Head Pose Estimation Using Spectral Regression Discriminant Analysis

Caifeng Shan and Wei Chen
Philips Research
High Tech Campus 36, 5656AE Eindhoven, The Netherlands
{caifeng.shan, w.chen}@philips.com

Abstract

In this paper, we investigate a recently proposed efficient subspace learning method, Spectral Regression Discriminant Analysis (SRDA), and its kernel version SRKDA for head pose estimation. One important unsolved issue of SRDA is how to automatically determine an appropriate regularization parameter. The parameter, which was empirically set in the existing work, has great impact on its performance. By formulating it as a constrained optimization problem, we present a method to estimate the optimal regularization parameter in SRDA and SRKDA. Our experiments on two databases illustrate that SRDA, especially SRKDA, is promising for head pose estimation. Moreover, our approach for estimating the regularization parameter is shown to be effective in head pose estimation and face recognition experiments.

1. Introduction

Head pose estimation [15] is a key component in many applications for human-computer interaction and visual surveillance. For example, head pose can be used to analyze a person's focus of attention in smart environments. In practical face analysis systems, head pose estimation is crucial for high-level tasks such as face recognition and facial expression analysis. There have been a number of studies on head pose estimation, and many methods have been proposed [15].

One of successful methods is manifold or subspace learning [15], which seeks to model continuous head pose variation in the low-dimensional space [7, 1]. Traditional subspace methods such as Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) have been exploited for modeling head pose variation [6, 13]. Recently a number of graph-based subspace learning techniques have been proposed, such as Locality Preserving Projections (LPP) [12] and Locally Embedded Analysis (LEA) [7]. These methods have shown to be effective for head pose estimation. One common problem of these methods is the high computational cost due to the eigen-decomposition of dense matrices. To address this problem, recently Cai *et al.* [4, 3, 5] proposed an efficient subspace learning algorithm, Spectral Regression Discriminant Analysis (SRDA). By casting projective function learning into a regression framework, SRDA avoids eigen-decomposition of dense matrices. Compared to other subspace learning algorithms with the cubic-time complexity, SRDA has the linear-time complexity. SRDA has shown promising performance in different applications including face recognition [4], text clustering and categorization [3], spoken letter recognition [5], and handwritten digit classification [5]. SRDA has also been extended for nonlinear problems using the kernel trick, called Spectral Regression Kernel Discriminant Analysis (SRKDA) [2]. In the PASCAL VOC challenge 2008[1], SRKDA provides the best results on object recognition.

In this paper, we aim to investigate SRDA and SRKDA for head pose estimation, which has not be studied in the existing work. One important unsolved issue of SRDA is how to automatically determine an appropriate regularization parameter α [3]. The parameter α, which was empirically set in the existing work, controls the smoothness of the estimator. Experiments in [3, 4, 5] implies that the performance of SRDA is closely related to the choice of α. Therefore, estimating an optimal α is an essential problem for SRDA. In this work, by formulating the problem as a constrained optimization problem, we present a method to estimate the optimal regularization parameter in SRDA. Compared to the existing regularization parameter estimation methods including General Cross-Validation (GCV) [9] and the L-curve [11], our approach is much more efficient, and provides more accurate estimation. Our experiments on two databases illustrate that SRDA, especially SRKDA, is effective for head pose estimation. We also test our approach for estimating the regularization parameter in head pose estimation and face recognition experiments; its effectiveness and efficiency are evidently verified.

[1] pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/index.html

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

2. Spectral Regression Discriminant Analysis

Given a data set $\{\mathbf{x}_i\}_{i=1}^m$ in $\mathbb{R}^n$, dimensionality reduction methods aim to find a low-dimensional representation of $\{\mathbf{x}_i\}$. In the graph-based subspace learning methods [18], a symmetric matrix $W(= [w_{ij}]_{m\times m})$ is built, where w_{ij} is the weight of the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let $\mathbf{y} = [y_1, \cdots, y_m]^T$ be the 1-dimensional projection of $X = [\mathbf{x}_1, \cdots, \mathbf{x}_m]$, the optimal $\mathbf{y}$ is given by minimizing [12]

$$\sum_{i,j}(y_i - y_j)^2 w_{ij}. \qquad (1)$$

Eqn. (1) can be rewritten in the matrix form:

$$\sum_{i,j}(y_i - y_j)^2 w_{ij} = 2\mathbf{y}^T(D - W)\mathbf{y} = 2\mathbf{y}^T L\mathbf{y} \qquad (2)$$

where D is a diagonal matrix whose entries are column (or row) sums of W. A constraint $\mathbf{y}^T D\mathbf{y} = 1$ can be imposed [12], and the minimization problem reduces to find the optimal $\mathbf{y}^*$

$$\mathbf{y}^* = \arg\min_{\mathbf{y}^T D\mathbf{y}=1} \mathbf{y}^T L\mathbf{y} = \arg\min_{\mathbf{y}} \frac{\mathbf{y}^T L\mathbf{y}}{\mathbf{y}^T D\mathbf{y}}. \qquad (3)$$

Notice that $L = D - W$, the above optimization problem is equivalent to

$$\mathbf{y}^* = \arg\max_{\mathbf{y}^T D\mathbf{y}=1} \mathbf{y}^T W\mathbf{y} = \arg\max_{\mathbf{y}} \frac{\mathbf{y}^T W\mathbf{y}}{\mathbf{y}^T D\mathbf{y}}. \qquad (4)$$

which is solved as the maximum eigen-problem:

$$W\mathbf{y} = \lambda D\mathbf{y}. \qquad (5)$$

To obtain a projective mapping for all samples, including new testing samples, a linear function $y_i = f(x_i) = \mathbf{a}^T\mathbf{x}_i$ is chosen, i.e., $\mathbf{y} = X^T\mathbf{a}$, Eqn. (4) can be rewritten as

