Square Loss based Regularized LDA for Face Recognition Using Image Sets

Yanlin Geng[1], Caifeng Shan[2], and Pengwei Hao[1,3]
[1]Center for Information Science, Peking University, Beijing 100871, China
[2]Philips Research, High Tech Campus 36, 5656 AE Eindhoven, The Netherlands
[3]Dept. Computer Science, Queen Mary, University of London, London E1 4NS, UK
gengyanlin@cis.pku.edu.cn, caifeng.shan@philips.com, phao@dcs.qmul.ac.uk

Abstract

In this paper, we focus on face recognition over image sets, where each set is represented by a linear subspace. Linear Discriminant Analysis (LDA) is adopted for discriminative learning. After investigating the relation between regularization on Fisher Criterion and Maximum Margin Criterion, we present a unified framework for regularized LDA. With the framework, the ratio-form maximization of regularized Fisher LDA can be reduced to the difference-form optimization with an additional constraint. By incorporating the empirical loss as the regularization term, we introduce a generalized Square Loss based Regularized LDA (SLR-LDA) with suggestion on parameter setting. Our approach achieves superior performance to the state-of-the-art methods on face recognition. Its effectiveness is also evidently verified in general object and object category recognition experiments.

1. Introduction

Face recognition has been studied in computer vision for decades. Depending on the data available, face recognition can be performed on a single image or an image set. Recently attention has been shifted towards image set based recognition, as it is easy to acquire and handle large quantities of image data nowadays. By exploring information from multiple images, better performance can be achieved compared to recognition based on a single image. Many of the previous work exploited the temporal continuity between images [12, 19], assuming the images were recorded from consecutive observations. Following recent work [9], in this paper we focus on face recognition over sets of images that may be unordered.

By representing each image set as a linear subspace, Kim *et al*. [9] exploited canonical correlations for comparing face image sets. They proposed Discriminant Analysis of Canonical Correlations (DCC), which extends Fisher Linear Discriminant Analysis (LDA) [4] for learning over image sets that maximizes canonical correlations of within-class sets and minimizes canonical correlations of between-class sets. By mapping the subspaces into an empirical feature space using the kernel method, Hamm and Lee [7] recently proposed Grassmann Discriminant Analysis (GDA) for image set classification, which applies regularized Fisher LDA to the subspace data.

In this work, we also represent each set of face images by a linear subspace [9, 17]. LDA is adopted for discriminative learning, where Fisher Criterion and Maximum Margin Criterion (MMC) [11] are discussed. We investigate the relation between regularization on Fisher Criterion and MMC, and present a unified framework for regularized LDA. With the framework, the ratio-form maximization of regularized Fisher LDA can be reduced to the difference-form optimization with an additional constraint. By incorporating the empirical loss as the regularization term, we introduce a generalized Square Loss based Regularized LDA (SLR-LDA) with suggestion on appropriate parameter setting. We apply SLR-LDA over image sets for face recognition, which achieves superior performance to the state-of-the-art methods. Its effectiveness is also evidently verified in general object and object category recognition experiments.

2. Key Ingredients of Our Approach

2.1. Subspace Representation of Image Sets

An image (or vector) set can be considered to span a linear subspace, and represented by the orthonormal basis matrix. Figure 1 illustrates some face images from one image set, and the corresponding subspace representation using the first 5 leading bases. To measure the similarity between image sets, a distance measure of linear subspaces is needed. Most of the used distances are based on canonical correlations [9], which are cosines of principal angles $0 \leq \theta_1 \leq ... \leq \theta_d \leq \pi/2$ between two linear subspaces. Given the orthonormal basis matrices of two subspaces, $B_1, B_2 \in R^{m \times d}$, where m is the dimension of fea-

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

Figure 1. Top two rows: some images from an image set in our face database. Bottom row: the 5 leading bases of the subspace that spanned by the image set.

ture vectors and d is the number of basis vectors, canonical correlations are the singular values of $B_1^T B_2$. In [7], Hamm and Lee discussed several distances and adopted the Projection metric

$$d_P(B_1, B_2) = (d - \sum \cos^2 \theta_i)^{0.5} \quad (1)$$

which produces a positive definite kernel function

$$k(B_1, B_2) = \sum \cos^2 \theta_i = \|B_1^T B_2\|_F^2 \quad (2)$$

thus it can be easily used for kernel methods in Hilbert spaces (see details in [7]). This distance is also computationally efficient, as only the Frobenius norm needs to be calculated, without explicitly computing the canonical correlations. Therefore, we also use this distance in our work.

2.2. Linear Discriminant Analysis

LDA seeks a linear transformation $f(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$ that maximizes the between-class distance d_B and minimizes the within-class distance d_W simultaneously, where $\mathbf{x} \in R^m$ is a data sample and $\mathbf{W} = (\mathbf{w_1}, \cdots, \mathbf{w_r}) \in R^{m\times r} (r \leq m)$ is the transformation matrix. There are mainly two criteria to implement this idea. Fisher LDA is to maximize the Fisher Criterion [4]

$$\frac{d_B}{d_W} = \frac{tr(\mathbf{W}^T S_B \mathbf{W})}{tr(\mathbf{W}^T S_W \mathbf{W})} \quad (3)$$

where S_B and S_W are the between-class and within-class scatter matrices respectively. The solution is the eigenvectors of the generalized eigen-problem $S_B\mathbf{w} = \lambda S_W\mathbf{w}$ associated with the largest eigenvalues. The main problem of Fisher Criterion is that the matrix inverse S_W^{-1} is involved, while S_W could be singular, especially in small sample size problems or in kernel methods [14]. To address this problem, PCA is usually adopted to reduce the dimension of the data [1]; however, this step may remove discriminative information.

