Joint Estimation of Age, Gender and Ethnicity: CCA vs. PLS

Guodong Guo*
LCSEE, West Virginia University

Guowang Mu
School of Science, Hebei University of Technology

***Abstract*— Human age, gender and ethnicity are valuable demographic information about a population. These measures are also considered important soft biometric traits for human recognition or search. Usually the three traits are studied separately. A recent study [9] shows that the three traits can be estimated simultaneously based on a multi-label regression formulation. The linear and kernel partial least squares (PLS) models are adopted to solve the multi-label regression problem in [9]. In this study, we investigate the canonical correlation analysis (CCA) based methods, including linear CCA, regularized CCA (rCCA), and kernel CCA (KCCA), and compare to the PLS models in solving the joint estimation problem. Interestingly, we found a consistent ranking of the five methods in estimating age, gender, and ethnicity. More importantly, we found that the CCA based methods can derive an extremely low dimensionality in estimating age, gender and ethnicity, which has not been shown in previous research, to the best of our knowledge. The experiments are conducted on a very large database of more than 55,000 face images.**

I. Introduction

Human age, gender and ethnicity are valuable demographic information about a population. Automated estimation of the demographics is of great value in practice, such as business intelligence, local community planning, and new school locating. Age, gender and ethnicity are also useful soft biometric traits to help human identification or verification. However, current computational techniques are still not robust enough for practical uses. Thus it is demanding to develop a robust and effective system to recognize age, gender and ethnicity for a given population.

In the literature, there are different methods for the estimation of each single trait, e.g., age estimation [19], [5], [7], [4], [25], [11], [3], gender classification [6] [2] [23] [32] [1], or ethnicity estimation [13] [16] [21] [12]. However, there are very few approaches to estimate all three traits together, excepting [37] [9]. In [37], a classification of ethnicity, gender, and age groups was executed, with each trait classified independently. In [9], a multi-label regression formulation is proposed to estimate age, gender, and ethnicity simultaneously. The partial least squares (PLS) models, both linear and kernel, are used to solve the multi-label regression problem. It is reported in [9] that a small number of feature dimensions, e.g, 20 or 30, reduced from thousands (in an original feature space), can achieve a good performance to estimate all three traits.

In our study, we use the same multi-label regression formulation as in [9], but investigate the CCA based methods for joint estimation of age, gender, and ethnicity. We compare the CCA based methods with the PLS based, and derive a ranking of five methods under consideration to deal with the joint estimation problem. More interestingly, we want to explore if a minimum number of dimensions can be obtained in estimating the three traits on a large database.

Specifically, we found that the canonical correlation analysis (CCA) based methods, including linear CCA, regularized CCA, and kernel CCA, can find only three basis vectors to project the original features of several thousand dimensions. Thus only three dimensions of features are needed (after transformation) to estimate all three traits. This is a *novel* finding in estimating age, gender and ethnicity. Further, we use the rank theory to analyze the feature dimensionality problem in using the CCA based methods for our specific task. Hopefully our analysis may inspire more investigations for other recognition problems to derive similar results with a minimum number of feature dimensions.

Based on the multi-label regression formulation, the framework is very simple even for estimating all three traits, as shown in Figure 1. Thousands of features are extracted from face images using the biologically-inspired features proposed in [11] for age estimation, and then the CCA based methods are used to project the feature into a very low dimensional space, i.e., three dimensions, and age, gender and ethnicity estimation are performed simultaneously.

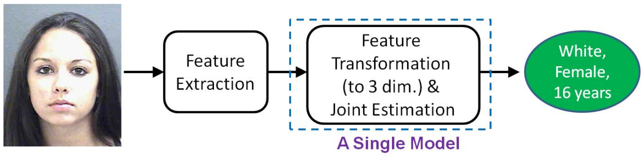

Fig. 1. Illustration of the framework in our approach. After feature extraction, the new model uses one method but accomplishes several things: (1) feature projection from thousands into only 3 dimensions, and (2) estimate of age, gender, and ethnicity altogether.

Our major contributions in this paper include: 1) A novel finding on feature dimensionality in estimating age, gender and ethnicity; 2) A rank theory based analysis of the dimensionality problem in using the CCA based methods; 3) A ranking of the CCA and PLS based methods in joint estimation of age, gender, and ethnicity.

