IEICE Transactions on Information and Systems,
vol.E93-D, no.10, pp.2875–2878, 2010.

A Semi-supervised Approach to Perceived Age Prediction from Face Images

Kazuya Ueki
NEC Soft, Ltd., Japan

Masashi Sugiyama
Tokyo Institute of Technology, Japan

Yasuyuki Ihara
NEC Soft, Ltd., Japan

Abstract

We address the problem of perceived age estimation from face images, and propose a new semi-supervised approach involving two novel aspects. The first novelty is an efficient active learning strategy for reducing the cost of labeling face samples. Given a large number of unlabeled face samples, we reveal the cluster structure of the data and propose to label cluster-representative samples for covering as many clusters as possible. This simple sampling strategy allows us to boost the performance of a manifold-based semi-supervised learning method only with a relatively small number of labeled samples. The second contribution is to take the heterogeneous characteristics of human age perception into account. It is rare to misjudge the age of a 5-year-old child as 15 years old, but the age of a 35-year-old person is often misjudged as 45 years old. Thus, magnitude of the error is different depending on subjects' age. We carried out a large-scale questionnaire survey for quantifying human age perception characteristics, and propose to utilize the quantified characteristics in the framework of weighted regression. Consequently, our proposed method is expressed in the form of weighted least-squares with a manifold regularizer, which is scalable to massive datasets. Through real-world age estimation experiments, we demonstrate the usefulness of the proposed method.

Keywords

perceived age estimation, active learning, weighted regression, semi-supervised learning, manifold regularization, human age perception

1 Introduction

Demographic analysis in public places such as shopping malls and train stations is attracting a great deal of attention these days since it is useful for designing effective marketing strategies. Such demographic information is often collected manually, e.g., at convenience stores, sales clerks input customers' attributes such as age and gender to a point-of-sale (POS) system. However, such manual data collection requires a lot of human labor, and thus automating this process is highly desired.

In this paper, we address the problem of age estimation from face images using machine learning techniques. Most of the existing studies on age estimation try to predict subjects' *real* age [6, 5, 8, 3, 2]. However, the problem of estimating subjects' real age is highly ill-posed since the correspondence between appearance and real age is not clear even for humans.

When designing marketing strategies, analyzing *perceived* age is often preferred to real age. However, little attention has been paid to perceived age analysis so far. In this paper, we therefore propose a new method of perceived age estimation from face images. Perceived age of a subject is defined as the mean of estimated age by a large number of people. Thus the problem of perceived age estimation can be naturally formulated as a *regression* problem, which is aimed at estimating the conditional mean of outputs (estimated age) given inputs (face images).

Face images often contain complex variability due to diversity of individual characteristics, angles, lighting conditions, etc. Thus a large number of *labeled* face samples are usually needed to achieve good prediction performance. However, labeling face samples requires much time and effort, and thus it is desirable to reduce the number of labeled samples without performance degradation. In this paper, we first propose an *active learning* strategy for reducing the sampling cost. We focus on a *semi-supervised* setup where a large number of unlabeled face samples are available abundantly. Our active learning idea is to apply a clustering technique to reveal the cluster structure of the face data and to label cluster-representative samples for covering as many clusters as possible. This simple sampling strategy allows us to boost the performance of a *manifold-based* semi-supervised learning method [9] only with a relatively small number of labeled samples.

In order to further improve the estimation accuracy, we propose to take the heterogeneous characteristics of human age perception into account—the age of a 5-year-old child may not be misjudged as 15 years old, but the age of a 35-year-old person is often misjudged as 45 years old. Thus, deviation of the age estimation error is different depending on subjects' age (which is referred to as *heteroscedastic noise*). We carried out a large-scale questionnaire survey in order to quantify human age perception characteristics. Based on the survey results, we propose to take account of the quantified characteristics by *weighted* regression, which is shown to be able to cope with heteroscedastic noise.

Combining the above two ideas, we propose a kernel-based semi-supervised perceived age estimation method which is expressed in the form of kernel weighted least-squares with a manifold regularizer. Thanks to its simple formulation, the proposed method is scalable to large-scale datasets. Through real-world age estimation experiments, we demonstrate

the usefulness of the proposed method.

2 Semi-supervised Approach to Perceived Age Estimation

In this section, we describe the proposed procedure for perceived age estimation.

2.1 Clustering-based Active Learning Strategy

First, we explain our active learning strategy for reducing the cost of labeling face samples.

