# Semi-supervised Adaptive Label Distribution Learning for Facial Age Estimation

**Peng Hou, Xin Geng*, Zeng-Wei Huo, Jia-Qi Lv**
MOE Key Laboratory of Computer Network and Information Integration,
School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
{hpeng, xgeng, huozw, lvjiaqi}@seu.edu.cn

## Abstract

Lack of sufficient training data with exact ages is still a challenge for facial age estimation. To deal with such problem, a method called Label Distribution Learning (LDL) was proposed to utilize the neighboring ages while learning a particular age. Later, an adaptive version of LDL called ALDL was proposed to generate a proper label distribution for each age. However, the adaptation process requires more training data, which creates a dilemma between the performance of ALDL and the training data. In this paper, we propose an algorithm called Semi-supervised Adaptive Label Distribution Learning (SALDL) to solve the dilemma and improve the performance using unlabeled data for facial age estimation. On the one hand, the utilization of unlabeled data helps to improve the adaptation process. On the other hand, the adapted label distributions conversely reinforce the semi-supervised process. As a result, they can promote each other to get better performance. Experimental results show that SALDL performs remarkably better than state-of-the-art algorithms when there are only limited accurately labeled data available.

## Introduction

Age estimation has attracted more attention in recent years because of its potential applications in business intelligence, human-computer interaction and so on. However, obtaining enough data with exact ages is still difficult. Although with the rise of big data, it is easier to collect a large scale of images for age estimation task, for example, from the Internet, unfortunately, these images are mostly labeled very roughly and contain large outliers of ages. They are harmful to learn a good age estimator (Ni, Song, and Yan 2009). One solution to utilize these images is to re-annotate them by humans. Obviously, it is not only time-consuming and cost-consuming but unreliable. The reason is that different people may age quite differently, because of the gene differences as well as life style and living condition differences. Correspondingly, there was one research reported that annotating the face images with perceiving ages by humans is a challenging task (Zeng et al. 2011). The mean absolute error (MAE) is 8.58 years for human age perception on images from MORPH (Ricanek Jr and Tesafaye 2006), and is 8.13 years on images from FG-NET (Cootes and Lanitis 2008). The other possible solution is to design a robust age estimator to tolerate the ill-labeled images. However, it has been also proved to be impractical. In detail, applying the age estimator learned from the collected face aging data from the Internet to some standard datasets, the MAE could be 8.60 years on MORPH (Ricanek Jr and Tesafaye 2006) and 9.49 years on FG-NET (Cootes and Lanitis 2008), respectively.

To estimate ages as precisely as possible, it is more feasible to make the most of existing slight well-labeled aging data. Motivated by this, one novel method was proposed to use the neighboring ages while learning a particular age based on the fact that aging is a slow and gradual process (Geng, Yin, and Zhou 2013). This is achieved by assigning a label distribution instead of 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. The shape of label distributions is same at all ages. Later, they found that the aging process could be significantly different at different aging stages (Geng, Wang, and Xia 2014). Generally speaking, the change of facial appearance during the childhood and senior ages is more apparent than that during the middle age. To accord with the tendency of aging variation, an adaptive method called Adaptive Label Distribution Learning (ALDL), was proposed to generate the proper label distribution for each age. But the adaptation process itself requires more labeled training data. Specially, if the training data are extremely limited, the performance of ALDL will get worse with the increase of the number of the adaptation iterations. An example is shown in Fig. 2. The horizontal ordinate represents the label distribution adaptation step. The vertical ordinate represents the MAE on 5,000 test images. The number of the training images for ALDL is 500. As can be seen that the MAE increases with the increase of the adaptation step. This demonstrates the poor performance of ALDL with limited training images.

Semi-supervised learning is an efficient technique to make use of unlabeled data to improve performance (Zhu 2005). As discussed above, the images with faces are extremely easy to obtain today. Tens of millions of face images are produced by people all over the world every day. If the images are treated as unlabeled, semi-supervised learning can be used. Correspondingly, there was a semi-supervised

*Corresponding author.

approach proposed for age estimation task (Kazuya, Sugiyama, and Ihara 2010). However, it is about perceived ages by humans rather than real ages estimated by a computer. In other words, in this research, there is no ground truth age labels for face images. Another research attempted to apply the general semi-supervised algorithms to solve the age estimation problem (Zhang and Guo 2013). Nevertheless, the algorithms are not very appropriate because they are not designed specially for age estimation.

To solve the dilemma between the performance of ALDL and the available labeled training data, we combine the label distribution adaptation and semi-supervised learning together to propose a novel method called *Semi-supervised Adaptive Label Distribution Learning* (SALDL). Generally speaking, there are three advantages of SALDL:

1. It utilizes the face images at neighboring ages when modeling a particular age via the label distribution and accords with the tendency of facial aging at different ages via the label distribution adaptation.
2. The semi-supervised process solves the dilemma between the performance of label distribution adaptation and the labeled training data by utilizing the unlabeled data. With more training data, the label distributions can be better adapted to the reality.
3. The better adapted label distributions can conversely enhance the utilization of unlabeled data. The reason is that, for the unlabeled data, the label distribution reflects not only the aging process but also the uncertainty of label assignment, which are more flexible than many traditional semi-supervised methods, where the label for each unlabeled image must be explicitly determined.

