Simultaneous Dimensionality Reduction and Human Age Estimation via Kernel Partial Least Squares Regression

Guodong Guo*
West Virginia University
Lane Dept. of CSEE, Morgantown, WV
guodong.guo@mail.wvu.edu

Guowang Mu
Hebei University of Technology
School of Science, Tianjin, P. R. China
GuowangMu@gmail.com

Abstract

Human age estimation has recently become an active research topic in computer vision and pattern recognition, because of many potential applications in reality. In this paper we propose to use the kernel partial least squares (KPLS) regression for age estimation. The KPLS (or linear PLS) method has several advantages over previous approaches: (1) the KPLS can reduce feature dimensionality and learn the aging function simultaneously *in a single learning framework, instead of performing each task separately using different techniques; (2) the KPLS can find a small number of latent variables, e.g., 20, to project thousands of features into a very low-dimensional subspace, which may have great impact on real-time applications; and (3) the KPLS regression has an output vector that can contain multiple labels, so that several related problems, e.g., age estimation, gender classification, and ethnicity estimation can be solved altogether. This is the first time that the kernel PLS method is introduced and applied to solve a regression problem in computer vision with high accuracy. Experimental results on a very large database show that the KPLS is significantly better than the popular SVM method, and outperform the state-of-the-art approaches in human age estimation.*

1. Introduction

Recently human age estimation has become an active research in computer vision and pattern recognition, because of many potential applications in the real world. Age estimation is useful for creating an age-specific human-computer interaction (AS-HCI) system [7] or electronic customer relationship management (ECRM) [1].

Some earlier work on age estimation [13, 18, 12, 14, 28, 7] performed on small databases, demonstrating the possibility of computer-based age estimation on human faces. Later work [6, 5, 35, 33, 8, 11] used larger databases, such as the Yamaha Gender and Age (YGA) database that contains 8,000 images. Guo et al. [10] demonstrated that age estimation on YGA has to be performed for males and females separately because of the influence of gender on age estimation. To deal with the gender influence problem, a two-step procedure was proposed [10] that uses a classification module first and then does age estimation for males and females separately. Ni et al. [17] performed cross-database age estimation using faces collected from the Internet as training data, and tested age estimation performance on some aging databases, e.g., the FG-NET [4] and MORPH [19]. However, the reported mean absolute errors (MAE) are high. For example, the MAE is 8.60 years on MORPH (a large database), and as high as 9.49 years on FG-NET (a smaller one). In order to reduce the age estimation errors, a study of the influence of ethnicity and gender on age estimation was performed very recently [9]. The empirical study shows that the influence is significant, and a two-step procedure was proposed: (1) gender and ethnicity group classification and (2) age estimation performed on each classified group. A question may be asked: Is it possible to use a *single* method to estimate age and deal with gender and ethnicity affection?

On the other hand, even a pure age estimation procedure may contain three tasks: feature extraction, feature dimensionality reduction, and aging function learning [14, 33, 11, 17, 10]. For example, [11, 10, 9] extracted features using a biologically-inspired method [20, 26, 16], performed dimensionality reduction by manifold learning techniques [2, 3, 34], and learned an aging function using the SVM [29]. One may ask: Is there a method to perform dimensionality reduction and aging function learning altogether, rather than performing two separate tasks using different techniques?

In this paper, we propose to investigate a new method, called kernel Partial Least Squares (kernel PLS or simply KPLS) regression, for age estimation. It can address the

*This work was partially supported by an NSF CITeR grant and an NIJ grant 2010-DD-BX-0161.

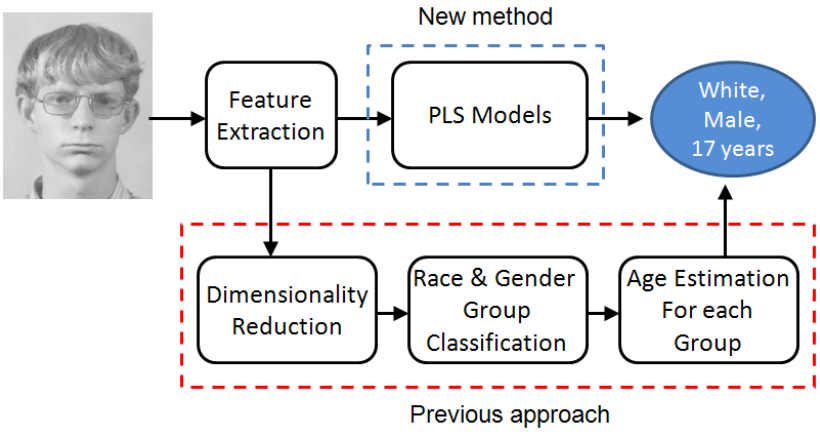

Figure 1. Procedures for age estimation under gender and ethnicity variations. After feature extraction, the new method proposed here can use a single step for age estimation, rather than taking three separate operations as in a previous approach [9].

