16

Perceived Age Estimation from Face Images

Kazuya Ueki[1], Yasuyuki Ihara[1] and Masashi Sugiyama[2]
[1]*NEC Soft, Ltd.*
[2]*Tokyo Institute of Technology*
Japan

1. Introduction

In recent years, demographic analysis in public places such as shopping malls and stations is attracting a great deal of attention. Such demographic information is useful for various purposes, e.g., designing effective marketing strategies and targeted advertisement based on customers' gender and age. For this reason, a number of approaches have been explored for age estimation from face images (Fu et al., 2007; Geng et al., 2006; Guo et al., 2009), and several databases became publicly available recently (FG-Net Aging Database, n.d.; Phillips et al., 2005; Ricanek & Tesafaye, 2006). It has been reported that age can be accurately estimated under controlled environment such as frontal faces, no expression, and static lighting conditions. However, it is not straightforward to achieve the same accuracy level in a real-world environment due to considerable variations in camera settings, facial poses, and illumination conditions. The recognition performance of age prediction systems is significantly influenced by such factors as the type of camera, camera calibration, and lighting variations. On the other hand, the publicly available databases were mainly collected in semi-controlled environments. For this reason, existing age prediction systems built upon such databases tend to perform poorly in a real-world environment.

In this chapter, we address the problem of perceived age estimation from face images, and describe our new approaches proposed in Ueki et al. (2010) and Ueki et al. (2011), which involve three novel aspects.

The first novelty of our proposed approaches 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.

The second 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 third contribution is to apply a recently proposed machine learning technique called *covariate shift adaptation* (Shimodaira, 2000; Sugiyama & Kawanabe, 2011; Sugiyama et al.,

2007; 2008) to alleviating lighting condition change between laboratory and practical environment.
Through real-world age estimation experiments, we demonstrate the usefulness of the proposed approaches.

2. Age estimation based on age perception characteristics

In this section, we mathematically formulate the problem of age estimation, and show how human age perception characteristics can be incorporated systematically.

2.1 Formulation

Throughout this chapter, we perform age estimation based not on subjects' real age, but on their *perceived* age. Thus, the 'true' age of the subject y is defined as the average perceived age evaluated by those who observed the subject's face images (the value is rounded-off to the nearest integer).
Let us consider a regression problem of estimating the age y^* of subject $\boldsymbol{x}$ (face features). Suppose we are given labeled training data

$$\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{l}.$$

We use the following kernel model for age regression.

$$f(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}}), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^\top$ is a model parameter, $^\top$ denotes the transpose, and $K(\boldsymbol{x}, \boldsymbol{x}')$ is a *positive definite kernel* (Schölkopf & Smola, 2002). We use the Gaussian kernel:

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

where σ^2 is the Gaussian variance.

A standard approach to learning the model parameter $\boldsymbol{\alpha}$ would be *regularized least-squares* (Hoerl & Kennard, 1970).

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{l}\sum_{i=1}^{l}(y_i^{\mathrm{tr}} - f(\boldsymbol{x}_i^{\mathrm{tr}};\boldsymbol{\alpha}))^2 + \lambda\|\boldsymbol{\alpha}\|^2\right], \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\lambda(>0)$ is the regularization parameter to avoid overfitting.
Below, we explain that merely using regularized least-squares is not appropriate in real-world perceived age prediction, and show how to cope with this problem.

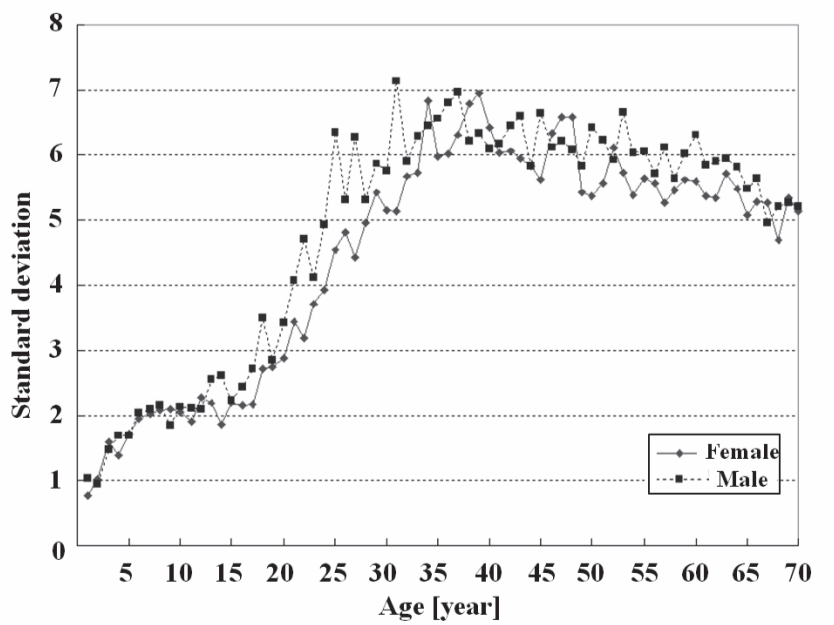

Fig. 1. The relation between subjects' true age y^* (horizontal axis) and the standard deviation of perceived age (vertical axis).

2.2 Incorporating age perception characteristics

Human age perception is known to have heterogeneous characteristics, e.g., 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.

In order to quantify this phenomenon, we investigated 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 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 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 described in Section 2.1. 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* (Duda et al., 2001), 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:

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{l} \sum_{i=1}^{l} \frac{(y_i^{\mathrm{tr}} - f(\boldsymbol{x}_i^{\mathrm{tr}}; \boldsymbol{\alpha}))^2}{w_{\mathrm{age}}(y_i^{\mathrm{tr}})^2} + \lambda \|\boldsymbol{\alpha}\|^2 \right], \tag{3}$$

where $w_{\mathrm{age}}(y)$ is the value given in Figure 1.