The rest of this paper is organized as follows. First, the algorithms about supervised age estimation and general semi-supervised learning are reviewed. Then, the SALDL algorithm is proposed. After that, the experimental results are reported. Finally, conclusions are drawn.

## Related Work

There have been many algorithms proposed for age estimation. The early works were the Weighted Appearance Specific (WAS) method and the Appearance and Age Specific (AAS) method in which the aging pattern was represented by a quadratic function (Lanitis, Draganova, and Christodoulou 2004). Then, the algorithm called AGES was proposed, which learned a subspace from the aging pattern vectors (Geng et al. 2006; Geng, Zhou, and Smith-Miles 2007). In (Fu and Huang 2008), multiple linear regression was used to the discriminative aging manifold of face images. After that, a locally adjusted robust regressor was designed for the prediction of human ages (Guo et al. 2008). Later, the feature extractor BIF (Biologically Inspired Features) and the KPLS (Kernel Partial Least Squares) regression method were used for age estimation (Guo et al. 2009). Besides, some work regarded age estimation as a regression problem with nonnegative label intervals and solved the problem through semidefinite programming (Yan et al. 2007a). Correspondingly, a method tried to model the age estimation as a multi-instance regression problem (Ni, Song, and Yan 2009). Ordinal Hyperplane Ranking (OHRank) algorithm was based on the transformation from the age estimation task into multiple cost-sensitive binary classification subproblems (Chang, Chen, and Hung 2011). A regressor for age estimation using a cumulative attribute was proposed (Chen et al. 2013) too. Especially, deep learning (Schmidhuber 2015) recently gets more and more attention and achieves great success in computer vision community. The corresponding deep algorithms for age estimation were also developed (Dong, Liu, and Lian 2015). However, almost all these algorithms require a large number of data with exact ages for training.

Semi-supervised learning is a relatively well-explored area (Chapelle et al. 2006). Self-training is the learning process which uses its own predictions to teach itself (Rosenberg, Hebert, and Schneiderman 2005). Co-training assumes the existence of two separate views in the feature space (Nigam and Ghani 2000). Then the two classifiers are applied to different views separately and they can enhance each other. S3VM is the extension of the popular support vector machine to semi-supervised learning condition (Joachims 1999). There are also some graph-based methods like Label Propagation (Wang and Zhang 2008) and mixture models like GMM (Gaussian Mixture Models) (Grandvalet and Bengio 2004). Although there are so many general semi-supervised algorithms proposed, it is not much appropriate to apply them to facial age estimation directly, because they cannot utilize the facial appearance characteristics.

## Semi-supervised Adaptive Label Distribution Learning

### Problem Foumulation

The name of SALDL, i.e., Semi-supervised Adaptive Label Distribution Learning, includes three key concepts: semi-supervised, adaptive and label distribution. Next we will introduce the concepts in the reverse order.

First, for a face image $\boldsymbol{x}$, its label distribution is defined as a vector $d_{\boldsymbol{x}}$, which contains the description degrees of a certain number of neighboring ages. The label distribution has two properties. One is that each element of the distribution vector $\boldsymbol{d}_{\boldsymbol{x},y}$ is a nonnegative real number representing the degree to which the age $y$ describes $\boldsymbol{x}$. The other is that all elements involved in the distribution vector sum up to 1, i.e., $\sum_y d_{\boldsymbol{x},y} = 1$, which means that the face image can always be fully described using all ages.

The form of the label distribution is a crucial issue, which should reflect the aging process correctly. According to the fact that aging is a slow and gradual process (Geng, Yin, and Zhou 2013), there are two constraints that the label distribution should satisfy. First, for the face image at the chronological age $\mu$, the description degree of $\mu$ should be the highest in the label distribution. Second, the description degrees of the neighboring ages should decrease with the increase of distance away from $\mu$. The discretized Gaussian distribution centered at the age $\mu$ might be a suitable choice, i.e.,

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

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

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

The other fact is then observed that the facial aging process could be significantly different at different aging stages (Geng, Wang, and Xia 2014). Fortunately, this difference can be reflected by the standard deviation $\sigma$ of the Gaussian label distribution at each age. For those ages where the facial appearance changes faster, the $\sigma$ should be smaller, resulting in sharper label distributions. On the contrary, for those ages where the facial appearance changes slower, the $\sigma$ should be larger, resulting in smoother label distributions. Thus, the standard deviations at different ages may be different. However, the label distributions are not available in the training data. This requires the algorithm to be able to learn the label distributions adapted to different ages. In this sense, label distribution adaptation means finding a proper $\sigma$ for each age.

