remain relatively steady before getting worse with too many LD adaptation steps. This reveals that the ALDL algorithms usually converge fast, and too many LD adaptation steps may cause overfitting.

To illustrate the adapted label distributions, the finally adapted standard deviations σ_α at the ages $\alpha \in \mathcal{Y}$ are shown in Fig. 6 as a gray-scale image. Each block in Fig. 6 represents

Fig. 6: The Adapted Standard Deviations σ_α Corresponding to Different Ages

one age, where higher gray scale (lighter) means larger σ_α, and further indicates slower facial appearance change. For easy comparison, a contrast stretching process is performed on the image to increase the contrast of different blocks, and some key ages are given under the corresponding gray-scale blocks. Note that the ages older than 60 are omitted because the training examples on these ages are too few to get a reasonable result. Roughly speaking, the σ_α before 45 is larger than that after 45, which indicates that the facial appearance changes slower in the age range 16-45 than in the age range 45-60. This is consistent with the fact mentioned in Section I that the facial appearance during the early adulthood changes slower than that during the senior age.

IV. Conclusion

This paper proposes a novel approach to facial age estimation based on adaptive label distribution learning (ALDL). There are two main targets in ALDL. The first is to create a proper label distribution for each training face image, which is adapted to its chronological age. The second is to learn a suitable conditional p.m.f., which can generate a label distribution similar to the adapted label distribution for each face image. In this way, ALDL can not only alleviate the problem of insufficient and incomplete training data in facial age estimation, but also accord with the general facial aging pattern of human beings. Experimental results show that the proposed ALDL algorithms (BFGS-ALDL and IIS-ALDL) perform significantly better than the compared state-of-the-art algorithms for facial age estimation.

Acknowledgment

This research was supported by the National Science Foundation of China (61273300, 61232007) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] A. M. Albert, K. Ricanek Jr, and E. Patterson, "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications," *Forensic Science International*, vol. 172, no. 1, pp. 1–9, 2007.

[2] A. L. Berger, S. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[3] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 585–592.

[4] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 2467–2474.

[5] 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, 2008.

[6] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, Beijing, China, 2007, pp. 1383–1386.

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

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

[9] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proc. the 14th ACM Int'l Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 307–316.

[10] X. Geng, K. Smith-Miles, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," in *Proc. 24th AAAI Conf. Artificial Intelligence*, Atlanta, GA, 2010, pp. 451–456.

[11] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 10, pp. 2401–2412, 2013.

[12] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[13] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011, pp. 657–664.

[14] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 112–119.

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

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

[17] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proc. the 6th Conf. Computational Natural Language Learning*, Taipei, Taiwan, 2002, pp. 49–55.

[18] B. Ni, Z. Song, and S. Yan, "Web image mining towards universal age estimator," in *Proc. the 17th ACM Int'l Conf. Multimedia*, New York, NY, 2009, pp. 85–94.

[19] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY: Springer, 2006.

[20] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 380–393, 1997.

[21] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, New York, NY, 2006, pp. 387–394.

[22] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Proc. the 7th Int'l Conf. Automatic Face and Gesture Recognition*, Southampton, UK, 2006, pp. 341–345.

[23] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimizationthat combines line search and trust region steps," *Mathematical Programming*, vol. 107, pp. 391–408, 2006.

[24] S. Yan, H. Wang, X. Tang, and T. S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. IEEE Int'l Conf. Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[25] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, 2007.