The other criterion is Maximum Margin Criterion introduced in [11], which is to maximize the difference $d_B - d_W$. With the constraint $\mathbf{w}^T\mathbf{w} = 1$, the solution is obtained by solving the eigen-problem $(S_B - S_W)\mathbf{w} = \lambda\mathbf{w}$. Compared to Fisher Criterion, no matrix inverse is involved in MMC, thus it avoids the singularity problem and is more computationally efficient. MMC is closely related to Fisher Criterion, which can be derived from MMC by incorporating the constraint $tr(\mathbf{W}^T S_W \mathbf{W}) = 1$. Recent papers [18, 20] have further extended MMC to weighted MMC, that is, to maximize $\beta d_B - \alpha d_W$.

2.3. Regularized LDA

Regularization has been considered to address the singularity problem in Fisher LDA. A common approach is adding μI to S_W [5, 14]. In [8], a diagonal matrix was introduced into S_W. Lu *et al.* [13] further modified S_W to $\eta S_W + S_B$. On the contrary, few studies have considered regularization on MMC, since MMC does not have the singularity problem. In a more recent work [16], Xue *et al.* presented Discriminatively Regularized Least-Squares Classification (DRLSC), which aims to maximize d_B, while at the same time minimize d_W and the square loss $L_{sq} = \sum \|f(\mathbf{x}_i) - \mathbf{y}_i\|^2$. Specifically, DRLSC is to minimize

$$J_{DR} = L_{sq} + \eta d_W - (1 - \eta) d_B \quad (4)$$

so can be regarded as the regularized (weighted) MMC.

An important but unsolved problem is, is there any relation between regularization on Fisher Criterion and MMC? As discussed above, with a suitable constraint MMC leads to Fisher LDA. Considering MMC doesn't has the singularity problem and is more computationally feasible, it is necessary to explore the relation between regularized Fisher LDA and regularized MMC. This can result in better understanding on regularized LDA. In the next section, we investigate their relation and present a general framework for regularized LDA, which unifies both regularized Fisher LDA and regularized MMC as a constrained optimization problem. A main problem of DRLSC is how to set the weighting coefficients of d_W and d_B. Without sound theoretical analysis, an ad hoc setting η and $1 - \eta$ were used [16], and the authors failed to provide a solution on how to determine η. With the proposed unified framework, we obtain the relation of the weighting coefficients, leading to a flexible approach to parameter setting.

3. A Unified Framework for Regularized LDA

Regularized Fisher LDA aims to maximize

$$J_{FC}(v) = \frac{d_B(v)}{d_W(v) + \frac{reg(v)}{\alpha}} \quad (5)$$
$$s.t. \quad v \in \Omega$$

where v is the variable, for example v can be $\mathbf{W}$ and $\mathbf{b}$ in the transformation $f(\mathbf{x}) = \mathbf{x}^T\mathbf{W} + \mathbf{b}$, $reg(v) \geq 0$ represents the regularization term, $\alpha > 0$ is the regularization parameter, and Ω is a closed set[1]. We use reg/α instead of $\alpha \cdot reg$ for the simplicity of the subsequent analysis. Denote the maximum of $J_{FC}(v)$ as $J_{FC,\max} = \max_{v\in\Omega} J_{FC}$. We have Theorem 3.1 (with proof in Appendix A).

Theorem 3.1. *(1). The maxima of $J_{FC}(v)$ also minimizes $J(v)$, and $\min J(v) = 0$, where*

$$J(v) = reg + \alpha d_W - \frac{\alpha}{J_{FC,\max}} d_B \quad (6)$$
$$s.t. \quad v \in \Omega$$

(2). The minima of $J(v)$ also maximizes $J_{FC}(v)$.
(3). $J_{\min}(\alpha, \beta) = 0$ if and only if $\beta = \frac{\alpha}{J_{FC,\max}}$, where

$$J_{\min}(\alpha, \beta) = \min_v (reg + \alpha d_W - \beta d_B) \quad (7)$$
$$s.t. \quad v \in \Omega$$

According to Theorem 3.1(1)(2), maximizing J_{FC} is equivalent to minimizing J. Thus we can reformulate the problem as

$$\min J(v) = reg + \alpha d_W - \beta d_B \quad (8)$$
$$s.t. \quad v \in \Omega$$
$$\beta = \frac{\alpha}{J_{FC,\max}}$$

$J(v)$ can be viewed as the regularization on weighted MMC, so regularized Fisher Criterion is equivalent to regularized weighted MMC with the constraint $\beta = \alpha/J_{FC,\max}$. Therefore, regularized Fisher LDA and regularized MMC can be unified in a framework. In this way, the ratio-form maximization of regularized Fisher LDA can be reduced to the difference-form optimization with an additional constraint, which is easy to be solved to a certain extend. Our framework also sheds some insight on the relationship between parameters α and β, rather than an ad hoc setting $\alpha + \beta = 1$ in [16]. Furthermore, the parameter setting becomes more flexible. That is, instead of predefining α and finding the optimal β, we can also predefine β and solve

$$\min J(v) = reg + \alpha d_W - \beta d_B \quad (9)$$
$$s.t. \quad v \in \Omega$$
$$\alpha = \beta J_{FC,\max}$$

This is very useful, as in some cases it is easy to predefine one of α or β.