two issues discussed above within a single learning framework. As illustrated in Figure 1, after feature extraction the newly proposed PLS models can use a single step for age estimation instead of taking three separate operations as in [9]. This single-step approach is beneficial in developing a practical age estimation system.

The PLS is a wide class of methods for modeling relations between sets of observed variables by means of latent variables. It was originally developed in chemometrics and has received a great deal of attention in other fields [22, 21], such as bioinformatics, medicine, pharmacology, and social science, to name a few. We are curious about the performance of the PLS methods for computer vision problems. Actually, Schwartz et al. recently applied the linear PLS method to deal with pedestrian detection [25] and face recognition [24]. However, the linear PLS was used only for reducing the dimensionality and is then followed by a classifier for two-class classification problems [25, 24]. Here our emphasis is that the PLS models can do *more things* than just dimensionality reduction. Further, we introduce the KPLS method, which is better than the linear PLS in terms of age estimation accuracy. However, to our best knowledge, the KPLS has not been explored before in any computer vision problem.

In the following, we will describe the linear and nonlinear PLS regression models first, discuss their relations to other learning techniques for age estimation, and then show the age estimation experiments. Finally, we draw conclusions.

2. Partial Least Squares

PLS models relationships between sets of observed variables by means of latent variables [32, 22, 21]. The underlying assumptions of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent variables. The projection of the observed data onto a small subspace of latent variables has been shown to be a powerful technique when observed variables are highly correlated, noisy and with high dimensionality. We describe the linear and nonlinear PLS regression models, and then compare them with other popular learning techniques.

2.1. Linear PLS Regression

Consider the general setting of a linear PLS algorithm [32, 22] to model the relation between two data sets or blocks of variables. Denote by $\mathcal{X} \subset \mathcal{R}^N$ an N-dimensional space of variables representing the first block and similarly by $\mathcal{Y} \subset \mathcal{R}^M$ a space representing the second block of variables. PLS models the relations between these two blocks by means of score vectors. After observing n data samples from each block of variables, PLS decomposes the $(n \times N)$ matrix of zero-mean variables $\mathbf{X}$ and the $(n \times M)$ matrix of zero-mean variables $\mathbf{Y}$ into the form

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F} \end{aligned} \tag{1}$$

where the $\mathbf{T}$ and $\mathbf{U}$ are $(n \times p)$ matrices of the p extracted score vectors (components, latent vectors), the $(N \times p)$ matrix $\mathbf{P}$ and the $(M \times p)$ matrix $\mathbf{Q}$ represent matrices of loadings, and the $(n \times N)$ matrix $\mathbf{E}$ and the $(n \times M)$ matrix $\mathbf{F}$ are the matrices of residuals. The PLS method, which in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm [31], finds weight vectors $\mathbf{w}, \mathbf{c}$ such that

$$\begin{aligned} [cov(\mathbf{t}, \mathbf{u})]^2 &= [cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 \\ &= max_{|r|=|s|=1} [cov(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]^2 \end{aligned} \tag{2}$$

where $cov(\mathbf{t}, \mathbf{u}) = \frac{\mathbf{t}^T \mathbf{u}}{n}$ denotes the sample covariance between the score vectors $\mathbf{t}$ and $\mathbf{u}$. The NIPALS algorithm

starts with random initialization of the $\mathcal{Y}$-space score vector $\mathbf{u}$ and repeats a sequence of iterations until convergence [31].

