REGULARIZED D-LDA FOR FACE RECOGNITION

Juwei Lu, K.N. Plataniotis, A.N. Venetsanopoulos

Bell Canada Multimedia Laboratory
The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto, Toronto, M5S 3G4, ONTARIO, CANADA

ABSTRACT

Linear Discriminant Analysis (LDA) is derived from the optimal Bayes classifier when classes are assumed to be Gaussian with identical covariance matrices. However, it is well known that the distribution of face images under a perceivable variation in viewpoint, illumination or facial expression, is highly nonlinear and complex. The Quadratic Discriminant Analysis (QDA) which relaxes the identical covariance assumption and allows for nonlinear discriminant boundaries to be formed, seems to be a better choice. However, the applicability of QDA to problems, such as face recognition, where the number of training samples is much smaller than the dimensionality of the sample space is problematic due to the increased number of parameters to be learned. In this paper, we propose a new regularized discriminant analysis method that effectively solves the so-called "small sample size" problem in very high-dimensional face image space. Extensive experimentation performed on the FERET database indicates that the proposed methodology outperforms traditional methods such as Eigenfaces, QDA and Direct LDA in a number of application scenarios.

1. INTRODUCTION

Face recognition (FR) systems, utilizing Linear Discriminant Analysis (LDA) techniques have been shown to be very successful [1, 2, 3, 4]. However, the so-called "plug-in" covariance matrix estimates widely used in these LDA-based approaches often suffer from the so-called "small sample size" (SSS) problem which exists in high-dimensional pattern recognition tasks where the number of available training samples is smaller than the dimensionality of the samples. The traditional solution to the SSS problem is to utilize principal component analysis (PCA) in conjunction with LDA (PCA+LDA) as it was done for example in Fisherfaces [1]. Recently, more effective solutions, called Direct LDA (D-LDA) methods, have been presented [2, 3, 4].

Although successful in many cases, LDA-based methods often fail to deliver good performance when face patterns are subject to large variations in viewpoints, illumination or facial expression, which result in a highly nonlinear and complex distribution. The limited success of these methods should be attributed to their linear nature [5, 6]. LDA can be considered as a special case of the optimal Bayes classifier when each class is subjected to a Gaussian distribution with identical covariance structure. Obviously, the assumption behind LDA is highly incorrect in practical FR tasks. As a result, it is reasonable to assume that a better solution to this inherent complex problem could be achieved using quadratic methods, such as the Quadratic Discriminant Analysis (QDA), which allows for complex discriminant boundaries to be formed. However, the SSS problem affects QDA more than LDA, since QDA requires much more training than LDA due to the increased number of parameters. To deal with such a situation, Friedman proposed a regularization technique of discriminant analysis (RDA) in the Gaussian framework [7]. The purpose of the regularization is to reduce the variance related to the sample-based estimates at the expense of potentially increased bias. Although RDA relieves to a great extent the SSS problem and performs well even when the number of training samples per class (L) is comparable to the dimensionality of the samples (D), it still fails when $L << D$, which is the case in most FR applications. For example, if only $L \in [2, 7]$ samples per subject are available for training while the dimensionality of the space is up to $D = 17154$, the RDA cannot be successfully implemented.

In the paper, we propose a new regularized discriminant analysis method called RD-LDA by incorporating the D-LDA technique into the RDA framework. The RD-LDA provides a comprehensive solution to the SSS problem hampering both LDA and QDA. It will be shown that, adjusting the parameters of the RD-LDA, we can obtain a number of new/traditional discriminant analysis methods such as Yang's D-LDA (YD-LDA) [3], Juwei's D-LDA (JD-LDA) [4], direct QDA (D-QDA), nearest center (NC) and weighted nearest center (WNC) classifiers.

2. METHODS

2.1. Determining the optimal discriminant features

Given a training set containing C classes $\{\mathbf{Z}_i\}_{i=1}^{C}$, with each class consisting of a number of face images: $\mathbf{Z}_i = \{\mathbf{z}_{ij}\}_{j=1}^{C_i}$, a total of $N = \sum_{i=1}^{C} C_i$ face images are available in the set. Each image is represented as a column vector of length $D(= I_w \times I_h)$, i.e. $\mathbf{z}_{ij} \in \mathbb{R}^D$, where $I_w \times I_h$ is the image size, and $\mathbb{R}^D$ denotes the D-dimensional real space.