2.3 Evaluation criterion

Conventionally, the performance of an age prediction function $f(\boldsymbol{x})$ for test samples $\{(\boldsymbol{x}_j^{\mathrm{te}}, y_j^{\mathrm{te}})\}_{j=1}^{t}$ was evaluated by the mean absolute error (MAE) (Geng et al., 2006; Lanitis et al., 2004; 2002; Ueki et al., 2008):

$$\mathrm{MAE} = \frac{1}{t} \sum_{j=1}^{t} \left| y_j^{\mathrm{te}} - f(\boldsymbol{x}_j^{\mathrm{te}}) \right|.$$

However, as explained above, 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):

$$\mathrm{WMSE} = \frac{1}{t} \sum_{j=1}^{t} \frac{(y_j^{\mathrm{te}} - f(\boldsymbol{x}_j^{\mathrm{te}}))^2}{w_{\mathrm{age}}(y_j^{\mathrm{te}})^2}. \tag{4}$$

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

3. Semi-supervised approach

In this section, we give an active learning strategy and a semi-supervised age regression method within the age-weighting framework described in the previous section.

3.1 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 (Fu et al., 2007; Guo et al., 2008; Ueki et al., 2008). 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 (MacQueen, 1967) to a large number of unlabeled samples. Since clustering of high-dimensional data is often unreliable, we first apply *principal component analysis* (PCA) (Jolliffe, 1986) to the face images for dimension reduction, which is a well-justified preprocessing for k-means clustering (Ding & He, 2004). The proposed active learning strategy is summarized as follows.

1. For a set of d-dimensional unlabeled face image samples $\{\boldsymbol{X}_i\}_{i=1}^{n}$, we compute $\{\boldsymbol{x}_i\}_{i=1}^{n}$ of r ($\ll d$) dimensions by the PCA projection.
2. Using the k-means clustering algorithm, we compute the l ($\ll n$) cluster centroids $\{\boldsymbol{m}_i\}_{i=1}^{l}$.

3. We choose $\{\boldsymbol{x}_i^{\mathrm{tr}} \mid \boldsymbol{x}_i^{\mathrm{tr}} = \boldsymbol{x}_{\tau(i)}\}_{i=1}^{l}$ from $\{\boldsymbol{x}_i\}_{i=1}^{n}$ as samples to be labeled, where

$$\tau(i) = \underset{i'}{\operatorname{argmin}} \|\boldsymbol{x}_{i'} - \boldsymbol{m}_i\|,$$

and $\|\cdot\|$ denotes the Euclidean norm.

Let $\{y_i^{\mathrm{tr}}\}_{i=1}^{l}$ be the labels for $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}$, and let the remaining samples of size u $(= n - l)$ that were not chosen to be labeled be denoted as

$$\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=l+1}^{n} = \{\boldsymbol{x}_i\}_{i=1}^{n} \backslash \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}.$$

Thus, the first l training samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}$ are labeled, and the remaining u training samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=l+1}^{l+u}$ are unlabeled.

3.2 Semi-supervised age regression with manifold regularization

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 (Sindhwani et al., 2006).

For age regression, we use the following kernel model:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l+u} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}}), \tag{5}$$

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 included $(l + u)$ kernels in the kernel regression model (5), but u can be very large in age prediction. In practice, we may only use c $(\leq u)$ elements randomly chosen from the set $\{k(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}})\}_{i=l+1}^{l+u}$ for reducing the computational cost; then the total number of basis functions is reduced to $b = l + c$. However, we stick to Eq.(5) below for keeping the explanation simple.

We employ a manifold regularizer (Sindhwani et al., 2006) in our training criterion, i.e., the parameter $\boldsymbol{\alpha}$ is learned so that the following criterion is minimized.

$$\frac{1}{l}\sum_{i=1}^{l} \frac{(y_i^{\mathrm{tr}} - f(\boldsymbol{x}_i^{\mathrm{tr}}))^2}{w_{\mathrm{age}}(y_i^{\mathrm{tr}})^2} + \lambda\|\boldsymbol{\alpha}\|^2 + \frac{\mu}{4(l+u)^2}\sum_{i,i'=1}^{l+u} A_{i,i'}(f(\boldsymbol{x}_i^{\mathrm{tr}}) - f(\boldsymbol{x}_{i'}^{\mathrm{tr}}))^2, \tag{6}$$

where λ and μ are non-negative regularization parameters. $A_{i,i'}$ represents the affinity between $\boldsymbol{x}_i^{\mathrm{tr}}$ and $\boldsymbol{x}_{i'}^{\mathrm{tr}}$, which is defined by

$$A_{i,i'} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2}{2\nu^2}\right) \tag{7}$$

if $\boldsymbol{x}_i^{\mathrm{tr}}$ is a h-nearest neighbor of $\boldsymbol{x}_{i'}^{\mathrm{tr}}$ or vice versa; otherwise $A_{i,i'} = 0$.

The first term in Eq.(6) 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 $A_{i,i'}$ tends to take large values if $\boldsymbol{x}_i^{\mathrm{tr}}$ and $\boldsymbol{x}_{i'}^{\mathrm{tr}}$ 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.