[1] $v \in \Omega$ is a short expression for some general constraints such as the equal constraint $ceq(v) = 0$ and the unequal one $cleq(v) \leq 0$. A closed Ω makes sure that the optima is within Ω.

Given α and β, $J(v) = reg + \alpha d_W - \beta d_B$ can be minimized. However, only α is predefined, and $\beta = \alpha/J_{FC,\max}$ depends on the unknown $J_{FC,\max}$ (or vice versa). To derive the optimal β, we have the following theorem.

Theorem 3.2.

$$\beta = \arg_{\beta'}\{J_{\min}(\alpha, \beta') = 0\} \quad (10)$$

Proof. From Theorem 3.1(3), we know $J_{\min}(\alpha, \beta) = 0$ if and only if $\beta = \alpha/J_{FC,\max}$. Thus the optimal β satisfies $J_{\min}(\alpha, \beta) = 0$, and any other β makes it nonzero. □

Similarly, if β predefined, the optimal α can be obtained from

$$\alpha = \arg_{\alpha'}\{J_{\min}(\alpha', \beta) = 0\} \quad (11)$$

Whether α or β predefined depends on which one is easy to select. In a specific application, if $J_{\min}(\alpha, \beta)$ can be formulated explicitly (for example, SLR-LDA in Section 4), we can obtain the optimal β according to Theorem 3.2 (or α from Eqn. (11)). In cases where $J_{\min}(\alpha, \beta)$ cannot be explicitly formulated, we have an iterative approach to compute the optimal β (or α). Due to the page limit, we omit the discussion in this paper.

4. Square Loss Regularized LDA (SLR-LDA)

For regularized LDA, it is critical to choose a suitable regularization term reg. As discussed above, a common approach is adding μI to S_W [5, 14]. This kind of regularization term is helpful for smoothing the function. However, for classification purposes, a smoothness constraint is not always useful. In supervised learning, the empirical loss $L = \sum V(f(\mathbf{x}_i), \mathbf{y}_i)$ between the desired and actual outputs provides the discriminative information for classification. Following DRLSC [16], we take the empirical loss as reg in this work. As shown above, DRLSC is regularized MMC $reg + \alpha d_W - \beta d_B$, and can be unified in our framework. The main problems of DRLSC include using an ad hoc setting $\alpha + \beta = 1$, and how to select the parameter η. Here, with the proposed framework, we provide a generalized Square Loss based Regularized LDA (SLR-LDA). We also give suggestion on how to choose the parameter. Our approach can be extended for image set classification using the kernel technique.

For C-class problems, let $\mathbf{y}_i \in R^{1\times C}$ be the class label for sample $\mathbf{x}_i \in R^m$. The indicator matrix $\mathbf{Y}$ is defined with its i-th row as $\mathbf{y}_i$. If $\mathbf{x}_i$ is in class c, $\mathbf{y}_i$ can be defined as a vector of all zeros except that its c-th element is one. For a linear classifier $f(\mathbf{x}) = \mathbf{x}^T\mathbf{W} + \mathbf{b}$, where $\mathbf{W} \in R^{m\times C}$ and $\mathbf{b} \in R^{1\times C}$, the square loss is defined as

$$L_{sq} = \sum \|f(\mathbf{x}_i) - \mathbf{y}_i\|^2 = \|\mathbf{X}^T\mathbf{W} + \mathbf{b}_N - \mathbf{Y}\|_F^2 \quad (12)$$

where the i-th column of $\mathbf{X} \in R^{m\times N}$ is the sample $\mathbf{x}_i$, N is the number of samples, and all the rows of $\mathbf{b}_N \in R^{N\times C}$ are $\mathbf{b}$. Denote $1_N \in R^N$ as the vector with all ones, we have $\mathbf{b}_N = 1_N\mathbf{b}$. The scatter distances d_B and d_W are defined using the kNN method [2, 16], which gives the distances as

$$d_B = tr(\mathbf{W}^T\mathbf{X}L_B\mathbf{X}^T\mathbf{W}) \tag{13}$$

$$d_W = tr(\mathbf{W}^T\mathbf{X}L_W\mathbf{X}^T\mathbf{W}) \tag{14}$$

See [16] for details. SLR-LDA is to maximize $\frac{d_B}{d_W+L_{sq}/\alpha}$, with a predefined α, which is reduced to minimizing

$$\begin{aligned} &J = L_{sq} + \alpha d_W - \beta d_B \\ &s.t. \quad \beta = \alpha/J_{FC,\max} \end{aligned} \tag{15}$$