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T XWX^T\mathbf{a}}{\mathbf{a}^T XDX^T\mathbf{a}} \qquad (6)$$

which can be solved as the maximum eigen-problem

$$XWX^T\mathbf{a} = \lambda XDX^T\mathbf{a} \qquad (7)$$

With different choices of W, the above framework leads to different subspace learning methods. A common problem of these methods is the high computational cost from the eigen-decomposition of dense matrices. To address this problem, Cai *et al.* [3, 4, 5] introduced SRDA which, instead of solving the eigen-problem in Eqn. (7), derives the linear projective functions via two steps:

1. Solve the eigen-problem in Eqn. (5) to get $\mathbf{y}$.
2. Find $\mathbf{a}$ which satisfies $X^T\mathbf{a} = \mathbf{y}$. In reality, such $\mathbf{a}$ might not exist. A possible solution is to find $\mathbf{a}$ which best fits the equation in the least squares sense:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} \sum_{i=1}^m (\mathbf{a}^T\mathbf{x}_i - y_i)^2 \qquad (8)$$

In the first step, SRDA constructs weight matrix W by incorporating the label information. Suppose c classes in the data set and m_t samples in the t-th class, i.e., $m_1 + \cdots + m_c = m$, W is defined as

$$w_{ij} = \begin{cases} 1/m_t & \text{, if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ both belong to the } t\text{-th class} \\ 0 & \text{, otherwise} \end{cases} \qquad (9)$$

In the second step, the minimization problem in Eqn. (8) is usually ill-posed in reality. Instead of using maximum likelihood estimation [8], which leads to the ordinary least squares (OLS) estimator

$$\hat{\mathbf{a}} = (XX^T)^{-1}X\mathbf{y}, \qquad (10)$$

SRDA adopts the regularization technique [17] to obtain the regularized estimator:

$$\hat{\mathbf{a}}^* = (XX^T + \alpha I)^{-1}X\mathbf{y} \qquad (11)$$

where $\alpha(\geq 0)$ is a regularization parameter to control the smoothness of the estimator $\hat{\mathbf{a}}^*$.

As illustrated in [3, 4, 5], the performance of SRDA varies greatly as α changes; a non-zero α was empirically set in these experiments. An inappropriate setting of α may result in poor performance in practice. In the next section, we present an efficient method to estimate the optimal regularization parameter for SRDA.

3. Optically Regularized SRDA

3.1. Regularization Parameter Estimation

Before we present our approach for estimating α, we first discuss two existing methods for regularization parameter estimation.

Generalized Cross-Validation — GCV is based on statistical consideration that a good regularization parameter should predict the missing data. More precisely, if an arbitrary data point y_i of $\mathbf{y}$ is left out, the corresponding regularized solution $\hat{\mathbf{a}}^*$ should be able to predict it correctly. Accordingly estimating the optimal regularization parameter α is reduced to minimizing the GCV function[9]:

$$\mathcal{G}(\alpha) \equiv \frac{\|X^T\hat{\mathbf{a}}^*(\alpha) - \mathbf{y}\|_2^2}{(trace(I - X^T(X^T)^{-1}(\alpha)))^2}, \qquad (12)$$

where $(X^T)^{-1}(\alpha)$ is an arbitrary matrix which maps the right-hand side $\mathbf{y}$ onto the estimator $\hat{\mathbf{a}}^*(\alpha)$, i.e.,

$\hat{\mathbf{a}}^*(\alpha) = (X^T)^{-1}(\alpha)\mathbf{y}$. Solving Eqn. (12) has the time complexity of $O(m^3)$, therefore GCV is not suitable for large datasets due to its computational complexity.

L-curve — The L-curve method is based on a log-log plot of the norm of a regularized solution $\|\hat{\mathbf{a}}^*\|_2$ versus the norm of the corresponding residual $\|X^T\hat{\mathbf{a}}^* - \mathbf{y}\|_2$. In the L-shaped plot, the point with the maximum curvature locates where the solution $\hat{\mathbf{a}}^*$ changes in nature from being dominated by regularization errors to being dominated by the errors in the right-hand side. Hence the α value corresponding to the corner suggests an solution wherein both the solution norm and the residual norm simultaneously attain low values. The corner is derived by examining the curvature of points, which is computational demanding. Hansen [11] proposed an heuristic algorithm which starts with a few points and adaptively adds more points when necessary. As the calculation of the eigenvectors of XX^T cannot be avoided, the L-curve method has the time complexity of $O(n^3)$.