Let $\mathbf{S}_{BTW}$ and $\mathbf{S}_{WTH}$ denote the between- and within-class scatter matrices of the training image set respectively. LDA determines a set of optimal discriminant basis vectors, denoted by $\{\psi_k\}_{k=1}^{M}$, so that the ratio of the between-

The authors would like to thank the FERET Technical Agent, the U.S. National Institute of Standards and Technology (NIST) for providing the FERET database.

and within-class scatters is maximized [8]. Assuming $\Psi = [\psi_1, \ldots, \psi_M]$, the maximization can be achieved by solving the following eigenvalue problem,

$$\Psi = \arg\max_{\Psi} \frac{|(\Psi^T \mathbf{S}_{BTW} \Psi)|}{|(\Psi^T \mathbf{S}_{WTH} \Psi)|} \tag{1}$$

Assuming that $\mathbf{S}_{WTH}$ is non-singular, the basis vectors Ψ correspond to the first M eigenvectors with the largest eigenvalues of $(\mathbf{S}_{WTH}^{-1}\mathbf{S}_{BTW})$. Due to the SSS problem, a degenerated $\mathbf{S}_{WTH}$ may be generated in FR tasks. Traditional methods, for example Fisherfaces [1], attempt to solve the SSS problem by using a PCA step to remove the null space of $\mathbf{S}_{WTH}$. However, it has been shown that the null space may contain the most significant discriminant information [2, 3].

Recently, the so-called direct LDA (D-LDA) approach have been introduced to avoid the shortcomings existing in traditional solutions to the SSS problem [2, 3, 4]. The basic premise behind the approach is that the null space of $\mathbf{S}_{WTH}$ may contain significant discriminant information if the projection of $\mathbf{S}_{BTW}$ is not zero in that direction, and that no significant information will be lost if the null space of $\mathbf{S}_{BTW}$ is discarded. Based on the finding, it can be concluded that the optimal discriminant features exist in the complement space of the null space of $\mathbf{S}_{BTW}$, which has a dimensionality $M = C - 1$. In [3, 4], the subspace denoted as $\mathcal{H}$ is scaled to have $\mathcal{H}^T\mathbf{S}_{BTW}\mathcal{H} = \mathbf{I}$, where $\mathbf{I}$ is the $(M \times M)$ identity matrix. The projection of $\mathbf{S}_{WTH}$ in $\mathcal{H}$, $\mathcal{H}^T\mathbf{S}_{WTH}\mathcal{H}$, is then estimated using sample analogs. However, when training sample number per class is small enough, even the projection $\mathcal{H}^T\mathbf{S}_{WTH}\mathcal{H}$ is ill- or poorly-posed. To this end, a modified optimization criterion represented as $\Psi = \arg\max_{\Psi} \frac{|\Psi^T \mathbf{S}_{BTW} \Psi|}{|\Psi^T \mathbf{S}_{BTW} \Psi + \Psi^T \mathbf{S}_{WTH} \Psi|}$, is proposed to use in JD-LDA [4] instead of Eq.1 used in YD-LDA [3]. The modified criterion introduced a considerable degree of regularization to reduce the variance of the plug-in estimate in ill- or poorly-posed situations. It will be shown later that such a regularization is only a special case of the proposed RD-LDA.

2.2. Regularized D-LDA (RD-LDA)

The number of face classes C is usually a small value, and comparable to the number of training samples N in most FR tasks, e.g. $C = 49$ and $N \in [98, 343]$ in the experiments reported here. Thus, it becomes appropriate to perform a RDA [7] in the low-dimensional subspace $\mathcal{H}$, where the most significant discriminant information are remained.

To this end, we firstly project the original face images into $\mathcal{H}$, obtaining a representation $\mathbf{y}_{ij} = \mathcal{H}^T\mathbf{z}_{ij}$ where $i = 1, \ldots, C$, $j = 1, \ldots, C_i$. The regularized sample covariance matrix estimate of class i in $\mathcal{H}$, $\hat{\Sigma}_i(\lambda, \gamma)$, can be expressed as,

$$\hat{\Sigma}_i(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_i(\lambda) + \frac{\gamma}{M} tr[\hat{\Sigma}_i(\lambda)]\mathbf{I} \tag{2}$$

where

$$\hat{\Sigma}_i(\lambda) = \frac{1}{C_i(\lambda)} [(1 - \lambda)\mathbf{S}_i + \lambda \mathbf{S}] \tag{3}$$

$$C_i(\lambda) = (1 - \lambda)C_i + \lambda N \tag{4}$$

$$\mathbf{S}_i = \sum_{j=1}^{C_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T \tag{5}$$

$$\mathbf{S} = \sum_{i=1}^{C} \mathbf{S}_i = N \cdot \mathcal{H}^T \mathbf{S}_{WTH} \mathcal{H} \tag{6}$$