We first provide the explicit form of $J_{\min}(\alpha,\beta) = \min_{\mathbf{W},\mathbf{b}}(L_{sq}+\alpha d_W-\beta d_B)$. Based on $\|A\|_F^2 = tr(A^TA)$, we have

$$\begin{aligned} J =& L_{sq} + \alpha d_W - \beta d_B \\ =& \|\mathbf{X}^T\mathbf{W} + \mathbf{b}_N - \mathbf{Y}\|_F^2 \\ &+ tr(\mathbf{W}^T\mathbf{X}(\alpha L_W - \beta L_B)\mathbf{X}^T\mathbf{W}) \\ =& tr(\mathbf{W}^T\mathbf{X}(I_N + \alpha L_W - \beta L_B)\mathbf{X}^T\mathbf{W}) \\ &- 2tr(\mathbf{W}^T\mathbf{X}(\mathbf{b}_N - \mathbf{Y})) + \|\mathbf{b}_N - \mathbf{Y}\|_F^2 \end{aligned} \tag{16}$$

where $I_N \in R^{N\times N}$ is the identity matrix. The derivatives with respect to $\mathbf{b}$ and $\mathbf{W}$ are

$$\begin{cases} \frac{1}{2}\frac{\partial J}{\partial \mathbf{b}} &= 1_N^T(\mathbf{X}^T\mathbf{W} + \mathbf{b}_N - \mathbf{Y}) \\ \frac{1}{2}\frac{\partial J}{\partial \mathbf{W}} &= \mathbf{X}L_{\alpha,\beta}\mathbf{X}^T\mathbf{W} - \mathbf{X}(\mathbf{Y} - \mathbf{b}_N) \end{cases} \tag{17}$$

where $L_{\alpha,\beta} := I_N + \alpha L_W - \beta L_B$. By setting the derivatives to zeros, we get the minima as

$$\begin{cases} \tilde{\mathbf{b}} &= \frac{1_N^T(R_{\alpha,\beta}-I_N)\mathbf{Y}}{1_N^T(R_{\alpha,\beta}-I_N)1_N} \\ \tilde{\mathbf{W}} &= (\mathbf{X}L_{\alpha,\beta}\mathbf{X}^T)^\dagger\mathbf{X}(\mathbf{Y} - \mathbf{b}_N) \end{cases} \tag{18}$$

where $R_{\alpha,\beta} := \mathbf{X}^T(\mathbf{X}L_{\alpha,\beta}\mathbf{X}^T)^\dagger\mathbf{X}$, and $\dagger$ denotes Moore-Penrose pseudo inverse. By substituting the minima back into J, we obtain the minimum

$$\begin{aligned} &J_{\min}(\alpha,\beta) = \\ &tr(\mathbf{Y}^T(R_{\alpha,\beta} - I_N)(\frac{1_N1_N^T(R_{\alpha,\beta} - I_N)\mathbf{Y}}{1_N^T(R_{\alpha,\beta} - I_N)1_N} - \mathbf{Y})) \end{aligned} \tag{19}$$

Given the predefined α, the optimal β can be obtained from Theorem 3.2. Similarly, we can also derive the optimal α with the predefined β.

4.1. Parameter β Selection

In reality, the minimum of J should not be $-\infty$. We show in Appendix B that this can be achieved by setting $\beta\lambda_{\max}(L_B) \le 1$, where $\lambda_{\max}(L_B)$ is the maximal eigenvalue of the matrix L_B. Considering $\beta \ge 0$ for the discrimination purpose, we have

$$\beta = \frac{1-\sigma}{\lambda_{\max}(L_B)} \text{ with } \sigma \in [0,1] \tag{20}$$

So in this case β can be easily predefined. With the defined β, the optimal α is solved from Eqn. (11).

4.2. SLR-LDA over Image Sets

Using the kernel trick, we extend SLR-LDA for image set classification using the subspace representation. Assume the classifier can be represented as $f(\mathbf{x}) = \mathbf{b} + \sum \mathbf{W}_{i.}k(\mathbf{x}_i,\mathbf{x})$, we have

$$L_{sq} = \|K^T\mathbf{W} + \mathbf{b}_N - \mathbf{Y}\|_F^2 \tag{21}$$

$$d_B = tr(\mathbf{W}^TKL_BK^T\mathbf{W}) \tag{22}$$

$$d_W = tr(\mathbf{W}^TKL_WK^T\mathbf{W}) \tag{23}$$

where K is the kernel matrix. The solution is

$$\begin{cases} \tilde{\mathbf{b}} &= \frac{1_N^T(R_{\alpha,\beta}-I_N)\mathbf{Y}}{1_N^T(R_{\alpha,\beta}-I_N)1_N} \\ \tilde{\mathbf{W}} &= (KL_{\alpha,\beta}K^T)^\dagger K(\mathbf{Y} - \mathbf{b}_N) \end{cases} \tag{24}$$

where $R_{\alpha,\beta} = K^T(KL_{\alpha,\beta}K^T)^\dagger K$. In our experiments, K is calculated using Eqn. (2).

5. Experiments

5.1. Face Recognition

We evaluate the proposed SLR-LDA for image set based face recognition, where the task is to classify an unknown set of face images to one of the training classes, each of which also represented by face image sets.