The linear PLS models can have variants based on the deflation difference [22]. Most PLS models assume that (1) the score vectors $\mathbf{t}_i$, $i = 1, \cdots, p$, are good predictors of $\mathbf{Y}$, and (2) a linear *inner relation* between the score vectors $\mathbf{t}$ and $\mathbf{u}$ exists; that is

$$\mathbf{U} = \mathbf{TD} + \mathbf{H} \tag{3}$$

where $\mathbf{D}$ is a $(p \times p)$ diagonal matrix and $\mathbf{H}$ is the matrix of residuals. Combining Eqns. (1) and (3), we can derive

$$\mathbf{Y} = \mathbf{TDQ}^T + (\mathbf{HQ}^T + \mathbf{F}) \tag{4}$$

and this defines the *linear PLS regression* model

$$\mathbf{Y} = \mathbf{TC}^T + \mathbf{F}^* \tag{5}$$

where $\mathbf{C}^T = \mathbf{DQ}^T$ denotes the $(p \times M)$ matrix of regression coefficients and $\mathbf{F}^* = \mathbf{HQ}^T + \mathbf{F}$ is the residual matrix.

From Eqn. (5), one can see that the linear PLS model can use only p latent variables contained in $\mathbf{T}$ for regression, and usually $p << N$ holds in practice. As a result, the linear PLS regression can be very fast in real applications.

The PLS models were originally derived for regression problems [31, 32, 22], but can be adapted to classification using a similar form as Eqn. (5). The difference is the $\mathbf{Y}$ matrix which is changed to encode the class membership. Some recent approaches attempt to apply the linear PLS for classification in computer vision applications [25, 24], but they can only deal with a two-class classification problem. A big effort has to be taken to deal with the multi-class classification problem in using the PLS method [24]. In addition, the linear PLS is only used for feature dimensionality reduction in [25, 24]. A classifier, e.g., SVM or others, has to follow the PLS for classification. Here we want to emphasize the nice property of the linear PLS model, i.e., the PLS model can do both *feature dimensionality reduction* and *age estimation* (or *regression*) simultaneously, which has not been presented in any previous approaches to human age estimation.

2.2. Kernel PLS Regression

When a strong nonlinear relation exists between two sets of data $\mathbf{X}$ and $\mathbf{Y}$, linear PLS models may not perform well. The nonlinear PLS models can be realized by using different extensions of the linear PLS models [21]. A mathematically elegant way is to use the kernel trick as in [23].

The linear PLS regression model (5) can also be expressed using the originally observed data $\mathbf{X}$ [23, 21] and written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}^* \tag{6}$$

where the matrix $\mathbf{B}$ has the following form

$$\mathbf{B} = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \tag{7}$$

Define the Gram matrix $\mathbf{K}$ of the cross dot products between all mapped input data points, i.e., $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$, where $\mathbf{\Phi}$ denotes the matrix of the mapped $\mathcal{X}$-space data $\{\Phi(\mathbf{x}_i) \in \mathcal{F}\}_{i=1}^n$, where $\mathcal{F}$ is the high-dimensional feature space. The kernel trick implies that the elements i, j of $\mathbf{K}$ are equal to the values of the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$.

Then the kernel variant of the linear PLS model (6) has the following form [21]

$$\mathbf{Y} = \mathbf{\Phi B} + \mathbf{F}^* \tag{8}$$

where the estimate of $\mathbf{B}$ is

$$\mathbf{B} = \mathbf{\Phi}^T\mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y} \tag{9}$$

Let $\mathbf{d}^m = \mathbf{U}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{y}^m$, $m = 1, \ldots, M$, where the $(n \times 1)$ vector $\mathbf{y}^m$ represents the m-th output variable. Then the *kernel PLS regression* estimate of the m-th output for a given input sample $\mathbf{x}$ will be

$$\hat{\mathbf{y}}^m = \Phi(\mathbf{x})^T\mathbf{\Phi}^T\mathbf{d}^m = \sum_{i=1}^{n} d_i^m k(\mathbf{x}, \mathbf{x}_i) \tag{10}$$

2.3. Relation to Other Learning Methods

Machine learning methods, such as support vector machines (SVM) and support vector regression (SVR) [29], have seen many successful applications in computer vision. The SVM and SVR have also been used successfully for human age estimation [8, 11], where age estimation can be viewed as either a multi-class classification or a regression problem. However, the SVM and SVR usually cannot do feature dimensionality reduction. While the dimension of the extracted features in representing aging faces could be as high as thousands, it is necessary to reduce the feature dimensionality before applying the SVM or SVR to age estimation. Dimensionality reduction has two impacts: (1) reduce the computational time, especially in real-time applications, and (2) more importantly, improve the estimation accuracy. Recent manifold learning techniques have the potentials to improve the classification performance by incorporating the label information in deriving a subspace for low-dimensional embedding. In [10], various manifold leaning methods were investigated for dimensionality reduction and age estimation accuracy improvement. The supervised manifold learning methods, such as OLPP (Orthogonal Locality Preserving Projections [2]), MFA (Marginal Fisher Analysis [34]), and LSDA (Locality Sensitive Discriminant Analysis [3]), can use the class label information (e.g., age label) to make the within-class examples closer, while different-class examples further away,