Face samples contain various diversity such as individual characteristics, angles, lighting conditions, etc. They often possess cluster structure, and face samples in each cluster tend to have similar ages [2, 4, 12]. Based on these empirical observations, we propose to label the face images which are closest to cluster centroids.

For revealing the cluster structure, we apply the k-means clustering method to a large number of unlabeled samples. Since clustering of high-dimensional data is often unreliable, we first apply principal component analysis (PCA) to the face images for dimension reduction. The proposed active learning strategy is summarized as follows.

1. For a set of n-dimensional unlabeled face image samples $\{\boldsymbol{X}_i\}_{i=1}^t$, we compute $\{\boldsymbol{x}_i\}_{i=1}^t$ of m ($\ll n$) dimensions by the PCA projection.
2. Using the k-means clustering algorithm, we compute the l ($\ll t$) cluster centroids $\{\boldsymbol{m}_j\}_{j=1}^l$.
3. We choose $\{\boldsymbol{x}_{\hat{i}_j}\}_{j=1}^l$ as samples to be labeled, where $\hat{i}_j = \operatorname*{argmin}_i \|\boldsymbol{x}_i - \boldsymbol{m}_j\|$ and $\|\cdot\|$ denotes the Euclidean norm.

For making the notation simple, we permute the order of samples $\{\boldsymbol{x}_i\}_{i=1}^t$ so that the first l samples $\{\boldsymbol{x}_i\}_{i=1}^l$ are labeled and the remaining u ($= t - l$) samples $\{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$ are unlabeled—this permutation is always possible without loss of generality. Let $\{y_i\}_{i=1}^l$ be the labels for $\{\boldsymbol{x}_i\}_{i=1}^l$.

2.2 Semi-supervised Age Regression with Manifold Regularization

As explained above, face images often possess cluster structure, and face samples in each cluster tend to have similar ages. Here we utilize this cluster structure by employing a method of semi-supervised regression with manifold regularization [9].

For age regression, we use the following kernel model:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l+u} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{l+u})^\top$ are parameters to be learned, $\top$ denotes the transpose, and $k(\boldsymbol{x}, \boldsymbol{x}')$ is a *reproducing kernel function*. We use the Gaussian kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/(2\sigma^2)\right),$$

where σ^2 is the Gaussian variance. We included $(l+u)$ kernels in the kernel regression model (1), but u can be very large in age prediction. In practice, we may only use c $(\leq u)$ elements randomly chosen from the set $\{k(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=l+1}^{l+u}$ for reducing the computational cost; then the total number of basis functions is reduced to $b = l + c$. However, we stick to Eq.(1) below for keeping the explanation simple.

We employ a manifold regularizer [9] in our training criterion, i.e., the parameter $\boldsymbol{\alpha}$ is learned so that the following criterion is minimized.

$$\frac{1}{l}\sum_{i=1}^{l}(y_i - f(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{\alpha}\|^2 + \frac{\mu}{4(l+u)^2}\sum_{i,j=1}^{l+u} W_{i,j}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2, \tag{2}$$

where λ and μ are non-negative constants. $W_{i,j}$ represents the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which is defined by

$$W_{i,j} = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(2\gamma^2)\right) \tag{3}$$

if $\boldsymbol{x}_i$ is a h-nearest neighbor of $\boldsymbol{x}_j$ or vice versa; otherwise $W_{i,j} = 0$.

The first term in Eq.(2) is the goodness-of-fit term and the second term is the ordinary regularizer for avoiding overfitting. The third term is the manifold regularizer. The weight $W_{i,j}$ tends to take large values if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster. Thus, the manifold regularizer works for keeping the outputs of the function $f(\boldsymbol{x})$ within the same cluster close to each other.

2.3 Incorporating Age Perception Characteristics

Next, we extend the above manifold regularization method so that human age perception characteristics can be taken into account.

First, we quantify human age perception characteristics through a large-scale questionnaire survey. We used an in-house face image database consisting of approximately 500 subjects whose age almost uniformly covers the range of our interest (i.e., age 1 to 70). For each subject, 5 to 10 face images with different face poses and lighting conditions were taken. We asked each of 72 volunteers to give age labels y to the subjects. The '*true*' age of a subject[1] is defined as the average of estimated age labels y (rounded-off to the nearest integer) for that subject, and denoted by y^*. Then the standard deviation of age labels y is calculated as a function of y^*, which is summarized in Figure 1.