Database of Face Image Sets — Since there is no large database of face image sets public available, we built our own face database. The database consists of face images extracted from various videos, including movies[2], TV programs [3], face video database[3], and face videos we captured about researchers in our group. There are in total 115 subjects with 62 females, and totally 1,862 image sets in our database. Each person has 4 to 40 image sets and each set has about 45 to 60 images. Some example face images from three sets are shown in Figure 2. All the faces were automatically detected using the Viola-Jones cascaded face detector [15], without any further processing such as alignment or background segmentation. As can be observed, the faces exhibit great variations in terms of head pose, illumination, expression, resolution, and occlusion. So our database is

[2] mi.eng.cam.ac.uk/ ~oa214/academic

[3] www.perceptual-vision.com/db/video/faces/cvglab

Figure 2. Example image sets in our face database.

much closer to the conditions for face recognition in real-life environments. In our experiments, following [9], each face image is first histogram equalized, and then re-scaled to 20×20 pixels, thus represented by a 400-dimensional feature vector.

Experimental Settings and Comparative Methods — In our experiments, p (=2, 3, 4, 5) image sets were randomly selected for training and the rest were used for testing[4]. For each p, we run experiments with 50 random splits and report the average performance. For each image set, PCA was performed to obtain the subspace representation. The subspace dimension was set as 15, which preserves about 98% of the data energy of each set.

We compare SLR-LDA with the state-of-the-art methods, including DRLSC [16], GDA1 [7], and DCC [9]. As discussed above, DRLSC is a special case of our approach with an ad hoc parameter setting. GDA1 is Kernel LDA with a smoothing regularization term μI, while DCC extends LDA for image sets using canonical correlations. For SLR-LDA, we use $\sigma = 0.5$ and thus $\beta = 0.5/\lambda_{\max}(L_B)$ in all experiments. For DRLSC, the optimal $\eta = 0.999$ was found by scanning through $[0, 1]$, which was used in our experiments. The number of nearest neighbors used in DRLSC and SLR-LDA is $k = 10$. For DCC, as recommended in [9], we reduced the dimension of the data using PCA: for training with 5 image sets, the reduced dimension was 150; for training with 2-4 sets, the reduced dimension was 120. The DCC experiments were repeated 20 times, because it is very time consuming.

Experimental Results — The recognition rates of different methods are shown in Figure 3. As can be observed, all the methods examined perform better when using more image sets for training. This is because, with more image sets used, face variations can be better represented, resulting in better discriminative functions. SLR-LDA consistently outperforms DCC and GDA1 with a clear margin. The main reason could be that the SLR-LDA incorporates the useful discriminative information by taking the loss as regularization term. In contrast, GDA1 adopts a smoothness regularization constraint that may not be sufficient for discrimination among classes, and DCC reduces the data dimension using PCA to avoid the singularity problem, which may remove the discriminative information. The benefit of using the loss as regularization term is also illustrated by DRLSC. With a properly chosen η, DRLSC also provides better performance than GDA1 and DCC. It is observed that SLR-LDA performs lightly better than DRLSC. This validates our theoretical analysis in Section 3, where we show that the relation should be $\beta = \alpha/J_{FC,\max}$ rather than $\alpha + \beta = 1$ in DRLSC. We further illustrate in the next paragraph that DRLSC is sensitive to η, while SLR-LDA is robust to $\sigma \in [0, 1]$.

[4]There are 12 persons in the database only having 4 image sets, so they are discarded when 5 sets are used for training.

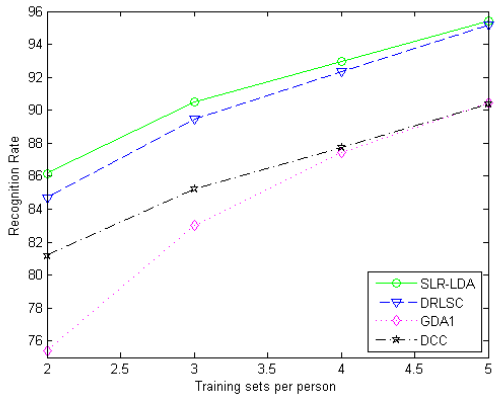

Figure 3. (Best viewed in color) Face recognition rates of different methods with different training sizes.

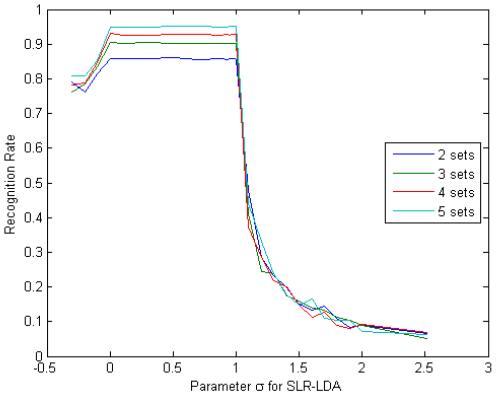

Figure 4. (Best viewed in color) Face recognition rates versus parameter σ for SLR-LDA.