Unfortunately, to find a good $\sigma$ for each age, a certain number of labeled training data are needed. This is against the general condition of limited labeled training data as discussed in the first section. Noticing the fact that the unlabeled data are abundant and quite easy to obtain, we turn to semi-supervised learning to utilize both labeled and unlabeled data. In one word, SALDL is a semi-supervised method aiming to enhance the label distribution adaptation using the unlabeled data for facial age estimation.

According to the above description, label distribution shares the same properties with probability distribution, i.e., $d_{\boldsymbol{x},y} \in [0,1]$ and $\sum_y d_{\boldsymbol{x},y} = 1$. Thus, many theories and methods can be borrowed from statistics to deal with label distributions. First of all, the description degree $d_{\boldsymbol{x},y}$ can be represented by the form of conditional probability, i.e., $d_{\boldsymbol{x},y} = p(y|\boldsymbol{x})$. This might be explained as that the probability of $y$ is equal to its description degree. Then we suppose $p(y|\boldsymbol{x})$ is a parametric model $p(y|\boldsymbol{x};\boldsymbol{\Theta})$, where $\boldsymbol{\Theta} \in \mathbb{R}^{r\times q}$ is the parameter matrix, $r$ and $q$ are the number of feature vector and age set respectively. Therefore, there are two optimization targets for SALDL, i.e., the parameter matrix $\boldsymbol{\Theta}$ in the conditional probability function and the standard deviation $\sigma$ for each age in Eq. (1). These targets can be optimized alternatively in a loop of four main steps as the following.

## SALDL Algorithm

**Label Distribution Initialization** The first step is to initilize the label distributions. Because the ages of unlabeled images are unknown, the initialization is only for the labeled images. The original label distributions are initialized by Eq. (1) with the same standard deviation $\sigma^0$ at all ages, i.e., $\sigma^0_\mu = \sigma^0, \forall \mu \in \mathcal{Y}$, where $\sigma^0$ is predefined and $\mathcal{Y}$ is the age set. For example, for the labeled image $\boldsymbol{x}_i$, the label distribution $\boldsymbol{d}^0_i = [d^0_{\boldsymbol{x}_i,y_1}, d^0_{\boldsymbol{x}_i,y_2}, \ldots, d^0_{\boldsymbol{x}_i,y_q}]$ is calculated by

$$d^0_{\boldsymbol{x}_i,y_j} = \frac{1}{\sigma^0_{\mu_i}\sqrt{2\pi}Z_i} \exp(-\frac{(y_j-\mu_i)^2}{2(\sigma^0_{\mu_i})^2}), \quad j = 1,2,\ldots,q. \tag{3}$$

**Algorithm 1** SALDL

**Input:**
The initial standard deviation $\sigma^0$;
The number of nearest neighbors $K$;
The balance parameter $C$;
The maximum number of iterations $T$;
The labeled image set $\mathcal{S}_l = \{(\boldsymbol{x}_1,\mu_1),\ldots,(\boldsymbol{x}_l,\mu_l)\}$;
The unlabeled image set $\mathcal{S}_u = \{\boldsymbol{x}_{l+1},\ldots,\boldsymbol{x}_{l+u}\}$.
**Output:** $p(y|\boldsymbol{x};\boldsymbol{\Theta})$.

$k \leftarrow 0$;
$\mathcal{S}^k \leftarrow \mathcal{S}_l$;
Initialize the label distributions in $\mathcal{S}^k$ by the real ages and $\sigma^0$ according to Eq. (3);
**repeat**
  $k \leftarrow k+1$;
  Train a LDL model $\mathrm{M}^k$ based on $\mathcal{S}^{k-1}$ by solving Eq. (4) for the model parameter matrix $\boldsymbol{\Theta}^k$;
  Based on the $\mathrm{M}^k$, predict the label distributions in $(\mathcal{S}_l \cup \mathcal{S}_u)$ according to Eq. (5);
  Based on the predicted label distributions, estimate the pseudo ages $\mu^k$ in $\mathcal{S}_u$ according to Eq. (9);
  Based on the real ages in $\mathcal{S}_l$, the pseudo ages $\mu^k$ in $\mathcal{S}_u$ and the predicted label distributions, select the confident images;
  Devide the selected confident images according to their real ages or pseudo ages, and estimate the standard derivation $\sigma^k$ for each age group according to Eq. (14);
  Update $\mathcal{S}_l$ and $\mathcal{S}_u$ by the real ages, pseudo ages and $\sigma^k$ according to Eq. (1);
  $\mathcal{S}^k \leftarrow (\mathcal{S}_l \cup \mathcal{S}_u)$;
**until** $k \geq T$.
$p(y|\boldsymbol{x};\boldsymbol{\Theta}) = \frac{1}{\Lambda}\exp((\boldsymbol{\theta}_y^T)^{\mathbf{T}}\boldsymbol{x})$

If we have $l$ labeled images, then the initial label distribution training set $\mathcal{S}^0$ is covered as $\mathcal{S}^0 = \{(\boldsymbol{x}_1,\boldsymbol{d}^0_{\boldsymbol{x}_1}),\ldots,(\boldsymbol{x}_l,\boldsymbol{d}^0_{\boldsymbol{x}_l})\}$.