In the remaining of the paper, we describe the extracted features briefly, and then introduce the CCA based methods and do dimensionality analysis using the rank theory in Section III. The PLS and kernel PLS methods are described in Section IV. The experimental evaluations are presented in Section V, and finally we draw conclusions.

*This work is partially supported by an NIJ grant 2010-DD-BX-0161.

II. Feature Extraction

Recently the biologically-inspired features (BIF) [27] have shown good performance in age estimation [11], [10], as well as object category recognition [31] [24] and face recognition [22]. A specially-designed BIF with two layers [11], [10], [8], the simple layer $S1$ and complex layer $C1$, shows much lower age estimation errors than previous approaches. The $S1$ layer contains a set of Gabor filters with parameters designed based on the visual cortex models [31], and the $C1$ layer contains some non-linear operations including the "MAX" pooling and an "STD" operation [11], in order to have some invariance to translation, rotation, and scaling, as well as a characterization of the aging details.

In our study, we want to use advanced features for facial pattern representation, and thus adopted the BIF method for feature extraction. Given the BIF features, our focus is to investigate the CCA based methods for the joint estimation problem. Further, we study the minimum number of features needed, and which methods can derive a minimum number of features.

III. Canonical Correlation Analysis

Canonical correlation analysis (CCA) is introduced by Hotelling [17] to describe the linear relation between two multidimensional variables as the problem of finding basis vectors for each set such that the projections of the two variables on their respective basis vectors are maximally correlated [17], [15].

The CCA methods have been applied to solve some computer vision problems, e.g., image annotation [14], action classification [18], and face recognition [20], [33]. But the CCA methods have not been exploited before for age estimation or recognizing all three traits (age, gender and ethnicity), to the best of our knowledge. More importantly, there is few work that studies the dimensionality issue in CCA based methods in solving computer vision or recognition problems.

A. Linear CCA

Let p-dimensional x and q-dimensional y denote the two sets of real-valued zero-mean random variables (i.e., $x \in R^p$ and $y \in R^q$). Let $p \times N$ matrix X be the data matrix of the first set, and $q \times N$ matrix Y be the data matrix of the second set. The canonical correlation analysis (CCA) computes two projection vectors, $w_x \in R^p$ and $w_y \in R^q$, such that the correlation coefficient

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}} \qquad (1)$$

is maximized [17], [15]. Since ρ is invariant to the scaling of w_x and w_y, CCA can be formulated equivalently as

$$\max_{w_x, w_y} w_x^T X Y^T w_y, \qquad (2)$$

subject to $w_x^T X X^T w_x = 1$, and $w_y^T Y Y^T w_y = 1$.

It can be shown [15] that w_x can be obtained by solving the following generalized eigenvalue problem,

$$X Y^T (Y Y^T)^{-1} Y X^T w_x = \lambda X X^T w_x, \qquad (3)$$

where λ is the eigenvalue corresponds to the eigenvector w_x. It has also been shown [15] that multiple projection vectors under certain orthonormality constraints consist of the top l eigenvectors of the generalized eigenvalue problem in (3). Thus the feature dimension of data X can be reduced to a lower value.

In our study, X represents the data matrix, and Y represents the label space. After the dimension of data X is reduced, we use a least square fitting to build the relation between the dimension reduced feature and label Y. Then the prediction of Y for the test data is based on the least square fitting result. This simple least square fitting method can work well, and is also applied to other CCA extensions, which will be introduced in this section later.

1) Dimensionality Analysis: Assuming XX^T is nonsingular, eqn. (3) can be written as

$$(XX^T)^{-1} X Y^T (Y Y^T)^{-1} Y X^T w_x = \lambda w_x, \qquad (4)$$

or

$$M w_x = \lambda w_x, \qquad (5)$$

with $M = (XX^T)^{-1} X Y^T (Y Y^T)^{-1} Y X^T$.

Now we analyze the rank of matrix M based on the linear algebra theory. Suppose the covariance matrices or XX^T and YY^T are positive definite. Then we have

$$\begin{aligned} rank(M) &= rank((XX^T)^{-1} X Y^T (Y Y^T)^{-1} Y X^T) \\ &\leq min\{rank(X), rank(Y)\} \\ &\leq min\{r(X), c(X), r(Y), c(Y)\} \\ &= min(p, q, N), \end{aligned} \qquad (6)$$

where $r(\cdot)$ and $c(\cdot)$ denote the number of rows and columns of a matrix, respectively. We have $r(X) = p$, $c(X) = N$, $r(Y) = q$, $c(Y) = N$.