Effect of Parameters in SLR-LDA and DRLSC — In Section 4.1, we make the suggestion $1 - \beta\lambda_{\max}(L_B) = \sigma \in [0, 1]$; σ was set as 0.5 in all the experiments. Here we verify our suggestion, and investigate the performance with different σ in SLR-LDA. The performance of SLR-LDA w.r.t different σ is shown in Figure 4. For comparison,

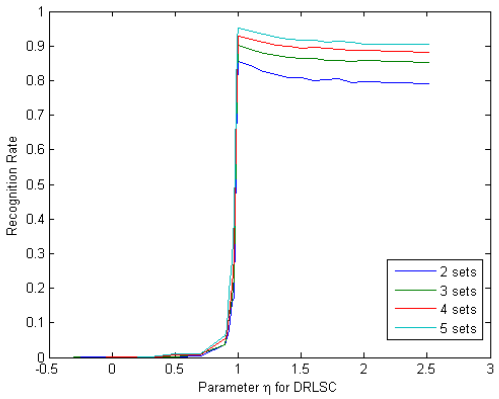

Figure 5. (Best viewed in color) Face recognition rates versus parameter η for DRLSC.

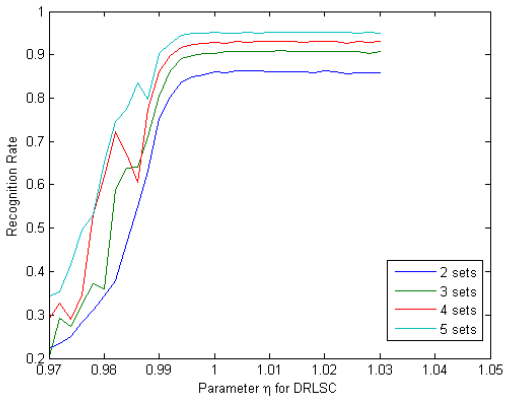

Figure 6. (Best viewed in color) Detailed face recognition rates versus parameter η close to 1 for DRLSC.

we also show the performance of DRLSC w.r.t η in Figure 5. We have several observations. (1) The suggestion $\sigma \in [0, 1]$ holds. For $\sigma < 0$ or $\sigma > 1$, the performance of SLR-LDA deteriorates sharply. (2) SLR-LDA is not sensitive to $\sigma \in [0, 1]$, and provides consistently stable results. While DRLSC is sensitive to the parameter $\eta \in [0, 1]$. (3) DRLSC performs well with η close to 1. In this case DRLSC is reduced to minimizing $L_{sq} + d_W$, which is conflict with the basic idea to maximize d_B. This is possibly due to the ad hoc setting $\alpha + \beta = 1$.

5.2. Object Recognition On ALOI Database

To further verify its effective, we also apply SLR-LDA for general object or object category recognition. ALOI database [6] contains 1000 objects captured from 72 views. In our experiments, we used images of 100 objects, which were segmented from the background using the masks provided in the database. For each object, every 12 images with consecutive view angles were set as an image set, resulting 6 image sets per object. The images were processed using the same procedure on the face database. The dimension of subspace representation was 10. Again p (=2, 3, 4, 5) image sets were randomly selected for training and the rest were used for testing. For each p, we run experiments with 50 random splits and report the average performance. We also compare SLR-LDA with DRLSC, GDA1, and DCC. In SLR-LDA, σ is set to 0.5, while in DRLSC, the optimal $\eta = 0.995$ was used.

Figure 7 shows the recognition results of different methods. Again, all the methods achieve higher recognition rates when more sets are used for training. Once more, SLR-LDA performs slightly better than the optimal DRLSC. Both SLR-LDA and DRLSC perform better than GDA1 and DCC, due to taking the empirical loss into account in the regularization term. We also show the effect of parameters in SLR-LDA and DRLSC in Figure 7. It is observed that SLR-LDA achieves stable performance with $\sigma \in [0, 1]$, but its performance deteriorates sharply for $\sigma < 0$ or $\sigma > 1$. For DRSLC, it also provides good performance with η close to 1, which is conflict with the basic idea to maximize d_B. Compared with the results on the face database, the recognition rates are much lower on the ALOI database. This is possibly because the images in ALOI have great variation on view angle, and each set contains much fewer images.

5.3. Object Category Recognition on ETH80

ETH80 database [10] contains 8 object categories, each of which has 10 image sets. Each set is composed of 41 images taken from different views. The dimension of the subspace representation was 10. Similarly, p (=2, 3, 4, 5) image sets were randomly selected for training and the rest were used for testing. For each p, we run experiments with 50 random splits and report the average performance. We also compare SLR-LDA with DRLSC, GDA1, and DCC. In SLR-LDA, σ is set to 0.5 in all experiments. In DRLSC, we examined different values of $\eta \in [0, 1]$ and use the optimal setting of 0.995.

The experimental results are shown in Figure 8. Again, SRL-LDA performs comparably or slightly better than the optimal DRLSC. Both SLR-LDA and DRLSC provide superior performance to GDA1 and DCC, and this reinforces our findings in experiments on face and ALOI databases. In most cases, GDA1 outperforms DCC, which is consistent with the observation in [7]. While 5 sets in each category are used for training, the performance among these methods is small. This is because the ETH80 database is built under well controlled conditions, and as more sets are used for training, characteristics of different categories could be well captured. The recognition rates are much higher than that on the ALOI database, mainly due to a smaller number of classes. The observed effect of parameters in SLR-LDA and DRLSC is also similar to our findings on face database and ALOI database.