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

$$\widehat{\boldsymbol{\alpha}} = \left(K^\top D K + l\lambda I_{l+u} + \frac{l\mu}{(l+u)^2} K^\top L K \right)^{-1} K^\top D \boldsymbol{y}, \tag{8}$$

where K is the $(l+u) \times (l+u)$ kernel Gram matrix whose (i,i')-th element is defined by

$$K_{i,i'} = k(\boldsymbol{x}_i^{\mathrm{tr}}, \boldsymbol{x}_{i'}^{\mathrm{tr}}).$$

D is the $(l+u) \times (l+u)$ diagonal weight matrix with diagonal elements defined by

$$w_{\mathrm{age}}(y_1^{\mathrm{tr}})^{-2}, \ldots, w_{\mathrm{age}}(y_l^{\mathrm{tr}})^{-2}, 0, \ldots, 0.$$

L is the $(l+u) \times (l+u)$ Laplacian matrix whose (i,i')-th entry is defined by

$$L_{i,i'} = \delta_{i,i'} \left(\sum_{i''=1}^{l+u} A_{i,i''} \right) - A_{i,i'},$$

where $\delta_{i,i'}$ 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

$$\boldsymbol{y} = (y_1^{\mathrm{tr}}, \ldots, y_l^{\mathrm{tr}}, 0, \ldots, 0)^\top.$$

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.(8) 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 above. Then the matrix K becomes an $(l+u) \times b$ rectangular matrix and the identity matrix in Eq.(8) 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-decent method* (Amari, 1967).

3.3 Empirical evaluation

Here, we apply the above age prediction method to in-house face-age datasets, and experimentally evaluate its performance.

3.3.1 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 were typically subtle, switching between neutral and 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 Tivive & Bouzerdoumi (2006a) and

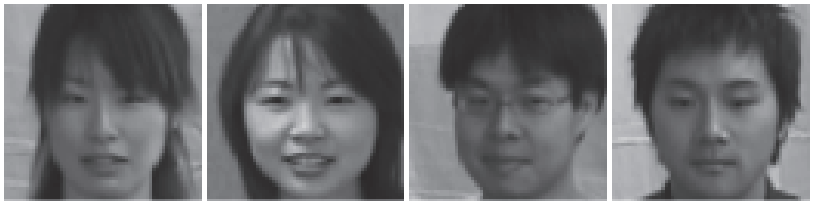

Fig. 2. Examples of face images.

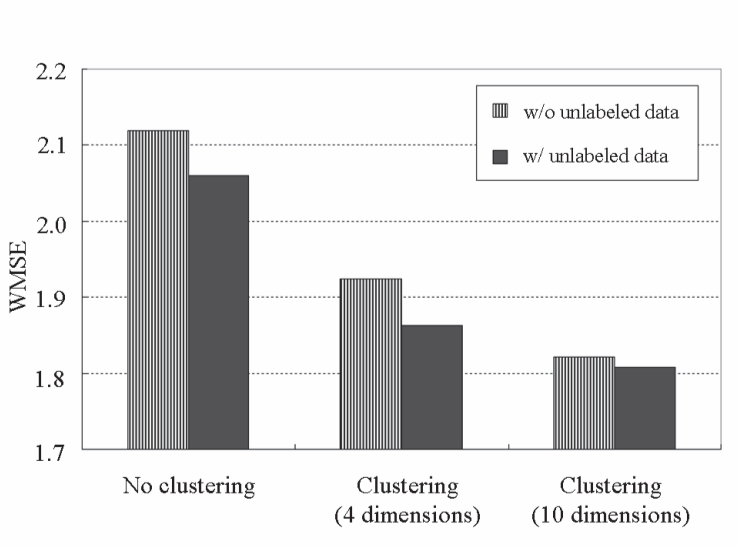

Fig. 3. Comparison of WMSE Eq.(4).

Tivive & Bouzerdoumi (2006b). In total, we have 28500 face samples in our database. Among them, $n = 27000$ are treated as unlabeled samples and the remaining $t = 1500$ 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 number of nearest neighbors and the decay rate of the similarity to $h = 5$ and $\nu = 1$, respectively (see Eq.(7)).

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-regularization method described in Section 2.2 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., $\mu = 0$).

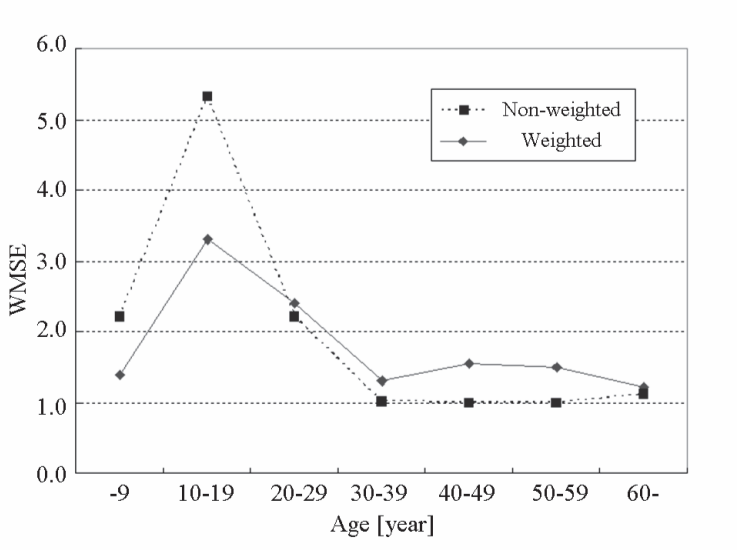

Fig. 4. WMSE for each age-group.