Based on the rank analysis in (6), we know that the rank of the matrix M is at most equal to the minimum value among the three numbers, $p, q,$ and N. Thus the number of non-zero eigenvalues of matrix M is less than or equal to $min\{p, q, N\}$. As a result, the eigenvalue problem in (4) has at most $min\{p, q, N\}$ eigenvectors, corresponding to non-zero eigenvalues.

In our case of estimating age, gender and ethnicity, we have $p = 4,376$, $q = 3$, and $N = 10,530$, based on our experimental setup in this study. It is obvious that $min\{p, q, N\} = 3$. So the maximum number of non-zero eigenvalues of (4) is three. Based on the assumption that XX^T is nonsingular, the generalized eigenvalue problem in (3) is equivalent to the problem in (4). Thus we can say that the generalized eigenvalue problem in (3) has at most three non-zero eigenvalues in our study, under the assumption that XX^T is nonsingular.

Thus we only need at most three eigenvectors corresponding to the three non-zero eigenvalues to project the data of X, even if it is a large scale problem with the number of training examples $N = 10,530$. The number of testing examples is even larger with more than 40,000. Our experiments (see Section V) show that it is true to use only three feature dimensions (or three projections of the data X) for estimation.

On the other hand, we found that we do need to use three feature dimensions to estimate the age, gender and ethnicity altogether in our experiments (see Section V). This indirectly indicates that the three traits are different, which means one trait cannot include another. So the rank of matrix Y is three, rather than less than three.

Another interesting thing that we observed in our experiments (see Section V) is that when more than three features are used for estimation, the estimation error or accuracy will almost keep the same (straight lines in Fig. 2). This kind of phenomenon has not been observed often in many computer vision problems. Further, the performance of the PLS methods has a very different behavior with respect to the feature dimension (see Fig. 2). Our work may inspire further exploration of the CCA based methods and comparison with the PLS for other computer vision problems.

We also found that the regularized CCA (or rCCA) performs much better than the standard CCA. A brief description of rCCA will be presented below.

B. Regularized CCA

In regularized CCA or rCCA, a regularization term is added to each data set to stabilize the solution [34]. The corresponding generalized eigenvalue problem is given by

$$\begin{aligned} &XY^T[(1-\gamma_y)YY^T+\gamma_y I]^{-1}YX^Tw_x \\ &= \lambda[(1-\gamma_x)XX^T+\gamma_x I]w_x \end{aligned} \tag{7}$$

It has been pointed out [29] that (1) when $\gamma_x = 0, \gamma_y = 0$ (7) is to solve the standard CCA; (2) when $\gamma_x = 1, \gamma_y = 1$ (7) is to solve the PLS eigenvalue problem; (3) by continuously changing of γ_x, γ_y a regularized CCA is solved. In our study, we set $\gamma_x = \gamma_y = 0.09$, and found that the rCCA is significantly better than the standard linear CCA.

Our rank analysis in Section III-A.1 can also be applied to the regularized CCA. We will not repeat it here. Experimentally, we found that only three feature dimensions are needed for rCCA in estimating the age, gender and ethnicity, similar to CCA, but with a lower error.

C. Kernel CCA

Kernel CCA, or KCCA, is to apply the kernel trick to the CCA [15]. In KCCA, the directions w_x and w_y can be rewritten as the projection of the data onto the direction α and β, as

$$w_x = X\alpha, \qquad w_y = Y\beta. \tag{8}$$

Let $K_x = X^TX, K_y = Y^TY$ be the kernel matrices corresponding to the two representation. Then the canonical correlation can be written as

$$\rho = \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \cdot \beta^T K_y^2 \beta}}, \tag{9}$$

which is to be maximized. To force non-trivial learning on the correlation by controlling the problem of overfitting, a regularization is applied to KCCA. Then the generalized eigenvalue problem becomes

$$(K_x + \kappa I)^{-1} K_y (K_y + \kappa I)^{-1} K_x \alpha = \lambda \alpha. \tag{10}$$

In our experiments, we set $\kappa = 0.05$. If κ is set to zero, it becomes the standard KCCA without regularization. In our study, this regularization is helpful, although the accuracy improvement is not big. We used the Gaussian kernel for X and linear kernel for Y, based on the property of our problem. Experimentally, we found the same thing as CCA or rCCA, that is, only three feature dimensions are needed for KCCA in estimating age, gender and ethnicity (see Fig. 2).