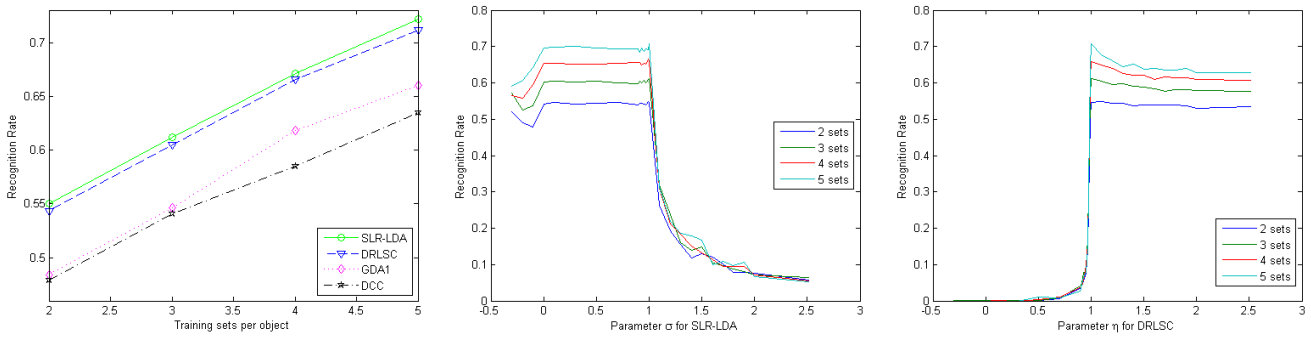

Figure 7. (Best viewed in color) (*Left*) Recognition rates of different methods using different training sizes on the ALOI database; (*Middle*) Recognition rates versus parameter σ for SLR-LDA; (*Right*) Recognition rates versus parameter η for DRLSC.

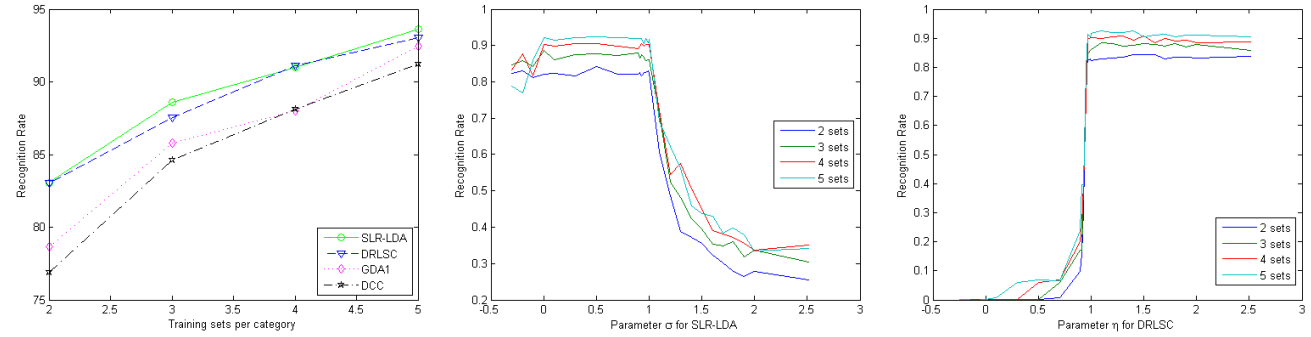

Figure 8. (Best viewed in color) (*Left*) Recognition rates of different methods using different training sizes on the ETH80 database; (*Middle*) Recognition rates versus parameter σ for SLR-LDA; (*Right*) Recognition rates versus parameter η for DRLSC.

6. Conclusions

Face recognition over image sets is addressed in this paper. A general framework for regularized LDA is presented, which unifies both Fisher Criterion and Maximum Margin Criterion. The ratio-form maximization of regularized Fisher LDA can now be reduced to the difference-form optimization with an additional constraint. By incorporating the empirical loss as the regularization term, we introduce a generalized Square Loss based Regularized LDA with suggestion on appropriate parameter setting. Our approach achieves superior performance to the state-of-the-art methods in face recognition and object (category) recognition experiments on several databases.

Appendix

A. Proof to Theorem 3.1

Eqn. (5) can be reformulated as the following equation

$$reg(v) + \alpha d_W(v) - \frac{\alpha}{J_{FC}(v)} d_B(v) = 0 \qquad (25)$$

which will be used in the following analysis.

(1) Suppose $u \in \Omega$ maximizes J_{FC}, for $\forall v \in \Omega$ we have

$$J(u) = reg(u) + \alpha d_W(u) - \frac{\alpha}{J_{FC,max}} d_B(u) \qquad (26)$$

$$= reg(u) + \alpha d_W(u) - \frac{\alpha}{J_{FC}(u)} d_B(u) = 0, \qquad (27)$$

$$J(v) = reg(v) + \alpha d_W(v) - \frac{\alpha}{J_{FC,\max}} d_B(v) \qquad (28)$$

$$\geq reg(v) + \alpha d_W(v) - \frac{\alpha}{J_{FC}(v)} d_B(v) = 0. \qquad (29)$$

So u also minimizes J and

$$\min J = J(u) = 0. \qquad (30)$$

(2) Suppose $u \in \Omega$ minimizes J, we have

$$0 = J(u) = reg(u) + \alpha d_W(u) - \frac{\alpha}{J_{FC,\max}} d_B(u) \qquad (31)$$

$$\Leftrightarrow \quad reg(u) + \alpha d_W(u) = \frac{\alpha}{J_{FC,\max}} d_B(u) \qquad (32)$$

$$\Leftrightarrow \quad J_{FC}(u) = \frac{d_B(u)}{d_W(u) + \frac{reg(u)}{\alpha}} = J_{FC,\max}, \qquad (33)$$

So u also maximizes J_{FC}.