**Label Distribution Learning** In this step, the label distribution for each training image in the set $\mathcal{S}^{k-1}$ is available, where $k$ is the iteration number. Thus, the aim of this step is to find a $\boldsymbol{\Theta}^k$ that can generate label distributions most similar to the ones in $\mathcal{S}^{k-1}$. This process is called LDL, i.e., Label Distribution Learning. If the Kullback-Leibler distance is used to measure the similarity between distributions, then the best parameter matrix is determined by

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

According to the maximum entropy criterion (Berger, Pietra, and Pietra 1996), $p(y|\boldsymbol{x};\boldsymbol{\Theta})$ can be expressed as

$$p(y_j|\boldsymbol{x};\boldsymbol{\Theta}) = \frac{1}{\Lambda}\exp(\boldsymbol{\theta}^{\mathbf{T}}_{y_j}\boldsymbol{x}) \tag{5}$$

where $\Lambda = \sum_j \exp(\boldsymbol{\theta}^{\mathbf{T}}_{y_j}\boldsymbol{x})$ is the normalization factor, and $\boldsymbol{\theta}_{y_j}$ is the $j$-th column vector of the matrix $\boldsymbol{\Theta}$ corresponding

to the age $y_j$. Substituting Eq. (5) into Eq. (4) yields

$$\boldsymbol{\Theta}^k = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{i,j} d_{\boldsymbol{x}_i,y_j}^{k-1} \boldsymbol{\theta}_{y_j}^{\mathbf{T}} \boldsymbol{x}_i - \sum_i \ln \sum_j \exp(\boldsymbol{\theta}_{y_j}^{\mathbf{T}} \boldsymbol{x}_i). \tag{6}$$

There are many optimization algorithms such as conjugate gradient and quasi-Newton methods to optimize Eq. (6). If the quasi-Newton method BFGS is used, then the optimization is mainly related to the first order gradient, which can be obtained through

$$\frac{\partial L(\boldsymbol{\Theta})}{\partial \boldsymbol{\theta}_{y_j}} = \sum_i \frac{\exp(\boldsymbol{\theta}_{y_j}^{\mathbf{T}} \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_j \exp(\boldsymbol{\theta}_{y_j}^{\mathbf{T}} \boldsymbol{x}_i)\boldsymbol{x}_i} - \sum_i d_{\boldsymbol{x}_i,y_j}^{k-1} \boldsymbol{x}_i. \tag{7}$$

**Estimate the Pseudo Ages for Unlabeled Images** The parameter matrix $\boldsymbol{\Theta}^k$ in the conditional probability mass function is available after the LDL step. Then, it can be used to predict the label distribution for each training image $\boldsymbol{x}$ according to Eq. (5). To keep consistent between the labeled and unlabeled data, we need to estimate a temporary pseudo age for each unlabeled image. Here, we adopt the simple and efficient algorithm *KNN*, i.e., K-Nearest Neighbor (Peterson 2009). In detail, for each unlabeled image, the pseudo age is estimated as the mean age of its $K$ nearest labeled image neighbors. Note that, because only the ages of the labeled images are known, the neighbors should be searched only in the labeled training set. Furthermore, to find the neighbors more appropriately, the similarity metric is modified as one balance between the Euclidean distance of the feature vectors and the Kullback-Leibler divergence of the predicted label distributions. For example, the similarity between the unlabeled image $\boldsymbol{x}_m$ and the labeled image $\boldsymbol{x}_n$ is

$$\lambda_{m,n} = \parallel \boldsymbol{x}_m - \boldsymbol{x}_n \parallel_2^2 + C \sum_j p(y_j|\boldsymbol{x}_m;\boldsymbol{\Theta}^k) \ln \frac{p(y_j|\boldsymbol{x}_m;\boldsymbol{\Theta}^k)}{p(y_j|\boldsymbol{x}_n;\boldsymbol{\Theta}^k)}, \tag{8}$$

where $C$ is the predefined balance parameter. Then, the pseudo age for $\boldsymbol{x}_m$ in the iteration $k$ is estimated as

$$\mu_m^k = \frac{1}{K} \sum_{\boldsymbol{x}_n \in \mathcal{N}_m} \mu_n. \tag{9}$$

where $\mathcal{N}_m$ is the labeled neighbor set for $\boldsymbol{x}_m$.

There are two interesting things needed bo be concerned. First, the uncertainty of label assignment for the unlabeled data can be exactly depicted by the label distribution. Second, the predicted label distribution for each image may change in each iteration, thus, the pseudo age $\mu_m^k$ is alterable. These makes SALDL more robust than other general semi-supervised learning methods.