In the next section, we briefly introduce the PLS and KPLS.

IV. Partial Least Squares

The partial least squares (PLS) algorithm [36], [29] can also model the relation between two data sets. It uses latent variables to learn a new space to make the data correlate to each other. Very recently, the linear and kernel PLS methods have been adapted to estimate age, gender and ethnicity [9]. For comparisons with the CCA, we briefly describe the PLS methods below.

Given two blocks of zero-mean variables, $\mathbf{X}$ and $\mathbf{Y}$, the PLS decomposes them into the form

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F} \end{aligned} \tag{11}$$

where the $\mathbf{T}$ and $\mathbf{U}$ are matrices of the extracted latent vectors, the matrices $\mathbf{P}$ and $\mathbf{Q}$ represent loadings, and the matrices $\mathbf{E}$ and $\mathbf{F}$ are residuals. The PLS method, which in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm [35], finds weight vectors $\mathbf{w}, \mathbf{c}$ such that

$$\begin{aligned} [cov(\mathbf{t},\mathbf{u})]^2 &= [cov(\mathbf{Xw},\mathbf{Yc})]^2 \\ &= max_{|r|=|s|=1}[cov(\mathbf{Xr},\mathbf{Ys})]^2 \end{aligned} \tag{12}$$

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 [35].

The PLS models can have variants based on the deflation difference [29]. Most PLS models assume that (1) the score vectors $\mathbf{t}_i$, $i = 1, \cdots, m$, are good predictors of $\mathbf{Y}$, and (2) a linear *inner relation* between the score vectors $\mathbf{t}$ and $\mathbf{u}$ exists; that is

$$\mathbf{U} = \mathbf{TD} + \mathbf{H} \tag{13}$$

where $\mathbf{D}$ is a $(m \times m)$ diagonal matrix and $\mathbf{H}$ is the matrix of residuals. Combining Eqns. (11) and (13), we can derive

$$\mathbf{Y} = \mathbf{TDQ}^T + (\mathbf{HQ}^T + \mathbf{F}) \tag{14}$$

and this defines the linear PLS model

$$\mathbf{Y} = \mathbf{TC}^T + \mathbf{F}^* \tag{15}$$

where $\mathbf{C}^T = \mathbf{D}\mathbf{Q}^T$ denotes the matrix of regression coefficients and $\mathbf{F}^* = \mathbf{H}\mathbf{Q}^T + \mathbf{F}$ is the residual matrix.

When a strong nonlinear relation exists between two sets of data $\mathbf{X}$ and $\mathbf{Y}$, the kernel trick can be used to derive the kernel PLS [30].

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 has the following form [28]

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{B} + \mathbf{F}^* \tag{16}$$

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

$$\mathbf{B} = \mathbf{\Phi}^T\mathbf{U}(\mathbf{T}^T\mathbf{K}\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \tag{17}$$

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

$$\hat{\mathbf{y}}^m = \Phi(\mathbf{x})^T\mathbf{\Phi}^T\mathbf{d}^m = \sum_{i=1}^{n} d_i^m k(\mathbf{x}, \mathbf{x}_i) \tag{18}$$

V. Experiments

To empirically investigate the issue of how many features are needed to estimate age, gender and ethnicity, and evaluate the estimation accuracy, we perform experiments on a very large database called MORPH [26].

We introduce the MORPH database with some details and our experimental setup. Then we present the results based on the CCA and its extensions, and compare with several other methods. The running time of the CCA and PLS based methods is also presented.

Probably MORPH [26] is the only large database that contains age, gender, and ethnicity. Since MORPH-I is too small, we used MORPH-II for our study, which contains about 55,000 face images. Following previous studies on MORPH II [8], [9], we divide the whole MORPH database $\mathcal{W}$ into three sets, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. The MORPH database has unbalanced gender and ethnicity distributions, which was also noticed in a recent study [8].