The standard deviation is approximately 2 (years) when the true age y^* is less than 15. The standard deviation increases and goes beyond 6 as the true age y^* increases from 15 to 35. Then the standard deviation decreases to around 5 as the true age y^* increases from 35 to 70. This graph shows that the perceived age deviation tends to be

[1]We confirmed that y^* almost uniformly covers the range of our interest (i.e., age 1 to 70).

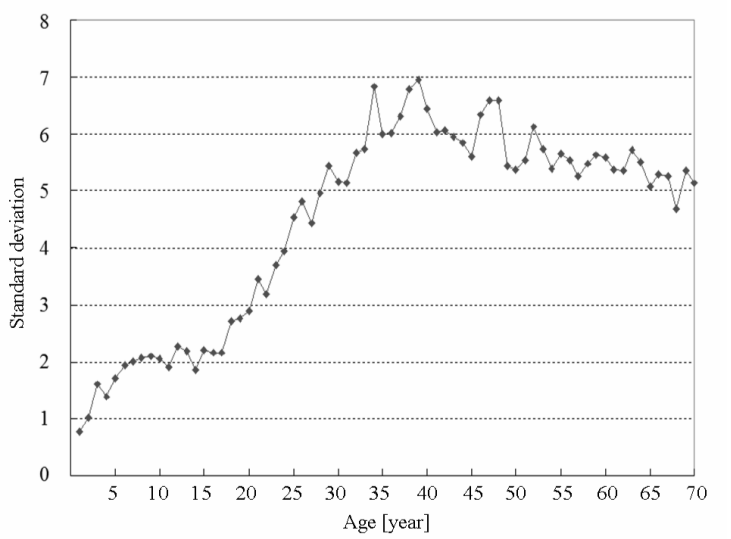

Figure 1: The relation between subjects' true age y^* (horizontal axis) and the standard deviation of perceived age (vertical axis).

small in younger age brackets and large in older age groups. This would well agree with our intuition considering the human growth process.

Now let us incorporate the above survey result into the perceived age estimation framework in Section 2.2. When the standard deviation is small (large), making an error is regarded as more (less) critical. This idea follows a similar line to the *Mahalanobis distance* [1], so it would be reasonable to incorporate the above survey result into the framework of *weighted regression analysis*. More precisely, weighting the goodness-of-fit term in Eq.(2) according to the inverse of the error variance optimally adjusts to the characteristics of human perception:

$$\frac{1}{l}\sum_{i=1}^{l}\frac{(y_i-f(\boldsymbol{x}_i))^2}{w(y_i)^2}+\lambda\|\boldsymbol{\alpha}\|^2+\frac{\mu}{4(l+u)^2}\sum_{i,j=1}^{l+u}W_{i,j}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2, \tag{4}$$

where $w(y)$ is the value given in Figure 1.

An important advantage of the above training method is that the solution can be obtained *analytically* by

$$\widehat{\boldsymbol{\alpha}}=\left(K^\top DK+l\lambda I_{l+u}+\frac{l\mu}{(l+u)^2}K^\top LK\right)^{-1}K^\top D\boldsymbol{y}, \tag{5}$$

where K is the $(l+u)\times(l+u)$ kernel Gram matrix whose (i,j)-th element is defined by $K_{i,j}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$. D is the $(l+u)\times(l+u)$ diagonal weight matrix with diagonal elements defined by $w(y_1)^{-2},\ldots,w(y_l)^{-2},0,\ldots,0$. L is the $(l+u)\times(l+u)$ Laplacian matrix whose (i,j)-th entry is defined by

$$L_{i,j}=\delta_{i,j}\sum_{j'=1}^{l+u}W_{i,j'}-W_{i,j},$$

where $\delta_{i,j}$ is the Kronecker delta. I_{l+u} denotes the $(l+u) \times (l+u)$ identity matrix. $\boldsymbol{y}$ is the $(l+u)$-dimensional label vector defined as $\boldsymbol{y} = (y_1, \ldots, y_l, 0, \ldots, 0)^\top$.

If u is very large (which would be the case in age prediction), computing the inverse of the $(l+u) \times (l+u)$ matrix in Eq.(5) is not tractable. To cope with this problem, reducing the number of kernels from $(l+u)$ to a smaller number b would be a realistic option, as explained in Section 2.2. Then the matrix K becomes an $(l+u) \times b$ rectangular matrix and the identity matrix in Eq.(5) becomes I_b. Thus the size of the matrix we need to invert becomes $b \times b$, which would be tractable when b is kept moderate. We may further reduce the computational cost by numerically computing the solution by a stochastic gradient-descent method.