3.2. Optically Regularized SRDA

The difference between the regularized estimator $\hat{\mathbf{a}}^*$ in Eqn. (11) and the OLS estimator $\hat{\mathbf{a}}$ in Eqn. (10) can be analyzed by using Singular Value Decomposition (SVD). Suppose X is a wide matrix ($m > n$), we have $X^T = USV^T$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are unitary matrices, and $S \in \mathbb{R}^{m\times n}$ is the singular value matrix with the rank of r ($r \leq n$). The solution $\hat{\mathbf{a}}^*$ of Eqn. (11) can be reduced as

$$\begin{aligned}
\hat{\mathbf{a}}^* &= \left(XX^T + \alpha I\right)^{-1} X\mathbf{y} && (13)\\
&= \left(VS^TU^TUSV^T + \alpha VV^T\right)^{-1} VS^TU^T\mathbf{y} && (14)\\
&= V\left(S^TS + \alpha I\right)^{-1} S^TU^T\mathbf{y} && (15)\\
&= \sum_{i=1}^{n} \mathbf{v}_i \left(s_i^2 + \alpha\right)^{-1} s_i\mathbf{u}_i^T\mathbf{y} && (16)\\
&= \sum_{i=1}^{n} \mathbf{v}_i \left(\frac{\mathbf{u}_i^T\mathbf{y}}{s_i} \cdot \frac{s_i^2}{s_i^2 + \alpha}\right), && (17)
\end{aligned}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the orthonormal column vectors in U and V respectively, s_i represents the i-th largest singular value of X^T (when $i > r$, $s_i = 0$), and $\mathbf{y}$ is the constant response calculated from Eqn. (5). $\frac{s_i^2}{s_i^2+\alpha} \in [0,1]$ is called *filter factor* in [11]. Similarly, the solution $\hat{\mathbf{a}}$ in Eqn. (10) can be reduced as

$$\hat{\mathbf{a}} = \sum_{i=1}^{r} \mathbf{v}_i \left(\frac{\mathbf{u}_i^T\mathbf{y}}{s_i}\right). \qquad (18)$$

By comparing Eqn. (17) and Eqn. (18), we can find that both $\hat{\mathbf{a}}^*$ and $\hat{\mathbf{a}}$ are linear combinations of basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, and the regularization technique in SRDA changes only the coefficients of the linear combination by adding a filter factor. The coefficients in $\hat{\mathbf{a}}^*$ can be seen as functions of the singular values of X and the regularization parameter α.

In order to estimate the optimal α, we first investigate the constraint on α itself. SRDA is solved as the multivariate linear regression problem

$$X^T\mathbf{a} + \boldsymbol{\varepsilon} = \mathbf{y}, \qquad (19)$$

where $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of random error with $E[\boldsymbol{\varepsilon}] = 0$ and $Var[\boldsymbol{\varepsilon}] = \sigma^2 I_n$. A good regularization parameter α should reduce the mean square error (MSE) of the regularized estimator $\hat{\mathbf{a}}^*$ [8]. Otherwise, $\hat{\mathbf{a}}^*$ will be far away from the $\mathbf{a}$ computed from Eqn. (19). In order to evaluate the MSE of $\hat{\mathbf{a}}^*$ with respect to α, it is necessary to derive $E[D^2(\alpha)]$, where $D(\alpha)$ denotes the distance from $\hat{\mathbf{a}}^*$ to $\mathbf{a}$. For the OLS estimator $\hat{\mathbf{a}}$, we have

$$\begin{aligned}
\hat{\mathbf{a}} &= \mathbf{a} + (XX^T)^{-1}X\varepsilon && (20)\\
E[\hat{\mathbf{a}}] &= \mathbf{a}. && (21)
\end{aligned}$$

From Eqn. (10) and Eqn. (11), we can obtain the relationship between $\hat{\mathbf{a}}$ and $\hat{\mathbf{a}}^*$ as follows:

$$\begin{aligned}
\hat{\mathbf{a}}^* &= (XX^T)(XX^T + \alpha I)^{-1}\hat{\mathbf{a}}\\
&= \left(I - \alpha(XX^T + \alpha I)^{-1}\right)\hat{\mathbf{a}}\\
&=: R\hat{\mathbf{a}} && (22)
\end{aligned}$$

where R is used for simplicity. Hence we have

$$\begin{aligned}
&E[D^2(\alpha)]\\
=&E\left[(\hat{\mathbf{a}}^* - \mathbf{a})^T(\hat{\mathbf{a}}^* - \mathbf{a})\right] && (23)\\
=&E\left[(R\hat{\mathbf{a}} - R\mathbf{a} + R\mathbf{a} - \mathbf{a})^T(R\hat{\mathbf{a}} - R\mathbf{a} + R\mathbf{a} - \mathbf{a})\right] && (24)\\
=&E\left[(\hat{\mathbf{a}} - \mathbf{a})^TR^TR(\hat{\mathbf{a}} - \mathbf{a})\right] + (R\mathbf{a} - \mathbf{a})^T(R\mathbf{a} - \mathbf{a}). && (25)
\end{aligned}$$

Substituting Eqn. (20) in the first term of Eqn. (25), we obtain

$$\begin{aligned}
&E\left[(\hat{\mathbf{a}} - \mathbf{a})^TR^TR(\hat{\mathbf{a}} - \mathbf{a})\right]\\
=&E[\varepsilon^TX^T(XX^T)^{-1}R^TR(XX^T)^{-1}X\varepsilon] && (26)\\
=&Trace(X^T(XX^T)^{-1}R^TR(XX^T)^{-1}X\,Var[\varepsilon])\\
&\quad + E[\varepsilon]^2Trace(X^T(XX^T)^{-1}R^TR(XX^T)^{-1}X) && (27)\\
=&\sigma^2Trace((XX^T)^{-1}R^TR) && (28)
\end{aligned}$$