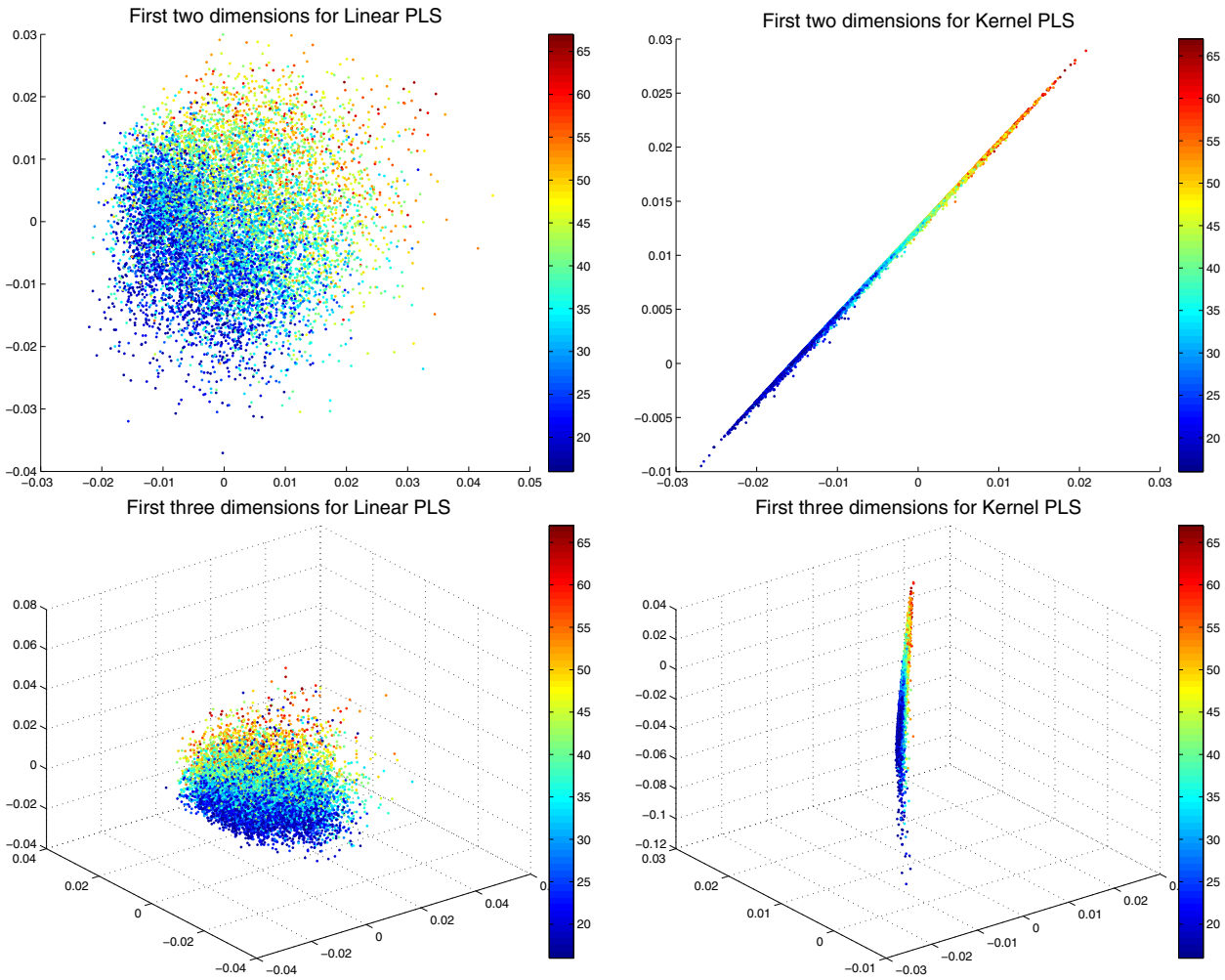

Figure 2. Visualization of the first two or three components learned by the linear and kernel PLS models. The age labels are colorized for a better view of the sample distributions after projection to the latent variables.

in computing the feature projection directions. These supervised manifold learning methods usually perform better than the unsupervised dimensionality reduction method such as the PCA [30]. It has been shown that the OLPP, LSDA, and MFA methods can improve the age estimation accuracies in comparison with the PCA (Principal Component Analysis) method [10]. However, the age estimation process has to take two steps [10, 11]: (1) dimensionality reduction by manifold learning, and (2) age estimation with SVM or SVR.