**Update the Adapted Label Distributions** Similar to ALDL, we assume that the aging process is different at different aging stages. This difference can be reflected by the standard deviations of the Gaussian label distributions. Unfortunately, the standard deviations can not be obtained directly from the traning data set. Thus, in this step, we expect to approximate the real standard deviation for each age.

According to the predicted label distribution, the age of image $\boldsymbol{x}$ can be estimated as the one which has the largest description degree

$$\hat{\mu} = \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x};\boldsymbol{\Theta}^k). \tag{10}$$

However, this age may be different from the real age for the labeled image or the pseudo age for the unlabeled image. To approximate the standard deviations as precisely as possible, we should firstly select the confident images whose predicted ages are accurate. In detail, we calculate the absolute error of the predicted age for each image as

$$e = |\mu - \hat{\mu}|. \tag{11}$$

Note that for the labeled images, the $\mu$ corresponds to the real chronological age. For the unlabeled images, it is substituted by the pseudo ages estimated in the previous step. Then, those images with the absolute age estimation errors lower than the MAE (Mean Absolute Error) of the whole image set are selected as the candidate set for the update of standard deviations. The MAE is calculated as

$$\text{MAE} = \frac{1}{l+u} \sum_i e_i. \tag{12}$$

where $l$ and $u$ are the number of labeled and unlabeled images, respectively.

After that, we assume all people have the same aging process. That means the label distributions are same for those images with the same age. Thus, the selected confident images can be divided into $q$ subsets by their real ages or pseudo ages. Let $\mathcal{S}_\mu^k$ denote the selected image set with the same age $\mu$ in the $k$-th iteration, then the corresponding label distributions of the images in $\mathcal{S}_\mu^k$ are all generated with the same standard deviation $\sigma_\mu^k$, i.e.,

$$d_{\boldsymbol{x}_r,y_j} = \frac{1}{\sigma_\mu^k \sqrt{2\pi} Z_\mu} \exp(-\frac{(y_j-\mu)^2}{2(\sigma_\mu^k)^2}), \tag{13}$$

However, even the images with the same age are in the same subset, their predicted label distributions calculated according to Eq. (5) may be different. To deal with this situation, we define the best $\sigma_\mu^k$ for age $\mu$ as the one that can generate a label distribution most similar to the all predicted label distributions in $\mathcal{S}_\mu^k$. If the Kullback-Leibler divergence is again used to measure the similarity between two distributions, then

$$\begin{aligned} \sigma_\mu^k = &\underset{\sigma_\mu}{\operatorname{argmin}} \sum_{\boldsymbol{x}_r \in \mathcal{S}_\mu^k} \sum_j d_{\boldsymbol{x}_r,y_j} \ln \frac{d_{\boldsymbol{x}_r,y_j}}{p(y_j|\boldsymbol{x}_r;\boldsymbol{\Theta}^k)}, \\ &s.t. \quad \sigma_\mu > 0. \end{aligned} \tag{14}$$

Substituting Eq. (13) into Eq. (14) yields a nonlinear programming problem, which can be effectively solved by the log barrier interior-point method (Waltz et al. 2006).

After $\sigma_\mu^k$ is determined for each age in $\mathcal{Y}$, the label distributions of all training images are updated with the new standard deviation $\sigma_\mu^k$ to get $\boldsymbol{d}_{\boldsymbol{x}}^k$ according to Eq. (1). Note that for the labeled images, the $\mu$ in Eq. (1) corresponds to the real chronological age. For the unlabeled images, the $\mu$ is again substituted by the estimated pseudo age. Finally, the training set with the updated label distributions $\mathcal{S}^k = \{(\boldsymbol{x}_1, \boldsymbol{d}_{\boldsymbol{x}_1}^k), \ldots, (\boldsymbol{x}_l, \boldsymbol{d}_{\boldsymbol{x}_l}^k), (\boldsymbol{x}_{l+1}, \boldsymbol{d}_{\boldsymbol{x}_{l+1}}^k), \ldots, (\boldsymbol{x}_{l+u}, \boldsymbol{d}_{\boldsymbol{x}_{l+u}}^k)\}$ is sent into the LDL step again to start the next iteration $k+1$. The whole process repeats until $k$ is not less than the predefined maximum number of iterations $T$. Since both the LDL step and the adaptation step iteratively reduce the K-L divergence, the iteration procedure will converge at last. The pseudo-code is shown as Algorithm. 1.

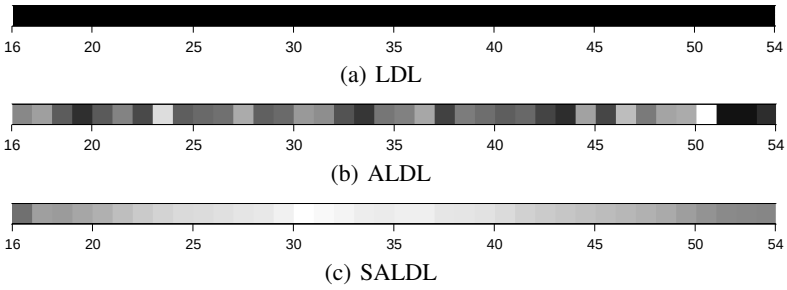


Figure 1: The standard deviations $\sigma_\mu$ at different ages estimated by LDL, ALDL and SALDL.