With Eqn. (22), we then have

$$\begin{aligned}
&E[D^2(\alpha)]\\
&= \sigma^2Trace\left((XX^T)^{-1}R^TR\right) + \mathbf{a}^T(R - I)^T(R - I)\mathbf{a} && (29)\\
&= \sigma^2Trace\left((XX^T + \alpha I)^{-1}(I - \alpha\,(XX^T + \alpha I)^{-1})\right)\\
&\quad + \mathbf{a}^T(\alpha^2(XX^T + \alpha I)^{-2})\mathbf{a} && (30)\\
&= \sigma^2\left(Trace(XX^T + \alpha I)^{-1} - \alpha\,Trace(XX^T + \alpha I)^{-2}\right)\\
&\quad + \alpha^2\mathbf{a}^T(XX^T + \alpha I)^{-2}\mathbf{a}. && (31)
\end{aligned}$$

Let $\mathbf{c} = [c_1, c_2, \cdots, c_n]^T$, which satisfies $\mathbf{c} = V^T\mathbf{a}$, we obtain

$$\begin{aligned} E[D^2(\alpha)] &= \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{(s_i^2+\alpha)} - \sum_{i=1}^{n} \frac{\alpha}{(s_i^2+\alpha)^2} \right) \\ &\quad + \alpha^2 \mathbf{c}^T (S^T S + \alpha I)^{-2} \mathbf{c} \qquad (32) \\ &= \sigma^2 \sum_{i=1}^{n} \frac{s_i^2}{(s_i^2+\alpha)^2} + \alpha^2 \sum_{i=1}^{n} \frac{c_i^2}{(s_i^2+\alpha)^2}. \qquad (33) \end{aligned}$$

It is obvious that, for any $\alpha > 0$, the first and second terms in Eqn. (33) are monotonically decreasing and increasing functions of α respectively. Taking the derivative of Eqn. (33) with respect to α, we have

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} = 2 \sum_{i=1}^{n} \frac{s_i^2(\alpha c_i^2 - \sigma^2)}{(s_i^2+\alpha)^3} \qquad (34)$$

Now we can see that

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} < 0, \quad for \quad 0 < \alpha < \min\left\{\frac{\sigma^2}{c_i^2}, \forall i\right\} \qquad (35)$$

and

$$\frac{\partial E[D^2(\alpha)]}{\partial \alpha} > 0, \quad for \quad \max\left\{\frac{\sigma^2}{c_i^2}, \forall i\right\} < \alpha < \infty. \qquad (36)$$

Thus, the minimum of MSE falls in the following interval of α

$$\left[\min\left\{\frac{\sigma^2}{c_i^2}\right\}, \max\left\{\frac{\sigma^2}{c_i^2}\right\}\right], \quad \forall i. \qquad (37)$$

Therefore, the optimal α should be neither too large nor too small.

The criteria we consider for estimating α is the robustness of the regularized estimator $\hat{\mathbf{a}}^*$ to noises in the data X. More precisely, $\hat{\mathbf{a}}^*$ with respect to the optimal α should be robust to the perturbation in the parameter space of the singular values $\{s_i\}$ of X, since $\hat{\mathbf{a}}^*$ can be seen as a function of α and $\{s_i\}$. In this way, estimating the optimal α is reduced to solve the minimization problem

$$\begin{aligned} &\alpha^* = \arg\min_{\alpha} E\left[\|\hat{\mathbf{a}}^*(\alpha, s+\epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\|_2^2\right] \\ &s.t. \ \ \alpha^* \in \left[\min\left\{\frac{\sigma^2}{c_i^2}\right\}, \max\left\{\frac{\sigma^2}{c_i^2}\right\}\right], \quad \forall i \qquad (38) \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, \delta^2)$ is the perturbation in the parameter space. Since

$$\begin{aligned} &\|\hat{\mathbf{a}}^*(\alpha, s+\epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\|_2^2 \\ &= \left(\hat{\mathbf{a}}^*(\alpha, s+\epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\right)^T \left(\hat{\mathbf{a}}^*(\alpha, s+\epsilon) - \hat{\mathbf{a}}^*(\alpha, s)\right) \qquad (39) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{y}^T \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_j \mathbf{u}_j^T \mathbf{y} \left(\frac{s_i}{s_i^2+\alpha} - \frac{s_i+\epsilon}{(s_i+\epsilon)^2+\alpha}\right)^2 \qquad (40) \end{aligned}$$

where $\mathbf{u}_i\mathbf{u}_j^T = 0$, $\mathbf{v}_i^T\mathbf{v}_j = 0$, when $i \neq j$, and $\mathbf{u}_i^T\mathbf{u}_j = 1$, $\mathbf{v}_i^T\mathbf{v}_j = 1$, when $i = j$. Thus, only the terms with $i = j$ remain in Eqn. (40). So we have

$$\alpha^* = \arg\min_{\alpha} E\left[\sum_{i=1}^{n} \mathbf{y}^T\mathbf{y} \left(\frac{s_i}{s_i^2+\alpha} - \frac{s_i+\epsilon}{(s_i+\epsilon)^2+\alpha}\right)^2\right]. \qquad (41)$$

Note that $\mathbf{y}^T\mathbf{y}$ is a positive constant, where $\mathbf{y}$ is calculated from Eqn. (5). Eqn. (41) is equivalent to

$$\alpha^* = \arg\min_{\alpha} E\left[\sum_{i=1}^{n} \left(\frac{(\alpha - s_i^2)\epsilon - s_i\epsilon^2}{(s_i^2+\alpha)((s_i+\epsilon)^2+\alpha)}\right)^2\right]. \qquad (42)$$