The face images in set $\mathcal{W}$ were preprocessed. The faces were detected and aligned, and also cropped and resized to 60×60, as suggested in [9]. Only the gray level images were used and the BIF features [11] were extracted.

The experimental results are shown in Table I and Fig. 2. We used the linear CCA, regularized CCA, and kernel CCA to estimate age, gender and ethnicity on the MORPH database. We also compare with the linear and nonlinear PLS models, as well as with some other approaches to age estimation on MORPH.

We found that all the CCA based methods only need three features to estimate all three traits simultaneously, as shown in Fig. 2. Interestingly, when more than three features are used, the accuracies or errors will stay almost the same, i.e., straight lines in the figures, without changing accuracies. If a less number of features are used, the performance drops steeply. This sudden change and then keeping steady is different from gradual changes in many feature selection or dimensionality reduction approaches. The behavior of CCA based methods is quite interesting, and it is not often to observe this kind of phenomenon in other computer vision problems. From Fig. 2, one can also see that when more eigenvectors are used for feature projection, corresponding to zero eigenvalues, there is no improvement for the task. Those feature dimensions should be discarded. As a result, only three feature dimensions are needed.

From Fig. 2, one can also observe that the linear and kernel PLS models have gradually changed accuracies or errors with respective to the feature dimensions, and need to use 30 and 35 “latent variables,” respectively, in order to get the highest accuracies or lowest errors. The accuracies of the PLS model may drop when more feature dimensions are used. The number of feature dimensions is determined by the latent variables in PLS and KPLS methods. The number of dimensions, 3, for the CCA based methods, is much smaller than the 30 or 35, based on the linear and kernel PLS methods, and is much smaller than the dimension of 200 selected by the OLPP method in [8], also is much smaller than the original dimension of 4,376 based on the BIF representation [11].

The CCA and 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 CCA and PLS models can estimate age, gender, and ethnicity using the single learning step. As shown in columns 5 and 6 in Table I, the accuracies of gender and ethnicity estimation are pretty high for both CCA and PLS models, and comparable to the complex three-step process (dimensionality reduction, race and gender group classification, and age estimation within each group) proposed in [8]. 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 I, 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 CCA obtains an MAE of 3.98 years, which is for the first time to have an MAE below 4 years on the MORPH database. This low MAE is even smaller than the kernel PLS which as an MAE of 4.04 years, and is smaller than any linear models, such as 5.37 years for CCA, 4.42 years for rCCA, and 4.56 years for PLS. It is also smaller than the 4.45 using a complex 3-step procedure in [8]. Compared with the MAE of 4.45 years in [8], the error reduction rate for KCCA is 10.6%. Considering the age estimation is performed on a very large database, this error reduction rate is 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 rCCA model. Comparing the kernel CCA with the linear and kernel SVMs, the error reduction rates are 21.8% and 18.9%, respectively. These two error reductions are very significant.

TABLE I

ESTIMATING AGE, GENDER AND ETHNICITY: THE PERFORMANCE OF VARIOUS METHODS ON THE WHOLE DATABASE OF MORPH-II, DENOTED BY $\mathcal{W}$. THE MAE IS IN YEARS. THE TEST SET IS $\mathcal{W} \setminus \mathcal{S}_1$, CORRESPONDING TO THE TRAINING SET $\mathcal{S}_1$.

Method	Tr. Set	Dim.	Gender Classif. Accu.	Race Classif. Accu.	Age Est. Err.	Avg. MAE
CCA	$\mathcal{S}_1$	3	95.2%	97.8%	5.39	5.37
	$\mathcal{S}_2$	3	95.2%	97.8%	5.35	
rCCA	$\mathcal{S}_1$	3	97.6%	98.7%	4.43	4.42
	$\mathcal{S}_2$	3	97.6%	98.6%	4.40	
KCCA	$\mathcal{S}_1$	3	98.5%	98.9%	4.00	**3.98**
	$\mathcal{S}_2$	3	98.4%	99.0%	3.95	
PLS[9]	$\mathcal{S}_1$	30	97.4%	98.7%	4.58	4.56
	$\mathcal{S}_2$	30	97.3%	98.6%	4.54	
KPLS[9]	$\mathcal{S}_1$	35	98.4%	99.0%	4.07	4.04
	$\mathcal{S}_2$	35	98.3%	99.0%	4.01	
3-Step[8]	$\mathcal{S}_1$	200	98.1%	98.9%	4.44	4.45
	$\mathcal{S}_2$	200	97.9%	98.8%	4.46	
LSVM[11]	$\mathcal{S}_1$	4,376	–	–	5.06	5.09
	$\mathcal{S}_2$	4,376	–	–	5.12	
KSVM	$\mathcal{S}_1$	4,376	–	–	4.89	4.91
	$\mathcal{S}_2$	4,376	–	–	4.92	