Figure 3 shows that the proposed active learning method gave smaller WMSE than the random sampling strategy; the use of unlabeled samples also improved 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 age-weighted learning method or the non-weighted learning method 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 2.2). 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. Coping with lighting condition change

In this section, we consider another semi-supervised learning setup where training and test samples follow different distributions. Such a situation often happens in real-world age prediction tasks, and we describe a systematic method to cope with such distribution change.

4.1 Lighting condition change as covariate shift

When designing age estimation systems, the environment of recording training face images is often different from the test environment in terms of lighting conditions. Typically, training data are recorded indoors such as a studio with appropriate illumination. On the other hand, in a real-world environment, lighting conditions have considerable varieties, e.g., strong

sunlight might be cast from a side of the face or there is no enough light. In such situations, age estimation accuracy is significantly degraded.

Let $p_{\mathrm{tr}}(\boldsymbol{x})$ be the probability density function of training face features and $p_{\mathrm{te}}(\boldsymbol{x})$ be the probability density function of test face features. When these two densities are different, it would be natural to emphasize the influence of training samples $(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})$ which have high similarity to data in the test environment. Such adjustment can be systematically carried out as follows (Shimodaira, 2000; Sugiyama & Kawanabe, 2011; Sugiyama et al., 2007; 2008):

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{l} \sum_{i=1}^{l} w_{\mathrm{imp}}(\boldsymbol{x}_i^{\mathrm{tr}}) \frac{(y_i^{\mathrm{tr}} - f(\boldsymbol{x}_i^{\mathrm{tr}}; \boldsymbol{\alpha}))^2}{w_{\mathrm{age}}(y_i^{\mathrm{tr}})^2} + \lambda \|\boldsymbol{\alpha}\|^2 \right], \tag{9}$$

i.e., the goodness-of-fit term in Eq.(3) is weighted according to the *importance function* (Fishman, 1996):

$$w_{\mathrm{imp}}(\boldsymbol{x}) = \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})}.$$

The solution of Eq.(9) can be obtained analytically by

$$\widehat{\boldsymbol{\alpha}} = (K^\top W K + l\lambda I_l)^{-1} K^\top W \boldsymbol{y}, \tag{10}$$

where K is the kernel matrix whose (i, i')-th element is defined by

$$K_{i,i'} = K(\boldsymbol{x}_i^{\mathrm{tr}}, \boldsymbol{x}_{i'}^{\mathrm{tr}}),$$

W is the l-dimensional diagonal matrix with (i, i)-th diagonal element defined by

$$W_{i,i} = \frac{w_{\mathrm{imp}}(\boldsymbol{x}_i^{\mathrm{tr}})}{w_{\mathrm{age}}(y_i^{\mathrm{tr}})^2},$$

I_l is the l-dimensional identity matrix, and

$$\boldsymbol{y} = (y_1^{\mathrm{tr}}, \ldots, y_l^{\mathrm{tr}})^\top.$$

When the number of training data l is large, we may reduce the number of kernels in Eq.(1) so that the inverse matrix in Eq.(10) can be computed with limited memory; or we may compute the solution numerically by a stochastic gradient-decent method (Amari, 1967).

4.2 Importance-Weighted Cross-Validation (IWCV)

In supervised learning, the choice of models (for example, the basis functions and the regularization parameter) is crucial for obtaining better prediction performance. *Cross-validation* (CV) would be one of the most popular techniques for model selection (Stone, 1974). CV has been shown to give an *almost* unbiased estimate of the generalization error with finite samples (Schölkopf & Smola, 2002), but such almost unbiasedness is no longer fulfilled under covariate shift.

To cope with this problem, a variant of CV called *importance-weighted CV* (IWCV) has been proposed (Sugiyama et al., 2007). Let us randomly divide the training set

$$\mathcal{Z} = \{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{l}$$

into M disjoint non-empty subsets $\{\mathcal{Z}_m\}_{m=1}^{M}$ of (approximately) the same size. Let $f_{\mathcal{Z}_m}(\boldsymbol{x})$ be a function learned from $\mathcal{Z}\backslash\mathcal{Z}_m$ (i.e., without $\mathcal{Z}_m$). Then the *M-fold IWCV* (IWCV) estimate of the generalization error is given by

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{|\mathcal{Z}_m|}\sum_{(\boldsymbol{x},y)\in\mathcal{Z}_m}\frac{w_{\mathrm{imp}}(\boldsymbol{x})}{w_{\mathrm{age}}(y)^2}(f_{\mathcal{Z}_m}(\boldsymbol{x})-y)^2,$$

where $|\mathcal{Z}_m|$ denotes the number of samples in the subset $\mathcal{Z}_m$.

It was proved that IWCV gives an *almost* unbiased estimate of the generalization error even under covariate shift (Sugiyama et al., 2007).

4.3 Kullback-Leibler Importance Estimation Procedure (KLIEP)

In order to compute the solution (10) or performing IWCV, we need to know the values of the importance weights

$$w_{\mathrm{imp}}(\boldsymbol{x}_i^{\mathrm{tr}}) = \frac{p_{\mathrm{te}}(\boldsymbol{x}_i^{\mathrm{tr}})}{p_{\mathrm{tr}}(\boldsymbol{x}_i^{\mathrm{tr}})},$$

which include two probability densities $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$.

In addition to the training samples $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{l}$, suppose we are given unlabeled test samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$ which are drawn independently from the density $p_{\mathrm{te}}(\boldsymbol{x})$. Then, performing density estimation of $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ gives an approximation of $w_{\mathrm{imp}}(\boldsymbol{x})$. However, since density estimation is a hard problem, the two-stage approach of first estimating $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ and then taking their ratio may not be reliable.