Here the PLS and kernel PLS methods can reduce the feature dimensions and do age estimation simultaneously, without using two separate steps.

Further, both the linear and nonlinear PLS models [21] can have a vector output $\mathbf{Y}$, which can contain multiple values in regression outputs. This nice property can help us do something more than age estimation. In [10], it showed that the age estimation is influenced significantly by gender. Age estimation should be performed for males and females separately, rather than in a mixed manner. To deal with the problem of gender influence in age estimation, the authors suggest a two-step process: gender classification and then age estimation. In [9], it shows that age estimation is influenced by both gender and ethnicity. To deal with this influence problem, a two-step procedure was proposed where age estimation was preceded by a gender and ethnicity group classification module.

Now, using the PLS or kernel PLS method, there is no need to have a classification module preceding age estimation. In a single step, the PLS or kernel PLS model can estimate age and take care of gender and ethnicity groups. It can be realized by simply putting the age, gender, and ethnicity labels into a vector $\mathbf{Y}$ for the PLS output.

In sum, the PLS models are the *right* method for human age estimation. It can perform dimensionality reduction, learn the aging function, and deal with gender and ethnic-

ity influences *simultaneously*. Because of this nice property, we believe that the linear and kernel PLS models are of great value for age estimation in practice. The only remaining question is the age estimation accuracy. We will evaluate both the linear and nonlinear PLS methods in our age estimation experiments, and compare with the state-of-the-art approaches.

3. Feature Extraction

Recently the biologically-inspired features (BIF) [20] have shown good performance in age estimation [11, 10], as well as object category recognition [27] [16] and face recognition [15]. A specially-designed BIF with two layers [11, 10, 9] shows much lower age estimation errors than previous approaches. In our study, we want to use advanced features for facial aging pattern representation, and thus adopted the BIF method for feature extraction. Here our focus is to study the performance of the linear and nonlinear PLS models in age estimation.

In previous use of the BIF representation for age estimation [11, 10, 9], the BIF has to be combined with manifold learning or subspace analysis in order to reduce the dimensionality and improve the estimation accuracy over the original high-dimensional BIF. The manifold learning methods that have been used before for the BIF representation include the PCA, OLPP [2], MFA [34], and LSDA [3]. After dimensionality reduction, the SVM method is usually used to estimate age [11, 10, 9]. Here we are interested in the linear and kernel PLS models since these methods can perform the two tasks simultaneously, given a feature representation. This is for the first time in computer vision to use *a single method* to reduce the feature dimensions and estimate age.

4. Experiments

We investigate the performance of PLS models for age estimation on a large database called MORPH [19]. A subset of the database is selected considering the gender and ethnicity distributions. The subset is divided into two halves that are used for training alternately, while the remaining in the database is used for testing. Note that only two ethnic groups (White and Black) were used for training, since the number of examples is too small in other groups. We evaluate the PLS models and compare with the state-of-the-art approaches.

4.1. The Database

The MORPH database [19] was used for our study. It is a large database containing two sections, I and II. Since MORPH-I is too small, we used MORPH-II that contains about 55,000 face images. The MORPH is a multi-ethnic database. It has about 77% Black faces and 19% White, while the remaining 4% includes Hispanic, Asian, Indian, and Other. The MORPH-II database is summarized in Table 1 based on the gender and ethnicity categories. One can see that the MORPH database is also very unbalanced in terms of the gender and ethnicity distributions, which was noticed in a recent study [9]. Further, it was showed in [9] that age estimation is influenced significantly by gender and ethnicity when evaluated in a cross-gender or cross-ethnicity situation.

Table 1. The MORPH-II database. All race and gender labels as well as age are provided with the database.