(3) We have proved in (1) that when $\beta' = \alpha/J_{FC,\max}$, $J_{\min}(\alpha, \beta') = 0$. Now suppose $J_{\min}(\alpha, \beta') = 0$ and $u \in \Omega$ is the minima, we need to prove $\beta' = \alpha/J_{FC,\max}$. For $\forall v \in \Omega$ we have

$$reg(v) + \alpha d_W(v) - \beta' d_B(v) \geq 0 \quad (34)$$

$$\Leftrightarrow \quad reg(v) + \alpha d_W(v) \geq \beta' d_B(v) \quad (35)$$

$$\Leftrightarrow \quad \frac{\alpha}{\beta'} \geq \frac{d_B(v)}{d_W(v) + reg(v)/\alpha} = J_{FC}(v) \quad (36)$$

The equality holds for the minima u, which means $\alpha/\beta' = J_{FC}(u)$. So we have $J_{FC}(v) \leq \alpha/\beta' = J_{FC}(u)$, which means $J_{FC}(u) = J_{FC,\max}$. Thus $\beta' = \alpha/J_{FC,\max}$

B. Suggestion on Parameter Selection

In order that it makes sense to minimize J in Eqn. (16), the minimum should not be $-\infty$. This suggests us to require that $L_{\alpha,\beta}$ is positive semi-definite (PSD).

The definition of a PSD matrix H is that $z^T H z \geq 0, \forall z$. According to Eqn. (16), J is quadratic. For simplicity we use $v = [\mathbf{W}^T \ \mathbf{b}^T]^T$ to represent the variables $\mathbf{b}$ and $\mathbf{W}$. Using Eqn. (17), we obtain the Hessian matrix

$$H := \frac{\partial^2 J}{\partial v^2} = \begin{bmatrix} \mathbf{X} L_{\alpha,\beta} \mathbf{X}^T & \mathbf{X} 1_N \\ 1_N^T \mathbf{X}^T & 1_N^T 1_N \end{bmatrix} \quad (37)$$

H should be PSD; otherwise if $z_0^T H z_0 < 0$, then $\lim_{\mu\to\infty} (\mu z_0)^T H (\mu z_0) = -\infty$, thus the minimum becomes $-\infty$. So now for $\forall p \in R^C$ and $z = [p^T \ 0]^T$, we have $z^T H z \geq 0$, which results into $(\mathbf{X}^T p)^T L_{\alpha,\beta} (\mathbf{X}^T p) \geq 0$. This suggests that we can require $L_{\alpha,\beta} = I_N + \alpha L_W - \beta L_B$ to be PSD.

As I_N, L_W and L_B are all PSD (remember that scatter distances $d_W \geq 0$ and $d_B \geq 0$), and the addition of two PSD matrices is also PSD, we just need to choose β so that $I_N - \beta L_B$ is PSD, which means $z^T (I_N - \beta L_B) z \geq 0, \forall z$. This can be achieved by selecting $\beta \lambda_{\max}(L_B) \leq 1$, where $\lambda_{\max}(L_B)$ is the maximal eigen-value of matrix L_B.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces versus fisherfaces: Recognition using class spefic linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 2

[2] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–713, 2007. 4

[3] E. Douglas-Cowie, R. Cowie, and M. Schroder. A new emotion database: Considerations, sources and scope. In *ISCA Workshop on Speech and Emotion: A conceptual Framework for Research*, pages 39–44, 2000. 4

[4] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. 1, 2

[5] J. H. Friedman. Regularized discriminant analysis. *J. Am. Stat. Assoc*, 84:165–175, 1989. 2, 3

[6] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 2005. 6

[7] J. Hamm and D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International Conference on Machine Learning (ICML)*, 2008. 1, 2, 5, 6

[8] J. P. Hoffbeck and D. A. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:763–767, 1996. 2

[9] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007. 1, 5

[10] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 6

[11] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Conference on Advances in Neural Information Processing Systems (NIPS)*, 2004. 1, 2

[12] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 1

[13] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26:181–191, 2005. 2

[14] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *IEEE Workshop on Neural Networks for Signal Processing*, 1999. 2, 3

[15] P. Vioa and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004. 4

[16] H. Xue, S. Chen, and Q. Yang. Discriminatively regularized least-squares classification. *Pattern Recognition*, 42:93–104, 2009. 2, 3, 4, 5

[17] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 318–323, 1998. 1

[18] W. Zheng, C. Zhou, and L. Zhao. Weighted maximum margin discriminant analysis with kernels. *Neurocomputing*, 67:357–362, 2005. 2

[19] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91:214–245, 2003. 1

[20] Y. Zhu and E. Sung. Margin-maximization discriminant analysis for face recognition. In *IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 609–612, 2004. 2