Here we describe a method called *Kullback-Leibler Importance Estimation Procedure* (KLIEP) (Sugiyama et al., 2008), which allows us to directly estimate the importance function $w_{\mathrm{imp}}(\boldsymbol{x})$ without going through density estimation of $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$.

Let us model $w_{\mathrm{imp}}(\boldsymbol{x})$ using the following model:

$$\widehat{w}_{\mathrm{imp}}(\boldsymbol{x}) = \sum_{k=1}^{b}\beta_k \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_k\|^2}{2\gamma^2}\right), \tag{11}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b)^{\top}$ is a parameter, and $\{\boldsymbol{c}_k\}_{k=1}^{b}$ is a subset of test input samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$ĄD Using the model $\widehat{w}_{\mathrm{imp}}(\boldsymbol{x})$, we can estimate the test input density $p_{\mathrm{te}}(\boldsymbol{x})$ by

$$\widehat{p}_{\mathrm{te}}(\boldsymbol{x}) = \widehat{w}_{\mathrm{imp}}(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x}). \tag{12}$$

We determine the parameter $\boldsymbol{\beta}$ in the model (12) so that the Kullback-Leibler divergence from p_{te} to $\widehat{p}_{\mathrm{te}}$ is minimized:

$$KL(p_{\mathrm{te}}\|\widehat{p}_{\mathrm{te}}) = \int p_{\mathrm{te}}(\boldsymbol{x}) \log \frac{p_{\mathrm{te}}(\boldsymbol{x})}{\widehat{p}_{\mathrm{te}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$
$$= \int p_{\mathrm{te}}(\boldsymbol{x}) \log \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} - \int p_{\mathrm{te}}(\boldsymbol{x}) \log \widehat{w}_{\mathrm{imp}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Since the first term is a constant with respect to the parameter $\boldsymbol{\beta}$, we ignore it and define the second term as

$$\mathrm{KL}' = \int p_{\mathrm{te}}(\boldsymbol{x}) \log \widehat{w}_{\mathrm{imp}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

We would like to determine the parameter $\boldsymbol{\beta}$ so that KL' is maximized. Let us impose $\widehat{w}_{\mathrm{imp}}(\boldsymbol{x})$ to be non-negative and normalized. Then we obtain the following convex optimization problem:

$$\max_{\boldsymbol{\beta}} \quad \left[\sum_{j=1}^{t} \log \left(\sum_{k=1}^{b} \beta_k \exp \left(-\frac{\|\boldsymbol{x}_j^{\mathrm{te}} - \boldsymbol{c}_k\|^2}{2\gamma^2} \right) \right) \right]$$
$$\text{s.t.} \quad \begin{cases} \beta_k \geq 0 \ \text{ for } k = 1, \ldots, b, \\ \dfrac{1}{l} \displaystyle\sum_{i=1}^{l} \left(\sum_{k=1}^{b} \beta_k \exp \left(-\frac{\|\boldsymbol{x}_i^{\mathrm{tr}} - \boldsymbol{c}_k\|^2}{2\gamma^2} \right) \right) = 1. \end{cases}$$

This is a convex optimization problem and the global solution—which tends to be sparse (Boyd & Vandenberghe, 2004)—can be obtained, e.g., by simply performing gradient ascent and feasibility satisfaction iteratively. A pseudo code of KLIEP is described in Table 1.

Input: Kernel width γ, training inputs $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}$, and test inputs $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$
Output: $\widehat{w}(\boldsymbol{x})$

Randomly choose $\{\boldsymbol{c}_k\}_{k=1}^{b}$ from $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$;
$B_{j,k} \leftarrow \exp \left(-\dfrac{\|\boldsymbol{x}_j^{\mathrm{te}} - \boldsymbol{c}_k\|^2}{(2\gamma^2)} \right)$
$b_k \leftarrow \dfrac{1}{l} \displaystyle\sum_{i=1}^{l} \exp \left(-\dfrac{\|\boldsymbol{x}_i^{\mathrm{tr}} - \boldsymbol{c}_k\|^2}{(2\gamma^2)} \right)$;
Initialize $\boldsymbol{\beta}$ $(> \mathbf{0})$ and ε $(0 < \varepsilon \ll 1)$;
Repeat until convergence
    $\boldsymbol{\beta} \leftarrow \varepsilon B^{\top}(\mathbf{1}./B\boldsymbol{\beta})$;
    $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + (1 - \boldsymbol{b}^{\top}\boldsymbol{\beta})\boldsymbol{b}/(\boldsymbol{b}^{\top}\boldsymbol{b})$;
    $\boldsymbol{\beta} \leftarrow \max(\mathbf{0}, \boldsymbol{\beta})$;
    $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}/(\boldsymbol{b}^{\top}\boldsymbol{\beta})$;
end

Table 1. Pseudo code of KLIEP. './' indicates the element-wise division. Inequalities and the 'max' operation for vectors are applied in an element-wise manner.