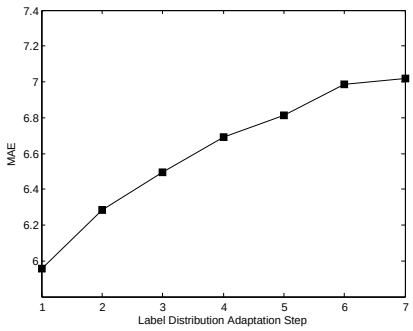


Figure 2: MAE variation of ALDL w.r.t. the label distribution adaptation steps using 500 labeled training images.

# Experiments

## Configuration

The dataset used in the experiments is MORPH (Ricanek Jr and Tesafaye 2006). There are 55,132 face images from more than 13,000 subjects in this database. In average, each subject has 4 face images. The ages range from 16 to 77 with a median age of 33. The faces are from different races, where about 77% are African faces, 19% are European faces, and remaining 4% includes Hispanic, Asian, Indian and other races. The feature extractor is the Biologically Inspired Features (BIF) (Guo et al. 2009). The dimensionality of the BIF vectors is further reduced to 200 using Marginal Fisher Analysis (MFA) (Yan et al. 2007b).

For SALDL, the initial standard deviation $\sigma^0$ is set to 3, the number of nearest neighbors $K$ is set to 10, the balance weight $C$ is set to 0.001. The maximum number of iterations $T$ decreases with the increase of the number of the labeled training images. All parameters are determined through the 10-fold cross validation process. After the optimal model parameter matrix $\mathbf{\Theta}^*$ is obtained, the predicted age for a test image $\boldsymbol{x}'$ is determined by $y^* = \operatorname{argmax}_y p(y|\boldsymbol{x}'; \mathbf{\Theta}^*)$.

Several existing facial age estimation algorithms are compared, which include KPLS (Guo and Mu 2011), OHRank (Chang, Chen, and Hung 2011), LDL (Geng, Yin, and Zhou 2013), and ALDL (Geng, Wang, and Xia 2014). As suggested in the papers, KPLS uses the RBF kernel with the inverse width of 1. OHRank uses the absolute cost function and the RBF kernel. For ALDL, the parameters $\sigma^0$ is set to 3 and the convergence threshold $\varepsilon$ is set to 0.02.

The age estimation can be treated as a multi-class problem in fact. There have been some semi-supervised multi-class algorithms proposed in the recent years, like Boosting (Valizadegan, Jin, and Jain 2008) and Support Vector Machine (Xu and Schuurmans 2005). However, LP (Label Propagation) is more scalable and commonly used in multi-class condition. Thus, we adopt the one proposed in (Wang and Zhang 2008) as a representer. For LP, the number of nearest neighbors is set to 10. In addition, to show the effect of the semi-supervised process, we modify SALDL by removing the adaptation process and call it as SLDL method. In SLDL, the initialization step, the LDL step and the pseudo age estimation are same as SALDL. Then, SLDL assigns the label distributions to the unlabeled images with the pseudo ages and $\sigma^0$ according to Eq. (1). After that, each image, either labeled or unlabeled, has a corresponding label distribution, but the standard deviations of the distributions are all equal to $\sigma^0$. Last, LDL step is executed again on the whole image set and the parameter matrix $\mathbf{\Theta}$ is got. The parameters in SLDL are set to be same as the ones in SALDL.

The test images are randomly selected firstly and fixed in all experiments. The number of the test images is 5,000. For the semi-supervised methods, i.e., LP, SLDL and SALDL, the number of training images they use in all experiments is always 50,000. In other words, if the number of labeled training images increase, then the number of unlabeled images will decrease, but the sum of them is still 50,000.

## Results

**Motivation Justification** To reveal the performance of ALDL with limited training images, the MAE variation with respect to the label distribution adaptation steps is shown in Fig. 2. Note that the number of labeled training images is 500. As shown in Fig. 2, the MAE increases with the increase of label distribution adaptation step. Thus, the adaptation process generates negative effect if the training images are rare for ALDL. On this occasion, ALDL degenerates to LDL with the best performance without adaptation.