In addition to the age estimation errors and gender and ethnicity classification accuracies, we would like to examine the running time for the CCA and PLS based methods. The real running time for both training and testing was recorded and is shown in Table II. Note that the time for feature extraction by the BIF is not shown. One can see that the linear methods (CCA, rCCA, and PLS) are much faster than the kernel extensions (KCCA, KPLS) in both training and testing stages. The reason why the the KCCA and KPLS are so slow is that the kernel matrix is too big, i.e., 10,530×10,530 in training, while 44,602×10,530 in testing. It is determined by the number of training and testing examples, respectively.

So, considering both the accuracies and running time, we recommend to use the rCCA method for practical applications. It is very fast, and its accuracy is higher than the CCA and PLS. The KCCA and KPLS have smaller errors, but the kernel computation is very heavy.

VI. DISCUSSION

Recently, the PLS and CCA based methods have shown great performance in solving computer vision problems. It is essential to evaluate and compare the PLS and CCA based methods in a variety of vision problems, so that we can have a deeper understanding about their behaviors for both general and specific vision applications.

Sharma and Jacobs [33] have done a nice job very recently by comparing the linear CCA and linear PLS methods (they neither discuss the kernel based methods, nor analyze minimum numbers of dimensions) for face recognition with respect to pose variation, image resolution change, and photo-sketch matching. Their results show that the linear PLS outperforms the linear CCA in most cases. Although their face recognition is very different from our problem of estimating age, gender, and ethnicity jointly, their observation is quite consistent with ours, that is, the linear PLS outperforms the linear CCA.

TABLE II

THE RUNNING TIME FOR DIFFERENT METHODS ON LENOVO THINKPAD X61 (INTEL CORE 2, DUO CPU T8100, @2.1GHZ, 4G RAM, WINDOWS VISTA 64). TO FOCUS ON COMPARISONS BETWEEN THE LEARNING METHODS, THE TIME FOR FEATURE EXTRACTION USING BIF IS NOT INCLUDED.

	Dimension	Training Time (second)	Test Time (second)
CCA	3	51.27	0.697
rCCA	3	49.49	0.623
PLS	30	87.21	0.858
KPLS	35	7,450.6	72,516.4
KCCA	3	15,610.5	72,515.6

However, we also found that the regularized CCA (or rCCA) performs better than the PLS, and the standard (linear) CCA as well, in our problem. Since there is no discussion on regularization or any equations related to rCCA in the presentation in [33], our guess is that only the standard CCA was evaluated in [33]. Based on our results, we conjecture that the regularized CCA may outperform the PLS for the face recognition problems discussed in [33]. If that is true, one has to evaluate and compare different extensions of the PLS and CCA methods for computer vision problems, rather than just the standard formulations.

So, our study about the CCA and PLS based methods in this paper, is not only important and useful for estimating demographics, but also helpful to inspire more careful and detailed explorations of these methods for other computer vision problems. As a result, better performance and less number of feature dimensions might be expected in more applications.

VII. CONCLUSIONS

We have presented a novel finding that only three dimensions of features are needed to estimate age, gender and ethnicity altogether. The experimental validations have been performed on a very large database with more than 55,000 face images. We have analyzed the feature dimensionality problem using the rank theory for the CCA based methods. A systematic comparison of the behaviors between the CCA and PLS based methods has been given, including accuracies or errors with respect to dimensions, as well as the running time. Experimentally, the rCCA has a comparable running time, but lower errors than the CCA and PLS. Based on an overall consideration, the regularized CCA is recommended for practical uses because of its fast speed and a relatively small error. Our dimensionality analysis of the CCA based