Race	Female	Male	Female and Male
Black	5,757	36,803	42,560
White	2,601	7,999	10,600
Hispanic	100	1,651	1,751
Asia	13	146	159
India	14	43	57
Other	2	3	5
Total	8,487	46,645	55,132

Because of the unbalanced distributions of ethnic groups, we randomly selected a subset of about 21,000 faces from MORPH-II that contains Black and White, Female and Male faces, denoted as set $\mathcal{S}$. The age range is from 16 to 67 years in set $\mathcal{S}$, almost the same as in $\mathcal{W}$, the whole database. So $\mathcal{S} \subset \mathcal{W}$. In set $\mathcal{S}$, there are half White and half Black, while the gender distribution is $F : M \approx 1 : 3$, where F is for Female and M is for Male. The selection is to use all available females and make the subset $\mathcal{S}$ as large as possible. In order to compare with the method and results in [9], we selected $\mathcal{S}$ similarly. Then the subset $\mathcal{S}$ is divided into $\mathcal{S}_1$ and $\mathcal{S}_2$ equally, so each $\mathcal{S}_i$ is half of $\mathcal{S}$. The purpose is to be able to alternate between $\mathcal{S}_1$ and $\mathcal{S}_2$ for training, and then the remaining in the whole database, i.e., $\mathcal{W} \setminus \mathcal{S}_1$ or $\mathcal{W} \setminus \mathcal{S}_2$ is used for testing.

The face images in set $\mathcal{W}$ were preprocessed. The faces were detected and aligned, and also cropped and resized to 60×60, similar to many previous approaches. Only the gray level images were used and the BIF features [11] were extracted from each face patch.

Now our focus is to evaluate the performance of linear and kernel PLS models for dimensionality reduction and age estimation. We also want to investigate the nice property of the PLS models that can use a vector as the regression output, so that the gender and ethnicity can be estimated together with age information. The Yamaha database used in [35, 11] and the very small FG-NET [4] database are not good for us to fully explore the advantages of the PLS models.

4.2. Visualization of the Regression Results

Before performing the age estimation experiments, let us first look at the regression results of linear and nonlinear partial least squares (PLS) models. Given a training set $\mathcal{S}_1$ containing about 10,000 face images, the BIF features were extracted on each face. Then the linear and kernel PLS methods are applied to the training set $\mathcal{S}_1$ with the labels of age, gender, and ethnicity. As introduced in the beginning of this section, the age range is from 16 to 67 years. The ethnicity has only White and Black in the selected subset, either $\mathcal{S}_1$ or $\mathcal{S}_2$. Experimentally we found that the linear PLS can work well with 30 hidden variables, while the nonlinear PLS work well with 20 hidden variables. Having more hidden variables does not necessarily improve the performance. To show the distribution of the regression results, the first two or three hidden variables are used for projection and the samples are displayed in Figure 2. One can see that the points projected by the linear PLS model can display the age trend although they are scattered. The points projected by the kernel PLS display a tighter and clearer distribution of the aging trend. This visualization shows intuitively that the kernel PLS model is probably better than the linear PLS in learning the aging regression function.

4.3. Results and Comparisons

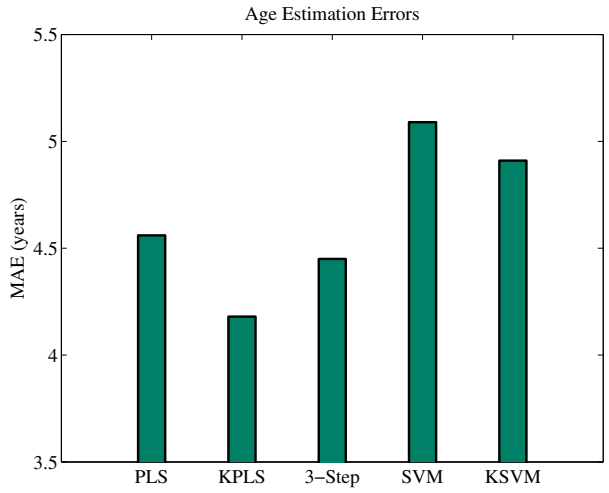

Figure 3. The age estimation errors using different methods. KPLS for Kernel PLS, KSVM for Kernel SVM, and 3-Step for the approach [9] using three steps – dimensionality reduction using OLPP, gender and ethnicity group classification, and then age estimation. The Kernel PLS model has the smallest error in age estimation.

Now we perform age estimation using the linear and nonlinear PLS models. We also compare the new PLS models with the state-of-the-art methods on age estimation. The experimental results are shown in Table 2. We found that both the linear and kernel PLS models can learn a very small number of "latent variables," e.g., 30 for linear PLS and 20 for kernel PLS, respectively. The number of latent variables determines the reduced dimensionality for the input features. The specific numbers of the latent variables were observed from a range of different numbers in experiments. The numbers, 20 and 30, are much smaller than the dimension of 200 selected by the OLPP method in [9], and much smaller than the original dimension of 4,376 based on the BIF representation [11].