Further, Fig. 1(b) shows the estimated adapted standard deviations $\sigma_\mu$ by ALDL with 500 labeled training images. Each block in Fig. 1 represents one age, where higher gray scale (lighter) means larger $\sigma_\mu$, and further indicates slower facial appearance change. Note that the ages older than 54 are omitted because the training examples on these ages are too few to get a reasonable result. Generally speaking, the

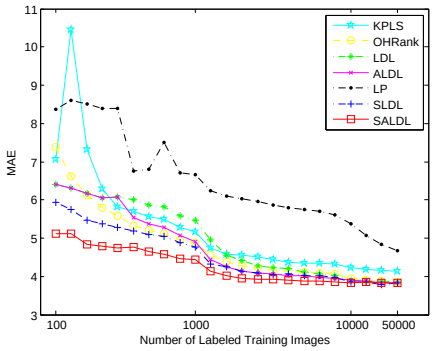


Figure 3: Comparison among all methods

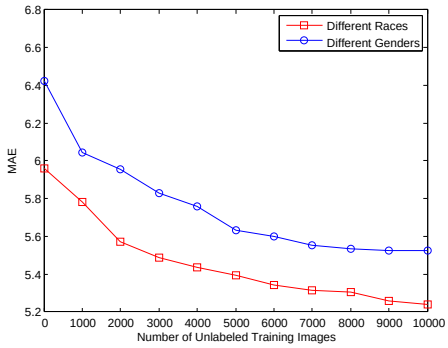


Figure 4: Using unlabeled images from different sources.

facial appearance changes slower during middle ages than that during childhood and senior ages. Thus, the gray should be lighter at age 30 than that at age 16 and 54. However, limited by the training images, the estimated standard deviations by ALDL are chaotic and can not correctly accord with the tendency of aging variation as shown in Fig. 1(b). As a contrast, the standard deviations $\sigma_\mu$ in LDL are same at all ages, thus the gray bar is completely black as shown in Fig. 1(a). For SALDL, as shown in Fig. 1(c), the estimated standard deviations perfectly accord with the aging process.

**Effects of the Semi-supervised and Adaptation Processes** To demonstrate the effects of the semi-supervised process more sufficiently, some experiment results of ALDL, SLDL and SALDL are shown in Fig. 3. The horizontal ordinate represents the number of labeled training images, and the vertical ordinate represents the MAE. Note that the results can be divided into three parts according to the value of horizontal axis. The first part ranges from 100 to 1,000, and the step interval is 100. The second part ranges from 1,000 to 10,000, and the step interval is 1,000. The last part ranges from 10,000 to 50,000, and the step interval is 10,000. As can be seen from the figure, SLDL and SALDL perform better than ALDL, especially when the number of labeled training images is less than 1,000. This strongly demonstrates the effect of semi-supervised process under the condition of limited training data.

Furthermore, SALDL performs better than SLDL, and this shows the positive effect of the label distribution adaptation. Besides, because of the diversity of intervals in different parts, the change rates of MAEs are also different. As shown in Fig. 3, the MAE changes faster in the first part than that in the third part. In particular, the MAEs in the third part, i.e., when the number of labeled training images is larger than 10,000, have little variation with the increase of labeled training images. This is probably because the training data are already enough, and adding either the labeled data or the unlabeled data will not have much effects.

**Comparison between SALDL and State-of-the-art Methods** The comparison results between SALDL and other state-art-of methods are also shown in Fig. 3. First, the green line for LDL and the magenta line for ALDL in Fig. 3 are overlapped when the number of labeled training images is less than 500. It is consistent with the former conclusion that when the training images are rare, ALDL degenerates to LDL. Then the semi-supervised method LP performs worst. The reason might be that other compared methods are all designed for the facial age estimation task, so the general semi-supervised method has no advantages even there are enough unlabeled data used. Moreover, there are several fluctuations for the methods LP and KPLS, this illustrates that they may be unstable when the training data are rare. Last, SALDL performs remarkably better than the compared methods especially when the labeled training images are extremely rare. For example, the MAE of SALDL with 100 labeled training images is 5.124. To get similar performance, the numbers of labeled training images required are 1,000 for KPLS, 800 for OHRank, 2,000 for LDL, 900 for ALDL, and 30,000 for LP, respectively. This essentially illustrates the advantages and effectiveness of SALDL. In particular, if the training images are all labeled, i.e., the number of labeled training images is 50,000, SALDL will degenerate to ALDL.

**Utilization of Unlabeled Data from Different Sources** In the above experiments, the sources of labeled and unlabeled images used in SALDL are same. However, this may be not satiable in real applications. In more general conditions, they are likely different. Thus, to illustrate the performance of SALDL fully, we design two more experiments with different races and genders, respectively. In the first one, we use 500 Caucasian face images as the labeled training data, the rest Caucasian face images, about 10,167 images, as the test data, and randomly select images from 1,000 to 10,000 images from other races as the unlabeled training data. The second one is almost same except that the labeled and test data are female face images and the unlabeled data are male face images. The results are shown in Fig. 4. We can see that the MAE decreases with the increase of the number of the unlabeled data in both experiments. It can well demonstrate the effectiveness of semi-supervised process in SALDL as the above part.

## Conclusion

This paper proposes a novel semi-supervised method called SALDL for facial age estimation. The motivation of SALDL is to solve the dilemma between the performance of label distribution adaptation and the limited training data. This

is achieved by combining the semi-supervised process and the adaptation process uniformly via the label distribution. In SALDL, the two procedures can promote each other and get better performance. Experimental results show that the proposed SALDL algorithm can make good use of the unlabeled data and perform significantly better than state-of-the-art algorithms when the available labeled images are limited.