The tuning parameter γ in KLIEP can be optimized based on *cross-validation* (CV) as follows (Sugiyama et al., 2008). First, divide the test samples $\mathcal{X}^{\mathrm{te}} = \{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$ into M disjoint subsets

Input: Kernel width candidates $\{\gamma\}$, training inputs $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}$, and test inputs $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$
Output: $\widehat{w}(\boldsymbol{x})$

Split $\mathcal{X}^{\mathrm{te}} = \{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$ into M disjoint subsets $\{\mathcal{X}_m^{\mathrm{te}}\}_{m=1}^{M}$;
for each model γ
  for each split $m = 1, \ldots, M$
    $\widehat{w}_{\mathcal{X}_m^{\mathrm{te}}}(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(\gamma, \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}, \mathcal{X}^{\mathrm{te}} \backslash \mathcal{X}_m^{\mathrm{te}})$;
    $\widehat{\mathrm{KL}}'_m(\gamma) \longleftarrow \frac{1}{|\mathcal{X}_m^{\mathrm{te}}|} \sum_{\boldsymbol{x} \in \mathcal{X}_m^{\mathrm{te}}} \log \widehat{w}_{\mathcal{X}_m^{\mathrm{te}}}(\boldsymbol{x})$;
  end
  $\widehat{\mathrm{KL}}'(\gamma) \longleftarrow \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{KL}}'_m(\gamma)$;
end
$\widehat{\gamma} \longleftarrow \underset{\gamma}{\mathrm{argmax}}\, \widehat{\mathrm{KL}}'(\gamma)$;
$\widehat{w}(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(\widehat{\gamma}, \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{l}, \mathcal{X}^{\mathrm{te}})$;

Table 2. Pseudo code of CV-based model selection for KLIEP.

$\{\mathcal{X}_m^{\mathrm{te}}\}_{m=1}^{M}$ of (approximately) the same size. Then obtain an importance estimate $\widehat{w}_{\mathcal{X}_m^{\mathrm{te}}}(\boldsymbol{x})$ from $\mathcal{X}^{\mathrm{te}} \backslash \mathcal{X}_m^{\mathrm{te}}$ (i.e., without $\mathcal{X}_m^{\mathrm{te}}$), and approximate KL′ using $\mathcal{X}_m^{\mathrm{te}}$ as

$$\widehat{\mathrm{KL}}'_r := \frac{1}{|\mathcal{X}_m^{\mathrm{te}}|} \sum_{\boldsymbol{x} \in \mathcal{X}_m^{\mathrm{te}}} \log \widehat{w}_{\mathcal{X}_m^{\mathrm{te}}}(\boldsymbol{x}).$$

This procedure is repeated for $m = 1, \ldots, M$, and the average $\widehat{\mathrm{KL}}'$ is used as an estimate of KL′:

$$\widehat{\mathrm{KL}}' := \frac{1}{m} \sum_{m=1}^{M} \widehat{\mathrm{KL}}'_m. \tag{13}$$

For model selection, we compute $\widehat{\mathrm{KL}}'$ for all model candidates (the Gaussian kernel width γ in the current setting), and choose the one that minimizes $\widehat{\mathrm{KL}}'$. A pseudo code of the CV procedure is summarized in Figure 2.

One of the potential limitations of CV in general is that it is not reliable in small sample cases since data splitting by CV further reduces the sample size. On the other hand, in our CV procedure, the data splitting is performed only over the *test input samples* $\mathcal{X}^{\mathrm{te}} = \{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{t}$, not over the training samples. Therefore, even when the number of training samples is small, our CV procedure does not suffer from the small sample problem as long as a large number of test input samples are available.

4.4 Empirical evaluation

Here, we experimentally evaluate the performance of the proposed method using in-house face-age datasets.

We use the face images recorded under 17 different lighting conditions: for instance, average illuminance from above is approximately 1000 lux and 500 lux from the front in the standard lighting condition, 250 lux from above and 125 lux from the front in the dark setting, and 190 lux from left and 750 lux from right in another setting (see Figure 5). Note that these

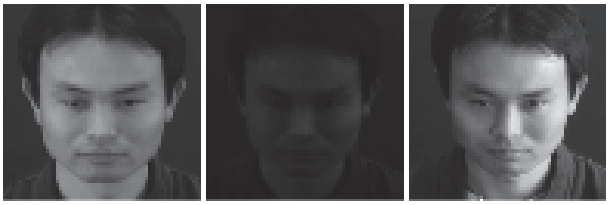

Fig. 5. Examples of face images under different lighting conditions (left: standard lighting, middle: dark, right: strong light from a side)

17 lighting conditions are diverse enough to cover real-world lighting conditions. Images were recorded as movies with camera at depression angle 15 degrees. The number of subjects is approximately 500 (250 for each gender). We used a face detector for localizing the two eye-centers, and then rescaled the image to 64×64 pixels. The number of face images in each environment is about 2500 (5 face images $\times$ 500 subjects).

As pre-processing, a neural network feature extractor (Tivive & Bouzerdoumi, 2006a;b) was used to extract 100-dimensional features from 64×64 face images. We constructed the male/female age prediction models only using male/female data, assuming that gender classification had been correctly carried out.

We split the 250 subjects into the *training set* (200 subjects) and the *test set* (50 subjects). The training set was used for training the kernel regression model (1), and the test set was used for evaluating its generalization performance. For the test samples $\{(\boldsymbol{x}_j^{\mathrm{te}}, y_j^{\mathrm{te}})\}_{j=1}^{t}$ taken from the test set in the environment with strong light from a side, age-weighted mean square error (WMSE)

$$\mathrm{WMSE} = \frac{1}{t}\sum_{j=1}^{t} \frac{(y_j^{\mathrm{te}} - f(\boldsymbol{x}_j^{\mathrm{te}}; \widehat{\boldsymbol{\alpha}}))^2}{w_{\mathrm{age}}(y_j^{\mathrm{te}})^2}$$

was calculated as a performance measure. The training and test sets were shuffled 5 times in such a way that each subject was selected as a test sample once. The final performance was evaluated based on the average WMSE over the 5 trials.

We compared the performance of the proposed method with the two baseline methods:

Baseline method 1: Training samples were taken only from the standard lighting condition and age-weighted regularized least-squares (3) was used for training.

Baseline method 2: Training samples were taken from all 17 different lighting conditions and age-weighted regularized least-squares (3) was used for training.