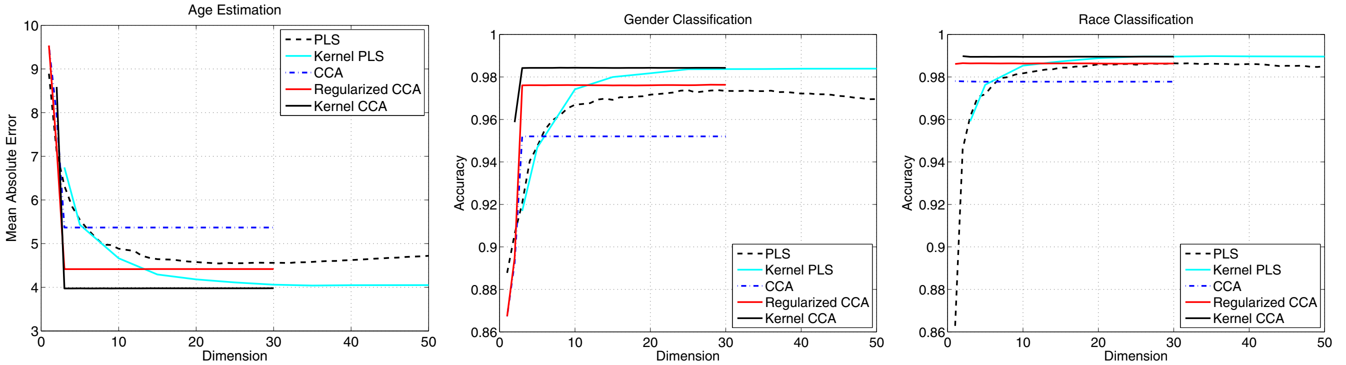

Fig. 2. Age estimation errors (MAEs in yrs.) and gender and ethnicity classification accuracies with respect to the feature dimensionality. Only three dimensions are sufficient for the CCA based methods, e.g., linear CCA, regularized CCA, and kernel CCA. The same features (e.g., the top three) are also used for gender and ethnicity classification. In contrast, both the linear and kernel PLS methods have errors reduced gradually w.r.t the dimensions, requiring more features (e.g., 30 or 35) to reach the highest accuracies, and even with performance drop when more features are used. A sorting of the methods in terms of age estimation errors is given as CCA > PLS > rCCA > KPLS > KCCA, which also hold for gender and ethnicity.

methods provides a new insight about the CCA, and may inspire further exploration of the behaviors of the CCA and PLS based methods in a broader range of vision problems.

References

[1] S. Baluja and H. A. Rowley. Boosting sex identification performance. *Intl. J. of Comput. Vision*, 71(1):111–119, 2007.

[2] R. Brunelli and T. Poggio. Hyperbf networks for gender classification. In *Proc. DARPA Image Understanding Workshop*, pages 311–314, 1992.

[3] Y. Fu, G.-D. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

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

[5] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM Conf. on Multimedia*, pages 307–316, 2006.

[6] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems 3*, pages 572–577, 1991.

[7] 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.

[8] 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.

[9] G.-D. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *IEEE Conf. on CVPR*, pages 657–664, 2011.

[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] S. Gutta, J. R. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Trans. on Neural Networks*, 11(4):948–960, 2000.

[13] S. Gutta and H. Wechsler. Gender and ethnic classification of face images. In *Intl. Conf. on AFGR*, pages 194–199, 1998.

[14] D. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor. A correlation approach for automatic image annotation. In *ADMA, LNAI 4093*, pages 681–692, 2006.

[15] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16:2639–2664, 2004.

[16] S. Hosoi, E. Takikawa, and M. Kawade. Ethnicity estimation with facial images. In *Intl. Conf. on AFGR*, 2004.

[17] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.

[18] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, 2007.

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

[20] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*, pages 605–611, 2009.

[21] X. Lu and A. Jain. Ethnicity identification from face images. In *Proc. SPIE Defense and Security Symposium*, 2004.

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

[23] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):707–711, 2002.

[24] 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.

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

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

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

[28] 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.

[29] 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.

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

[31] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE Conf. on CVPR*, 2005.

[32] G. Shakhnarovich, P. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *Intl. Conf. on Automatic Face and Gesture Recognition*, 2002.

[33] A. Sharma and D. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR*, pages 593–600, 2011.

[34] L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 33:194–200, 2011.

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

[36] 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.

[37] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *Intl. Conf. on Biometrics*, pages 464–473, 2007.