## Acknowledgements

This research was supported by National Science Foundation of China (61622203, 61273300, 61232007), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), Fundamental Research Funds for the Central Universities (SJLX15_0043), Research Innovation Program for College Graduates of Jiangsu Province, Collaborative Innovation Center of Novel Software Technology and Industrialization, and Collaborative Innovation Center of Wireless Communications Technology.

## References

Berger, A. L.; Pietra, V. J. D.; and Pietra, S. A. D. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22(1):39–71.

Chang, K.-Y.; Chen, C.-S.; and Hung, Y.-P. 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 585–592.

Chapelle, O.; Schölkopf, B.; Zien, A.; et al. 2006. Semi-supervised learning.

Chen, K.; Gong, S.; Xiang, T.; and Loy, C. 2013. Cumulative attribute space for age and crowd density estimation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2467–2474.

Cootes, T., and Lanitis, A. 2008. The fg-net aging database.

Dong, Y.; Liu, Y.; and Lian, S. 2015. Automatic age estimation based on deep learning algorithm. *Neurocomputing*.

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

Geng, X.; Zhou, Z.-H.; Zhang, Y.; Li, G.; and Dai, H. 2006. Learning from facial aging patterns for automatic age estimation. In *Proc. 14th ACM Int'l Conf. Multimedia*, 307–316.

Geng, X.; Wang, Q.; and Xia, Y. 2014. Facial age estimation by adaptive label distribution learning. In *Proc. 22nd IEEE International Conference on Pattern Recognition*, 4465–4470.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE Trans. Pattern Analysis and Machine Intelligence* 35(10):2401–2412.

Geng, X.; Zhou, Z.-H.; and Smith-Miles, K. 2007. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29(12):2234–2240.

Grandvalet, Y., and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, 529–536.

Guo, G., and Mu, G. 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 657–664.

Guo, G.; Fu, Y.; Dyer, C. R.; and Huang, T. S. 2008. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Processing* 17(7):1178–1188.

Guo, G.; Mu, G.; Fu, Y.; and Huang, T. S. 2009. Human age estimation using bio-inspired features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 112–119.

Huerta, I.; Fernández, C.; Segura, C.; Hernando, J.; and Prati, A. 2015. A deep analysis on age estimation. *Pattern Recognition Letters* 68:239–249.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, volume 99, 200–209.

Kazuya, U.; Sugiyama, M.; and Ihara, Y. 2010. A semi-supervised approach to perceived age prediction from face images. *IEICE Trans. Information and Systems* 93(10):2875–2878.

Lanitis, A.; Draganova, C.; and Christodoulou, C. 2004. Comparing different classifiers for automatic age estimation. *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics* 34(1):621–628.

Ni, B.; Song, Z.; and Yan, S. 2009. Web image mining towards universal age estimator. In *Proc. 17th ACM Int'l Conf. Multimedia*, 85–94. ACM.

Nigam, K., and Ghani, R. 2000. Analyzing the effectiveness and applicability of co-training. In *Proc. 9th ACM Int'l Conf Information and knowledge management*, 86–93.

Peterson, L. E. 2009. K-nearest neighbor. *Scholarpedia* 4(2):1883.

Ricanek Jr, K., and Tesafaye, T. 2006. Morph: A longitudinal image database of normal adult age-progression. In *Proc. 7th IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 341–345.

Rosenberg, C.; Hebert, M.; and Schneiderman, H. 2005. Semi-supervised self-training of object detection models.

Schmidhuber, J. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61:85–117.

Valizadegan, H.; Jin, R.; and Jain, A. K. 2008. Semi-supervised boosting for multi-class classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 522–537. Springer.

Waltz, R. A.; Morales, J. L.; Nocedal, J.; and Orban, D. 2006. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical programming* 107(3):391–408.

Wang, F., and Zhang, C. 2008. Label propagation through linear neighborhoods. *IEEE Trans. Knowledge and Data Engineering* 20(1):55–67.

Xu, L., and Schuurmans, D. 2005. Unsupervised and semi-supervised multi-class support vector machines. In *Proc. IEEE Conf. Association for the Advancement of Artificial Intelligence*, volume 5.

Yan, S.; Wang, H.; Tang, X.; and Huang, T. S. 2007a. Learning auto-structured regressor from uncertain nonnegative labels. In *Proc. 11th IEEE International Conference on Computer Vision*, 1–8.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H.-J.; Yang, Q.; and Lin, S. 2007b. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29(1):40–51.

Zeng, J.; Ling, H.; Latecki, L. J.; Fitzhugh, S.; and Guo, G. 2011. Analysis of facial images across age progression by humans. *Machine Vision* 2012.

Zhang, C., and Guo, G. 2013. Exploiting unlabeled ages for aging pattern analysis on a large database. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 458–464.

Zhu, X. 2005. Semi-supervised learning literature survey.