The linear and kernel PLS models can have a vector as the output. So it is convenient to put age, gender, and ethnicity labels altogether into the output vector. Then the PLS models can estimate age, gender, and ethnicity using the single learning step. As shown in columns 5 and 6 in Table 2, the accuracies of gender and ethnicity estimation are pretty high for both PLS models, and comparable to the complex three-step process proposed in [9]. Please note that for ethnicity estimation we only reported accuracies for the Black and White, since other race groups were not used in training because the number of samples is too small. The linear SVM method used in [11] (without dimensionality reduction) only deals with age estimation, without considering gender or ethnicity. The linear or kernel SVM cannot do age, gender and ethnicity *simultaneously*.

Let us look at the age estimation results shown in the last column in Table 2, which is the average of column 7 when two different sets, $\mathcal{S}_1$ and $\mathcal{S}_2$, were used for training alternately. The kernel PLS obtains an MAE of 4.18 years, which is smaller than the 4.56 by linear PLS, and is smaller than the 4.45 using the complex 3-step procedure in [9]. The error reduction rates are 8.3% and 6.1%, respectively. Considering the age estimation is performed on a very large database, these error reductions are statistically significant. The linear SVM has an MAE of 5.09 years when working on the original BIF features without dimensionality reduction. The kernel SVM gives an MAE of 4.91 which is even higher than the linear PLS model. Comparing the kernel PLS with the linear and kernel SVMs, the error reduction rates are 17.9% and 14.9%, respectively. These two error reductions are very significant. As a result, the kernel PLS model outperforms the linear PLS, the state-of-the-art approaches using a three-step procedure [9] and using the linear SVM in [11]. To visually show the errors based on different methods, we display the MAEs in Figure 3.

4.4. Aging Function Learning

We have shown that the kernel PLS model has the smallest age estimation errors in comparison with the state-of-the-art methods as well as the nice properties to do dimensionality reduction and gender and ethnicity estimation simultaneously. Now we are curious about the capability of the kernel PLS model *just* for the aging function learning. In order to explore this, we applied the linear PLS model

Table 2. The performance of various approaches to age estimation on the whole database of MORPH-II, denoted by $\mathcal{W}$.

Method	Training Set	Test Set	Dimension	Gender Classification Accuracy	Race Classification Accuracy	Age Estimation Error (yrs.)	Average MAE (yrs.)
PLS	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	30	97.35%	98.7%	4.58	4.56
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	30	97.33%	98.6%	4.54	
KPLS	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	20	98.20%	98.9%	4.21	**4.18**
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	20	98.20%	98.8%	4.15	
3-Step [9]	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	200	98.09%	98.9%	4.44	4.45
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	200	97.94%	98.8%	4.46	
Linear SVM [11]	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4,376	–	–	5.06	5.09
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4,376	–	–	5.12	
Kernel SVM	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4,376	–	–	4.89	4.91
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4,376	–	–	4.92	

Table 3. Evaluation of different methods for aging function learning, given the input data with reduced dimensionality using the linear PLS (30 latent variables) on BIF features.

Method	Training Set	Test Set	Age Estimation Error (yrs.)	Average MAE (yrs.)
KPLS	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4.43	**4.43**
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4.42	
Linear SVM	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4.54	4.55
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4.56	
Kernel SVM	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4.89	4.87
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4.85	
Linear SVR	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4.57	4.55
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4.53	
Kernel SVR	$\mathcal{S}_1$	$\mathcal{W} \setminus \mathcal{S}_1$	4.88	4.88
	$\mathcal{S}_2$	$\mathcal{W} \setminus \mathcal{S}_2$	4.87	

for dimensionality reduction using 30 latent variables, and then used the kernel PLS model for aging function learning. To compare the performance with other advanced learning techniques, we also used the linear SVM, kernel SVM, linear SVR, and kernel SVR (the kernel SVR [29] has been used for aging function learning before in [8] on different databases, therefore we selected the SVR for comparison, too). All kernels here use the RBF function [29]. The age estimation results are shown in Table 3. One can observe that the SVM and SVR have very similar age estimation errors, while the linear SVM (or SVR) performs even better than the kernel SVM (or SVR). The kernel PLS model gives the smallest error among the five methods, given the same input. This demonstrates that the kernel PLS is a great method for aging function learning, not just for dimensionality reduction.