The importance weights were not used in these baseline methods. The Gaussian width σ and the regularization parameter λ were determined based on 4-fold CV over WMSE, i.e., the training set was further divided into a training part (150 subjects) and a validation part (50 subjects).

In the proposed method, training samples were taken from all 17 different lighting conditions (which is the same as the baseline method 2). The importance weights were estimated by KLIEP using the training samples and additional *unlabeled* test samples; the hyper-parameter γ in KLIEP was determined based on 2-fold CV (Sugiyama et al., 2008). We then computed the average importance score over different samples for each lighting condition and used the average importance score for training the regression model. The Gaussian width σ and the

	Male	Female
Baseline method 1	2.83	6.51
Baseline method 2	2.64	4.40
Proposed method	**2.54**	**3.90**

Table 3. The test performance measured by WMSE.

regularization parameter λ in the regression model were determined based on 4-fold IWCV (Sugiyama et al., 2007).
Table 3 summarizes the experimental results, showing that, for both male and female data, the baseline method 2 is better than the baseline method 1 and the proposed method is better than the baseline method 2. This illustrates the effectiveness of the proposed method. Note that WMSE for female subjects is substantially larger than that for male subjects. The reason for this would be that female subjects tend to have more diversity such as short/long hair and with/without makeup, which makes prediction harder (Ueki et al., 2008).

5. Conclusion

We introduced three novel ideas for perceived age estimation from face images: taking into account the human age perception for improving the prediction accuracy (Section 2), clustering-based active learning for reducing the sampling cost (Section 3), and alleviating the influence of lighting condition change (Section 4).
We have incorporated the characteristics of human age perception as weights—error in younger age brackets is treated as 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.
Lighting condition change is one of the critical causes of performance degradation in age prediction from face images. In this chapter, we proposed to employ a machine learning technique called *covariate shift adaptation* for alleviating the influence of lighting condition change. We demonstrated the effectiveness of our proposed method through real-world perceived age prediction experiments.
In the experiments in Section 4.4, test samples were collected from a particular lighting condition, and samples from the same lighting condition were also included in the training set. Although we believe this setup to be practical, it would be interesting to evaluate the performance of the proposed method when no overlap in the lighting conditions exists between training and test data. Following the theoretical study by Cortes et al. (2010) would be a promising direction for further addressing this issue.
In principle, the covariate shift framework allows us to incorporate not only lighting condition change but also various types of environment change such as face pose variation and camera setting change. In our future work, we will investigate whether the proposed approach is still useful in such challenging scenarios.

Recently, novel approaches to importance estimation for high-dimensional problems have been explored (Kanamori et al., 2009; Sugiyama, Kawanabe & Chui, 2009; Sugiyama, Yamada, von Bünau, Suzuki, Kanamori & Kawanabe, 2011; Yamada et al., 2010). In our future work, we would like to incorporating these new ideas into our framework of perceived age estimation, and see how the prediction performance can be further improved.

In the context of covariate shift adaptation, the importance weights played a central role for systematically adjusting the difference of distributions in the training and test phases. Beyond covariate shift adaptation, it has been shown recently that the ratio of probability densities can be used for solving various machine learning tasks (Sugiyama, Kanamori, Suzuki, Hido, Sese, Takeuchi & Wang, 2009; Sugiyama, Suzuki & Kanamori, 2012). This novel machine learning framework includes multi-task learning (Bickel et al., 2008; Simm et al., 2011), privacy-preserving data mining (Elkan, 2010), outlier detection (Hido et al., 2011), conditional density estimation (Sugiyama et al., 2010), and probabilistic classification (Sugiyama, 2010). Furthermore, mutual information—which plays a central role in information theory (Cover & Thomas, 2006)—can be estimated via density ratio estimation (Suzuki et al., 2008; Suzuki, Sugiyama & Tanaka, 2009). Since mutual information is a measure of statistical independence between random variables, density ratio estimation can be used also for variable selection (Suzuki, Sugiyama, Kanamori & Sese, 2009), dimensionality reduction (Suzuki & Sugiyama, 2010), independent component analysis (Suzuki & Sugiyama, 2011), and causal inference (Yamada & Sugiyama, 2010). In our future work, we will apply those novel machine learning tools in perceived age prediction.

6. References

Amari, S. (1967). Theory of adaptive pattern classifiers, *IEEE Transactions on Electronic Computers* EC-16(3): 299–307.

Bickel, S., Bogojeska, J., Lengauer, T. & Scheffer, T. (2008). Multi-task learning for HIV therapy screening, *in* A. McCallum & S. Roweis (eds), *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)*, pp. 56–63.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press, Cambridge, UK.

Cortes, C., Mansour, Y. & Mohri, M. (2010). Learning bounds for importance weighting, *in* J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor & A. Culotta (eds), *Advances in Neural Information Processing Systems 23*, pp. 442–450.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edn, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Ding, C. & He, X. (2004). K-means clustering via principal component analysis, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML2004)*, ACM Press, New York, NY, USA, pp. 225–232.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification*, Wiley, New York.

Elkan, C. (2010). Privacy-preserving data mining via importance weighting, *ECML/PKDD Workshop on Privacy and Security Issues in Data Mining and Machine Learning (PSDML2010)*.

FG-Net Aging Database (n.d.).
URL: *http://sting.cycollege.ac.cy/ alanitis/fgnetaging/*

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, Berlin, Germany.

Fu, Y., Xu, Y. & Huang, T. S. (2007). Estimating human age by manifold analysis of face pictures and regression on aging features, *Proc. of IEEE Multimedia and Expo* pp. 1383–1386.