2.4 Evaluation Criteria

Conventionally, the performance of an age prediction function $f(\boldsymbol{x})$ for test samples $\{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i^*)\}_{i=1}^m$ was evaluated by the mean absolute error (MAE) [8, 7, 3, 12]:

$$\mathrm{MAE} = \frac{1}{m} \sum_{i=1}^{m} \left| \widetilde{y}_i^* - f(\widetilde{\boldsymbol{x}}_i) \right|.$$

However, as explained in Section 2.3, this does not properly reflect human age perception characteristics. Here we propose to use the weighted criterion also for performance evaluation in experiments. More specifically, we evaluate the prediction performance by the weighted mean squared error (WMSE):

$$\mathrm{WMSE} = \frac{1}{m} \sum_{i=1}^{m} \frac{(\widetilde{y}_i^* - f(\widetilde{\boldsymbol{x}}_i))^2}{w(\widetilde{y}_i^*)^2}. \tag{6}$$

The smaller the value of WMSE is, the better the age prediction function is.

3 Empirical Evaluation

In this section, we apply the proposed age prediction method to in-house face-age datasets, and experimentally evaluate its performance.

3.1 Data Acquisition and Experimental Setup

Age prediction systems are often used in public places such as shopping malls or train stations. In order to make our experiments realistic, we collected face image samples from video sequences taken by ceiling-mounted surveillance cameras with depression angle 5–10 degrees. The recording method, image resolution, and the image size are diverse depending on the recording conditions—for example, some subjects were illuminated by dominant light sources, walking naturally, seated on a stool, and keeping their heads still. The subjects' facial expressions are typically subtle, switching between neutral and

Figure 2: Examples of face images.

smiling. We used a face detector for localizing the two eye-centers, and then rescaled the image to 64×64 pixels. Examples of face images are shown in Figure 2. Faces whose age ranges from 1 to 70 were used in our experiments.

As pre-processing, we extracted 100-dimensional features from the 64×64 face images using a neural network feature extractor proposed in [11, 10]. In total, we have 28500 face samples in our database. Among them, $u = 27000$ are treated as unlabeled samples and the remaining $m = 1500$ are used as test samples. From the 27000 unlabeled samples, we choose $l = 200$ samples to be labeled by active learning. The Gaussian-kernel variance σ^2 and the regularization parameters λ and μ were determined so that WMSE for the test data is minimized (i.e., they are optimally tuned). For manifold regularization, we fixed the nearest neighbor number and the decay rate of the similarity to $h = 5$ and $\gamma = 1$, respectively (see Eq.(3)).

3.2 Results

We applied the k-means clustering algorithm to 27000 unlabeled samples in the 4-dimensional or 10-dimensional PCA subspace and extracted 200 clusters. We chose 200 samples that are closest to the 200 cluster centroids and labeled them; then we trained a regressor using the weighted manifold method proposed in Section 2.3 with the 200 labeled samples and 5000 unlabeled samples randomly chosen from the pool of 26800 $(= 27000 - 200)$ unlabeled samples. We compared the above method with random sampling strategy. Figure 3 summarizes WMSE obtained by each method; in the comparison, we also included supervised regression where unlabeled samples were not used (i.e., $\mu = 0$).

Figure 3 shows that the proposed active learning method gave smaller WMSE than the random sampling strategy; the use of unlabeled samples for learning also improved

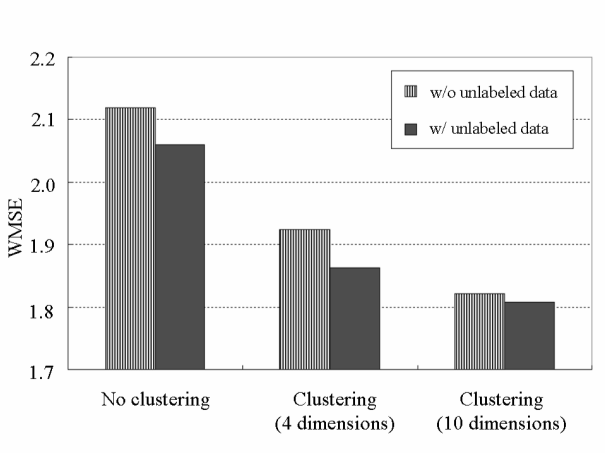

Figure 3: Comparison of WMSE (6).