Another reason to perform this experiment is that the linear PLS is very fast in learning (and also in testing). It can quickly reduce the original high dimensionality into a small number, e.g., 30. Then various methods for aging function leaning can be evaluated and compared conveniently on a very large database.

5. Conclusions

We have investigated the elegant methods called partial least squares models for age estimation. Both the linear and nonlinear PLS models can reduce the dimensionality from thousands into a very small number, e.g., 30 or 20. The PLS models are useful not only for dimensionality reduction, but also to learn the aging function as well. They can even deal with age, gender, and ethnicity altogether within a single learning step. All these nice properties make them great methods for human age estimation. When performing experiments on a very large database of more than 50,000 face images, the kernel PLS model gives the smallest error of 4.18 years, which outperforms the linear PLS model and the state-of-the-art methods. We expect to see more applica-

tions of the PLS models in other computer vision problems.

References

[1] *Electronic Customer Relationship Management (ECRM)*. http://en.wikipedia.org/wiki/ECRM.

[2] D. Cai, X. He, J. Han, and H. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE Trans. on Image Processing*, 15:3608–3614, 2006.

[3] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *Proc. Int. Joint Conf. on Artificial Intell.*, 2007.

[4] FGNET. The fg-net aging database. In *http://www.fgnet.rsunit.com/*, 2002.

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

[6] Y. Fu, Y. Xu, and T. S. Huang. Estimating human ages by manifold analysis of face pictures and regression on aging features. In *IEEE Conf. on Multimedia and Expo*, pages 1383–1386, 2007.

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. on PAMI*, 29(12):2234–2240, 2007.

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

[9] G.-D. Guo and G. Mu. Human age estimation: what is the influence across race and gender? In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2010.

[10] G.-D. Guo, G. Mu, Y. Fu, C. Dyer, and T. S. Huang. A study on automatic age estimation on a large database. In *IEEE International Conference on Computer Vision*, pages 1986–1991, 2009.

[11] G.-D. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009.

[12] J. Hayashi, M. Yasumoto, H. Ito, and H. Koshimizu. A method for estimating and modeling age and gender using facial image processing. In *Seventh Int. Conf. on Virtual Systems and Multimedia*, pages 439–448, 2001.

[13] Y. Kwon and N. Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[14] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. on SMC-B*, 24(4):621–628, 2002.

[15] E. Meyers and L. Wolf. Using biologically inspired features for face processing. *Int. J. Comput. Vis.*, 76:93–104, 2008.

[16] J. Mutch and D. Lowe. Object class recognition and localization using sparse features with limited receptive fields. In *IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 11–18, 2006.

[17] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *ACM Multimedia*, 2009.

[18] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Trans. on Image Processing*, 15(11):3349–3361, 2006.

[19] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *IEEE Conf. on AFGR*, pages 341–345, 2006.

[20] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[21] R. Rosipal. Nonlinear partial least squares: An overview. In H. Lodhi and Y. Yamanishi, editors, *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, pages 169–189. ACCM, IGI Global, 2011.

[22] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*, pages 34–51. Springer, 2006.

[23] R. Rosipal and L. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *J. of Machine Learning Research*, 2:97–123, 2001.

[24] W. Schwartz, H. Guo, and L. Davis. A robust and scalable approach to face identification. In *ECCV*, 2010.

[25] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *ICCV*, 2009.

[26] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007.

[27] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2005.

[28] K. Ueki, T. Hayashida, and T. Kobayashi. Subspace-based age-group classification using facial images under various lighting conditions. In *IEEE conf. on AFGR*, 2006.

[29] V. N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[30] A. R. Webb. *Statistical Pattern Recognition, 2nd Edition*. John Wiley, 2002.

[31] H. Wold. Path models with latent variables: The nipals approach. In H. M. Blalock and et al, editors, *Quantitative Sociology: Internnational perspectives on mathematical and statistical model building*, pages 307–357. Academic Press, 1975.

[32] H. Wold. Partial least squares. In S. Kotz and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. Wiley, New York, 1985.

[33] S. Yan, M. Liu, and T. Huang. Extracting age information from local spatially flexible patches. In *IEEE conf. on ICASSP*, pages 737–740, 2008.

[34] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:40–51, 2007.

[35] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang. Regression from patch-kernel. In *IEEE conf. on CVPR*, 2008.