Considering ϵ is very small, we neglect the term $s_i\epsilon^2$. Thus, we have

$$\begin{aligned} \alpha^* &= \arg\min_{\alpha} \sum_{i=1}^{n} \left(\frac{\alpha - s_i^2}{(s_i^2+\alpha)((s_i+\epsilon)^2+\alpha)}\right)^2 E(\epsilon^2) \qquad (43) \\ &= \arg\min_{\alpha} \sum_{i=1}^{n} \frac{(\alpha - s_i^2)^2}{((s_i^2+\alpha)((s_i+\epsilon)^2+\alpha))^2} \cdot \delta^2 \qquad (44) \\ &= \arg\min_{\alpha} \sum_{i=1}^{n} \frac{(\alpha - s_i^2)^2}{\rho(\alpha)^2} \qquad (45) \end{aligned}$$

where $\rho(\alpha) := (s_i^2+\alpha)((s_i+\epsilon)^2+\alpha)$. Now the minimization problem (38) can be rewritten as follows:

$$\begin{aligned} &\alpha^* = \arg\min_{\alpha} \sum_{i=1}^{n} \frac{(\alpha - s_i^2)^2}{\rho(\alpha)^2} \\ &s.t. \ \ \alpha^* \in \left[\min\left\{\frac{\sigma^2}{c_i^2}\right\}, \max\left\{\frac{\sigma^2}{c_i^2}\right\}\right], \quad \forall i \qquad (46) \end{aligned}$$

It is difficult to solve the minimization problem in Eqn (46) analytically, as α^* and c_i are coupled. Considering the intrinsic bound of α and $\rho(\alpha)$, the problem is relaxed to a simple form

$$\alpha^* = \arg\min_{\alpha} \sum_{i=1}^{n} (\alpha - s_i^2)^2 \qquad (47)$$

By setting the derivative with respect to α equal to 0, we obtain the solution

$$\alpha^* = \frac{1}{n} \sum_{i}^{n} s_i^2. \qquad (48)$$

In practice, the above solution can be computed without extra computational cost, because $\{s_1^2, \cdots, s_n^2\}$ are the eigenvalues of the symmetric data matrix XX^T and their sum is the trace of XX^T, which has been calculated in SRDA.

3.3. Extension to SRKDA

Spectral regression was extended for Kernel Discriminant Analysis (KDA) [2]. Considering the problem in a feature space $\mathcal{F}$ induced by a nonlinear mapping ϕ: $\mathbb{R}^n \to \mathcal{F}$, KDA seeks the optimal projective function **v** in the feature space by solving the following optimization problem:

$$\mathbf{v}* = \arg\max_{\mathbf{v}} \frac{\mathbf{v}^T S_b^{\phi} \mathbf{v}}{\mathbf{v}^T S_w^{\phi} \mathbf{v}} = \arg\max_{\mathbf{v}} \frac{\mathbf{v}^T S_b^{\phi} \mathbf{v}}{\mathbf{v}^T S_t^{\phi} \mathbf{v}} \tag{49}$$

where S_b^{ϕ}, S_w^{ϕ} and S_t^{ϕ} are the between-class, within-class, and total scatter matrices in the feature space respectively. The solution is a linear combination of $\phi(\mathbf{x}_i)$ such that $\mathbf{v}* = \sum_{i=1}^{m} p_i \phi(\mathbf{x}_i)$. Let $\mathbf{p} = [p_1, \cdots, p_m]^T$, Eqn. (49) is equivalent to:

$$\mathbf{p}* = \arg\max_{\mathbf{p}} \frac{\mathbf{p}^T K W K \mathbf{p}}{\mathbf{p}^T K K \mathbf{p}} \tag{50}$$

where K is the kernel matrix, W is defined in Eqn. (9), and $\mathbf{v}^T\mathbf{v} = \mathbf{p}^T K \mathbf{p} = 1$. SRKDA first solves the eigenproblem $W\mathbf{y} = \lambda \mathbf{y}$ to get **y**, and then finds **p** by solving $(K + \alpha I)\mathbf{p} = \mathbf{y}$, where I is the identify matrix.

Considering the Gaussian kernel, it can be proved that the kernel matrix K is strictly positive definite when all vectors in K are different [14]. Thus, if the triangular matrix R obtained by the Cholesky decomposition of K, i.e., $K = R^T R$, we have $\mathbf{v} = R\mathbf{p}$. The optimal projective function **v** can then be calculated as

$$\mathbf{v} = R(R^T R + \alpha I)^{-1}\mathbf{y} = (RR^T + \alpha I)^{-1} R\mathbf{y} \tag{51}$$

Therefore, we can use the technique proposed above to determine the regularization parameter for SRKDA.

4. Experiments

4.1. Head Pose Estimation

We carried out experiments on two databases: (1) the FacePix database [1] contains head pose images of 30 subjects, each of which has 181 images of different head poses. In our experiments, for each subject, we selected 91 images representing pose angles from −90° to +90° at increments of 2°. (2) the Pointing'04 database [10] consists of 15 subjects. Each subject has 2 series of 93 images at different poses, including 13 yaw poses and 7 pitch poses, plus two extreme cases with yaw angle 0° and pitch angle 90° and −90°. We used all these images in our experiments, which are manually cropped and aligned based on nose. Some example images of two databases are shown in Figure 1 and Figure 2. All face images were down-scaled to 32×32 pixels in the gray-scale space, thus represented as 1024-dimensional vectors.

Figure 1. Example images from the FacePix database.