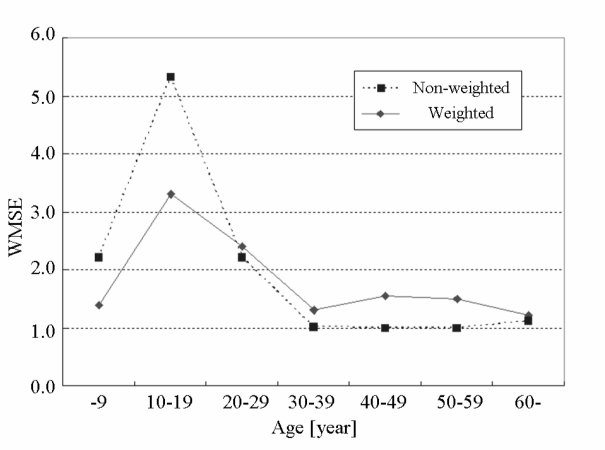

Figure 4: WMSE for each age-group.

the performance. Thus the proposed active learning method combined with manifold-based semi-supervised learning is shown to be effective for improving the age prediction performance.

In order to more closely understand the effect of age weighting, we investigated the prediction error for each age bracket. Figure 4 shows age-bracket-wise WMSE when the weighted learning method (see Eq.(4)) or the non-weighted learning method (see Eq.(2)) is used. The figure shows that the error in young age groups (less than 20 years old) is significantly reduced by the use of the age weights, which was shown to be highly important in practical human evaluation (see Section 2.3). On the other hand, the prediction error for middle/older age groups is slightly increased, but a small increase of the error in these age brackets was shown to be less significant in our questionnaire survey. Therefore, the experimental result indicates that our approach qualitatively improves the age prediction accuracy.

4 Conclusions

We introduced two novel ideas for perceived age estimation from face images: clustering-based active learning for reducing the sampling cost and taking into account the human age perception for improving the prediction accuracy.

We have incorporated the characteristics of human age perception as weights—error in younger age brackets is more serious than that in older age groups. On the other hand, our framework can accommodate *arbitrary* weights, which opens up new interesting research possibilities. Higher weights lead to better prediction in the corresponding age brackets, so we can improve the prediction accuracy of arbitrary age groups (but the price we have to pay for this is a performance decrease in other age brackets). This property could be useful, for example, in cigarettes and alcohol retail, where accuracy around 20 years old needs to be enhanced but accuracy in other age brackets is not so important. Another possible usage of our weighted regression framework is to combine learned functions obtained from several different age weights, which we would like to pursue in our future work.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Wiley, New York, 2001.

[2] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features", Proc. of IEEE Multimedia and Expo, pp.1383-1386, 2007

[3] X. Geng, Z. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation", Proc. of ACM International Conf. on Multimedia, pp.307-316, 2006.

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

[5] W. B. Horng, C. P. Lee, and C. W. Chen, "Classification of age groups based on facial features", Tamkang Journal of Science and Engineering, vol.4, no.3, pp.183-192, 2001.

[6] Y. H. Kwon, and N. V. Lobo, "Age classification from facial images", Computer Vision and Image Understanding, vol.74, no.1, pp.1-21, 1999.

[7] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation", IEEE Trans. on Systems, Man, and Cybernetics Part B, vol.34, no.1, pp.621-628, 2004.

[8] A. Lanitis, C. J. Taylor, and T. F. Cootes, (2002). "Toward automatic simulation of aging effects on face images", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.24, no.4, pp.442-455, 2002.

[9] V. Sindhwani, M. Belkin, and P. Niyogi, "The geometric basis of semi-supervised learning", in Semi-Supervised Learning, MIT Press, Cambridge, 2006.

[10] F. H. C. Tivive, and A. Bouzerdoum, "A shunting inhibitory convolutional neural network for gender classification", Proc. of International Conf. on Pattern Recognition, vol.4, pp.421-424, 2006.

[11] F. H. C. Tivive, and A. Bouzerdoumi, "A gender recognition system using shunting inhibitory convolutional neural networks", Proc. of International Joint Conf. on Neural Networks, pp.5336-5341, 2006.

[12] K. Ueki, M. Miya, T. Ogawa, and T. Kobayashi, "Class distance weighted locality preserving projection for automatic age estimation", Proc. of IEEE International Conf. on Biometrics: Theory, Applications and Systems, pp.1-5, 2008.