Geng, X., Zhou, Z., Zhang, Y., Li, G. & Dai, H. (2006). Learning from facial aging patterns for automatic age estimation, *Proc. of ACM International Conf. on Multimedia* pp. 307–316.

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

Guo, G., Mu, G., Fu, Y., Dyer, C. & Huang, T. (2009). A study on automatic age estimation using a large database., *International Conference on Computer Vision in Kyoto (ICCV 2009)* pp. 1986–1991.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M. & Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation, *Knowledge and Information Systems* 26(2): 309–336.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12(3): 55–67.

Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer-Verlag, New York, NY, USA.

Kanamori, T., Hido, S. & Sugiyama, M. (2009). A least-squares approach to direct importance estimation, *Journal of Machine Learning Research* 10: 1391–1445.

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

Lanitis, A., Taylor, C. J. & Cootes, T. F. (2002). Toward automatic simulation of aging effects on face images, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(4): 442–455.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, CA., USA, pp. 281–297.

Phillips, P. J., Flynn, P. J., Scruggs, W. T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J. & Worek, W. J. (2005). Overview of the face recognition grand challenge., *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)* pp. 947–954.

Ricanek, K. J. & Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression., Proceedings of the IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR 2006) pp. 341–345.

Schölkopf, B. & Smola, A. J. (2002). Learning with Kernels, MIT Press, Cambridge.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function, *Journal of Statistical Planning and Inference* 90(2): 227–244.

Simm, J., Sugiyama, M. & Kato, T. (2011). Computationally efficient multi-task learning with least-squares probabilistic classifiers, *IPSJ Transactions on Computer Vision and Applications*, 3: 1–8.

Sindhwani, V., Belkin, M. & Niyogi, P. (2006). The geometric basis of semi-supervised learning, *Semi-Supervised Learning*, MIT Press, Cambridge.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B* 36: 111–147.

Sugiyama, M. (2010). Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting, *IEICE Transactions on Information and Systems* E93-D(10): 2690–2701.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I. & Wang, L. (2009). A density-ratio framework for statistical data processing, *IPSJ Transactions on Computer Vision and Applications* 1: 183–208.

Sugiyama, M. & Kawanabe, M. (2011). *Covariate Shift Adaptation: Toward Machine Learning in Non-Stationary Environments*, MIT Press, Cambridge, MA, USA. to appear.

Sugiyama, M., Kawanabe, M. & Chui, P. L. (2009). Dimensionality reduction for density ratio estimation in high-dimensional spaces, *Neural Networks* 23(1): 44–59.

Sugiyama, M., Krauledat, M. & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation, *Journal of Machine Learning Research* 8: 985–1005.

Sugiyama, M., Suzuki, T. & Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning: A Versatile Tool for Statistical Data Processing*, Cambridge University Press, Cambridge, UK. to appear.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation, *Annals of the Institute of Statistical Mathematics* 60(4): 699–746.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H. & Okanohara, D. (2010). Least-squares conditional density estimation, *IEICE Transactions on Information and Systems* E93-D(3): 583–594.

Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T. & Kawanabe, M. (2011). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search, *Neural Networks*, 24(2): 183–198.

Suzuki, T. & Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation, *in* Y. W. Teh & M. Tiggerington (eds), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, Vol. 9 of *JMLR Workshop and Conference Proceedings*, Sardinia, Italy, pp. 804–811.

Suzuki, T. & Sugiyama, M. (2011). Least-squares independent component analysis, *Neural Computation* 23(1): 284–301.

Suzuki, T., Sugiyama, M., Kanamori, T. & Sese, J. (2009). Mutual information estimation reveals global associations between stimuli and biological processes, *BMC Bioinformatics* 10(1): S52.

Suzuki, T., Sugiyama, M., Sese, J. & Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation, *in* Y. Saeys, H. Liu, I. Inza, L. Wehenkel & Y. V. de Peer (eds), *Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008)*, Vol. 4 of *JMLR Workshop and Conference Proceedings*, Antwerp, Belgium, pp. 5–20.

Suzuki, T., Sugiyama, M. & Tanaka, T. (2009). Mutual information approximation via maximum likelihood estimation of density ratio, *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)*, Seoul, Korea, pp. 463–467.

Tivive, F. H. C. & Bouzerdoumi, A. (2006a). A gender recognition system using shunting inhibitory convolutional neural networks, *Proc. of International Joint Conf. on Neural Networks* pp. 5336–5341.

Tivive, F. H. C. & Bouzerdoumi, A. (2006b). A shunting inhibitory convolutional neural network for gender classification, *Proc. of International Conf. on Pattern Recognition* 4: 421–424.

Ueki, K., Miya, M., Ogawa, T. & Kobayashi, T. (2008). Class distance weighted locality preserving projection for automatic age estimation, *Proc. of IEEE International Conf. on Biometrics: Theory, Applications and Systems* pp. 1–5.

Ueki, K., Sugiyama, M. & Ihara, Y. (2010). A semi-supervised approach to perceived age prediction from face images, *IEICE Transactions on Information and Systems* E93-D(10): 2875–2878.

Ueki, K., Sugiyama, M. & Ihara, Y. (2011). Lighting condition adaptation for perceived age estimation, *IEICE Transactions on Information and Systems* E94-D(2): 392–395.

Yamada, M. & Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, The AAAI Press, Atlanta, Georgia, USA, pp. 643–648.

Yamada, M., Sugiyama, M., Wichern, G. & Simm, J. (2010). Direct importance estimation with a mixture of probabilistic principal component analyzers, *IEICE Transactions on Information and Systems* E93-D(10): 2846–2849.