Figure 2. Example images from the Pointing'04 database.

p subjects ($p = 3, 6, 9, 12, 15, 18, 21$ in the FacePix database and $p = 2, 4, 6, 8, 10, 12$ in the Pointing'04 database) were randomly selected for training and the rest were used for testing. For each p, we average the results over 30 random splits and report the average estimation error. We show in Figure 3 and Figure 5 the performance of SRDA and SRKDA as a function of the parameter α, where the dimension of the projection subspace is set to $c - 1$, and c is the number of classes (i.e., $c = 91$ in the FacePix data and $c = 93$ in the Pointing'04 data). We choose an exponentially incremental sampling of α to present the complete variation. It is evident that the performance of SRDA and SRKDA changes greatly with the variation of α. SRDA achieves significantly better performance when the projection space is smoothed (with $\alpha > 10^{-1}$) than with α close to 0. But SRKDA achieves the best performance when is smoothed with $\alpha = 10^0$. There always exists an optimal regularization parameter in all these experiments, although the performance of SRDA does not change much for large α on the FacePix data and Pointing'04 (Yaw) data.

We applied our method to estimate the optimal regularization parameter α for SRDA and SRKDA. The SRDA and SRKDA with the optimal α estimated by our approach are denoted as OR-SRDA and OR-SRKDA. The average performance of OR-SRDA and OR-SRKDA are ploted in Figure 4 and Figure 6. For comparison, we show the best performance of SRDA and SRKDA obtained by exhaustively examining different α, and the performance of SRDA and SRKDA with $\alpha = 0$ (denoted as Z-SRDA and Z-SRKDA). For SRDA, we also included results of PCA, LDA, and the supervised LPP. We can draw the following conclusions from these figures: (1) The regularization parameter is essential for SRDA and SRKDA. SRDA and SRKDA with $\alpha = 0$ provides much worse performance than that with a proper α. (2) With the α automatically estimated by our approach, OR-SRDA and OR-SRKDA achieves much similar results to the best performance of SRDA and SRKDA respectively. This evidently illustrates our approach is effective for estimating α in SRDA and SRKDA. (3) Regarding head pose estimation, OR-SRDA performs better than PCA, LDA, and LPP; OR-SRKDA provides better performance than OR-SRDA. These comparative experiments demon-

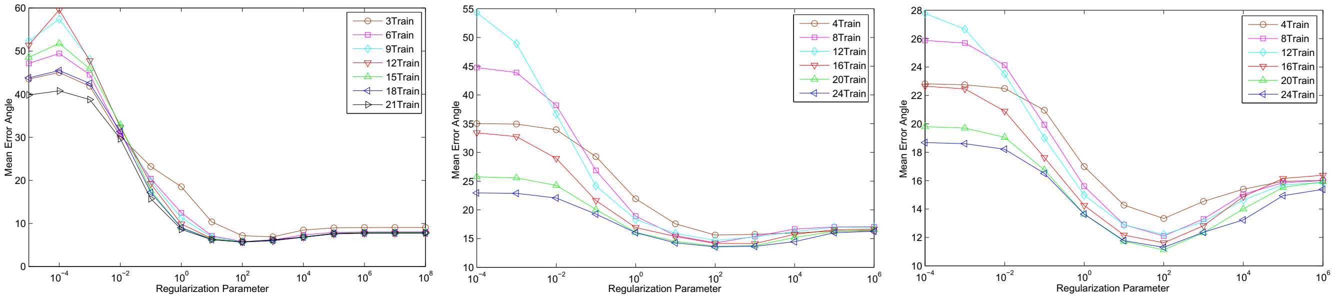

Figure 3. The mean error angle of SRDA with respect to different α. (Left) the FacePix database; (Middle) the Pointing'04 database in the Yaw direction); (Right) the Pointing'04 database in the Pitch direction.

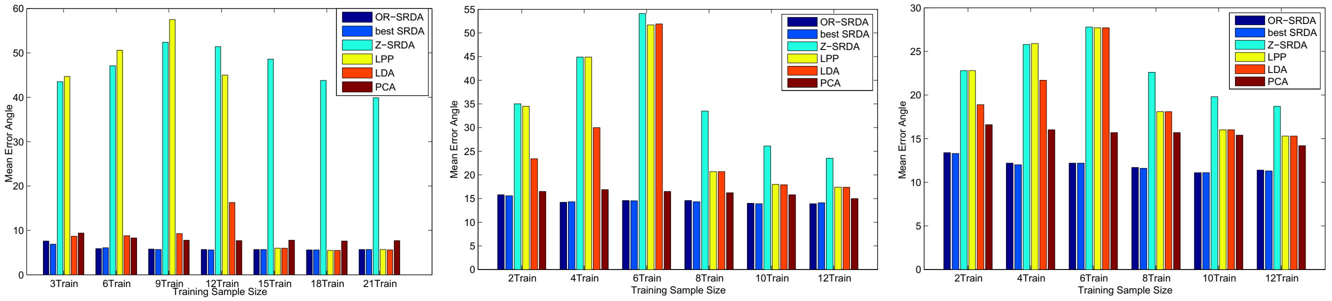

Figure 4. Head pose estimation performance of SRDA and other methods. (Left) the FacePix database; (Middle) the Pointing'04 database in the Yaw direction; (Right) the Pointing'04 database in the Pitch direction.

strate that SRDA and SRKDA are promising for head pose estimation.

We compare our approach with the GCV and L-curve methods for estimating the regularization parameter on the FacePix database. Figure 7 shows the performance of SRDA with the parameters estimated by these methods and also their computational cost [2]. The experiments were performed in a linux PC (CPU 3.0 GHz, cache 1024kb, RAM 4GB). It is observed that the GCV method fails to estimate a proper regularization parameter for SRDA on the database, while the L-curve method performs much better, and our method outperforms both. Regarding the computational cost, our approach is much more efficient than other methods.

The performance of OR-SRDA varies with the reduced dimension. In the above experiments, the reduced dimension is set as $c - 1$, and c is the number of classes. Figure 8 shows the average performance versus dimensionality reduction on FacePix database. It is observed that OR-SRDA indeed achieves the best performance with the reduced dimension of $c - 1$.

We implemented a real-time head pose estimation system based on the above approach. Some examples are shown in Figure 9.

4.2. Face Recognition

To further test our approach for estimating α, we conducted experiments on face recognition using the CMU PIE database [16], Following [4], we used the data online [3], which includes face images of 68 subjects with near-frontal poses and different illuminations and facial expressions, resulting 170 images for each subject. For each subject, p (=10,20,30,40,50,60) images were randomly selected for training and the rest were used for testing. For each p, we average the results over 30 random splits and report the mean.

Figure 10 shows the performance of SRDA as a function of the parameter α, where the dimension of the reduced subspace is set $c - 1$ ($c = 68$). Similarly, we observe that the regularization parameter has impact on the performance of SRDA. Compared to head pose estimation (shown in Figure 3), where the training images in the same class are from different subjects, the "oversmooth" effect due to a large α is very obvious for face recognition. For example, the error rate with $\alpha = 1e3$ is bigger than that of $\alpha = 0$. However, as observed before, there always exists an optimal regularization parameter in each experiment. We applied our method to estimate the optimal α, and compare different algorithms in Figure 11. Again we can see that, with the α estimated by

[2] The GCV and L-curve methods are implemented in Regularization Tools Version 4.1, http://www2.imm.dtu.dk/˜pch/Regutools/

[3] http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html

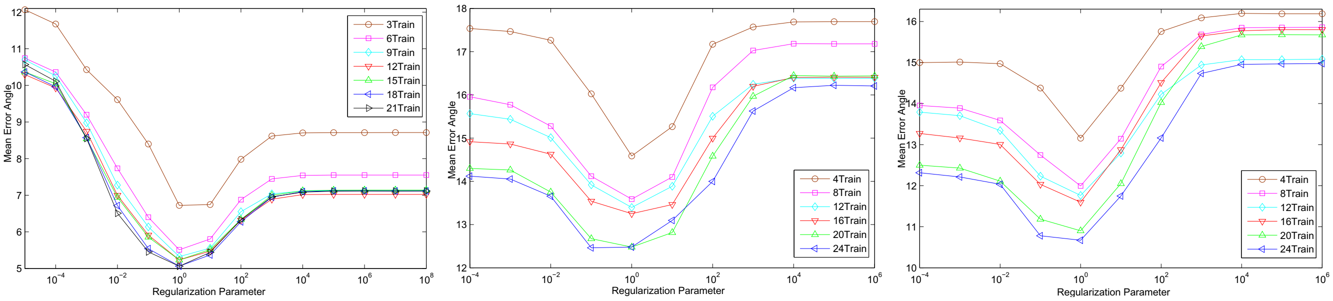

Figure 5. The mean error angle of SRKDA with respect to different α. (Left) the FacePix database; (Middle) the Pointing'04 database in the Yaw direction; (Right) the Pointing'04 database in the Pitch direction.

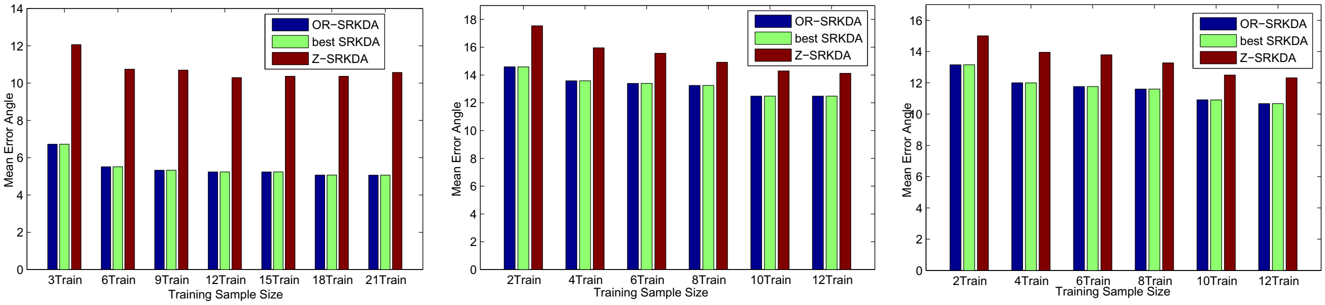

Figure 6. Head pose estimation performance of SRKDA. (Left) the FacePix database; (Middle) the Pointing'04 database in the Yaw direction; (Right) the Pointing'04 database in the Pitch direction.

our approach, OR-SRDA provides results that is much close to the best performance of SRDA obtained by exhaustively searching α. Moreover, the difference between OR-SRDA and the best SRDA becomes much smaller with larger training data. This further verify the validity of our approach for estimating the optimal regularization parameter in SRDA. It is also observed that OR-SRDA consistently provides superior performance to LDA and LPP for face recognition.

5. Conclusion

SRDA is an efficient subspace learning method, which has been proven powerful in different applications. In this paper, we investigate SRDA and its kernel version SRKDA for head pose estimation. Determining an appropriate regularization parameter is an important unsolved issue for SRDA. By formulating it as a constrained optimization problem, we present a method to estimate the optimal regularization parameter in SRDA and SRKDA. Our experiments on two databases illustrate the SRDA especially SRKDA, is promising for head pose estimation. Moreover, our approach for estimating the regularization parameter is shown to be effective in head pose estimation and face recognition experiments.

References

[1] V. N. Balasubramanian, S. Krishna, and S. Panchanathan. Person-independent head pose estimation using biased manifold embedding. *EURASIP Journal of Advances in Signal Processing*, 8(1):1–15, 2008. 1, 5

[2] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. In *IEEE International Conference on Data Mining (ICDM)*, Oct. 2007. 1, 5

[3] D. Cai, X. He, and J. Han. Regularized locality preserving indexing via spectral regression. In *ACM International Conference on Information and Knowledge Management*, Nov. 2007. 1, 2

[4] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. In *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2007. 1, 2, 6

[5] D. Cai, X. He, and J. Han. SRDA: An efficient algorithm for large scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1–12, Jan. 2008. 1, 2

[6] I. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang. Head pose estimation using fisher manifold learning. In *Proceeding IEEE International Workshop Analysis and Modeling of Faces and Gestures*, pages 203–207, 2003. 1

[7] Y. Fu and T. S. Huang. Graph embedded analysis for head pose estimation. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 3–8, 2006. 1

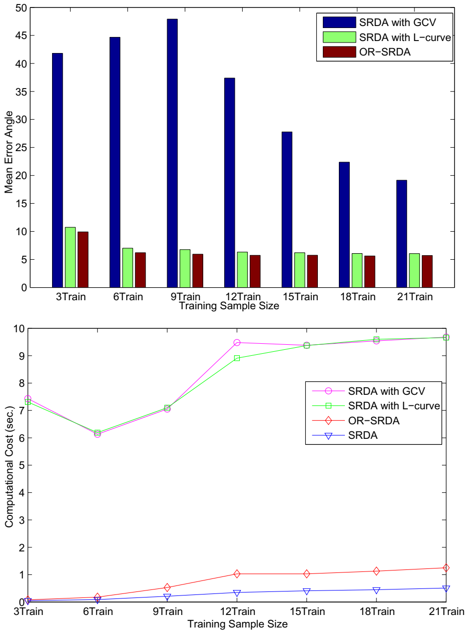

Figure 7. Comparison of different regularization parameter estimation methods on the FacePix database. (Top) The performance of SRDA with the regularization parameters estimated by different methods; (Bottom) Computational cost of different methods compared to SRDA.

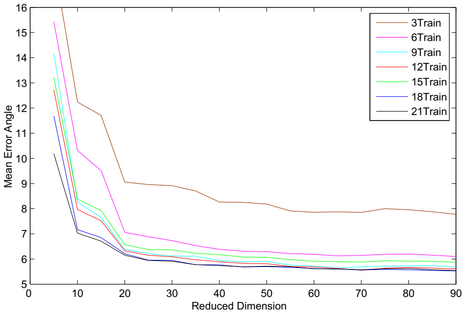

Figure 8. The performance of OR-SRDA versus dimensionality reduction on the FacePix database.

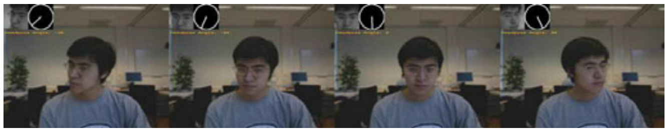

Figure 9. Examples of real-time head pose estimation.

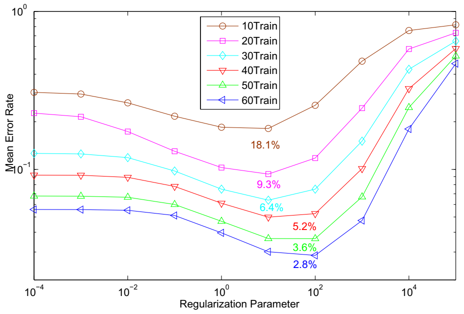

Figure 10. The average recognition results of SRDA with respect to different α on the PIE database.

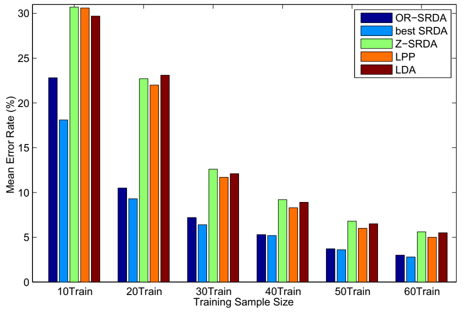

Figure 11. Face recognition performance on the PIE database.

[8] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, December 2006. 2, 3

[9] G. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979. 1, 2

[10] N. Gourier and J. Letessier. Estimating face orientation from robust detection of salient facial features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004. 5

[11] P. C. Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. Society for Industrial and Applied Mathematics, 1998. 1, 3

[12] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, Mar. 2005. 1, 2

[13] S. McKenna and S. Gong. Real-time face pose estimation. In *Real-Time Imaging*, number 5, pages 333–347, 1998. 1

[14] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(1):11–22, 1986. 5

[15] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1):34–58, 2008. 1

[16] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25, pages 1615–1618, 2003. 6

[17] A. N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Math*, 4:1624–1627, 1963. 2

[18] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, Jan. 2007. 2