(λ, γ) is a pair of regularization parameters, and $\bar{\mathbf{y}}_i$ is the projection of the mean of class i in $\mathcal{H}$.

In the FR procedure, any input query image $\mathbf{z}$ is firstly projected into the subspace $\mathcal{H}$: $\mathbf{y} = \mathcal{H}^T\mathbf{z}$. its class label i^* then can be inferred through $i^* = \arg\min_i d_i(\mathbf{y})$ based on QDA, where $d_i(\mathbf{y})$ is the well-known Mahalanobis (quadratic) distance between $\mathbf{y}$ and class $\bar{\mathbf{y}}_i$, and has the following expression,

$$d_i(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}}_i)^T \hat{\Sigma}_i^{-1}(\lambda, \gamma)(\mathbf{y} - \bar{\mathbf{y}}_i) + \ln \left|\hat{\Sigma}_i(\lambda, \gamma)\right| - 2 \ln \pi_i \tag{7}$$

where $\pi_i = C_i/N$ is the prior probability of class i.

The regularization parameter λ $(0 \leq \lambda \leq 1)$ controls the amount that the $\mathbf{S}_i$ are shrunk toward $\mathbf{S}$. The other parameter γ $(0 \leq \gamma \leq 1)$ controls shrinkage of the class covariance matrix estimates toward a multiple of the identity matrix. Under the regularization scheme, a QDA can be performed without suffering the high variance of the plug-in estimates even when the dimensionality of the subspace $\mathcal{H}$ is comparable to the number of available training samples. We refer to the approach as regularized D-LDA, hereafter RD-LDA.

Since the RD-LDA is derived from the D-LDA and RDA, it has close relationship with a series of traditional discriminant analysis methods. Firstly, the four corners defining the extremes of the (λ, γ) plane represent four well-known classification algorithms, as summarized in Table 1, where the prefix 'D-' means that all these methods are developed in the subspace $\mathcal{H}$ derived from the D-LDA technique. Due to the criterion of Eq.1 used in YD-LDA [3], it is obvious that YD-LDA is actually a standard LDA implemented in $\mathcal{H}$. Also, we have $\Sigma_i = \alpha\left(\frac{\mathbf{S}}{N} + I\right) = \alpha\left(\mathcal{H}^T\mathbf{S}_{WTH}\mathcal{H} + I\right)$ when $(\lambda = 1, \gamma = \eta)$, where $\alpha = \left(\frac{tr[\mathbf{S}/N]}{tr[\mathbf{S}/N]+M}\right)$ and $\eta = \frac{M}{tr[\mathbf{S}/N]+M}$. In this situation, it is not difficult to see that RD-LDA is equivalent to JD-LDA [4]. In addition, a set of intermediate discriminant classifiers between the five traditional ones can be obtained when we smoothly slip the two regularization parameters in their domains. The purpose of RD-LDA is to find the (λ^*, γ^*) that give the best correct recognition rate for a particular FR task.

Table 1. A series of algorithms derived from RD-LDA.

Algs.	D-NC	D-WNC	D-QDA	YD-LDA	JD-LDA
λ	1	0	0	1	1
γ	1	1	0	0	η

3. EXPERIMENTAL RESULTS

3.1. The FR Evaluation Design

A set of experiments are included in the paper to assess the performance of the proposed RD-LDA method. To show

the high complexity of the face patterns' distribution, a middle-size subset of the FERET database [9] is used in the experiments. The subset consists of 606 gray-scale images of 49 people, each one having more than 10 samples. These images cover a wide range of variations in illumination, facial expression/details, acquisition time, races and others. We follow the preprocessing sequence recommended in [9], which includes four steps: (1) images are translated, rotated and scaled so that the centers of the eyes are placed on specific pixels, (2) a standard mask is applied to remove the nonface portions, (3) histogram equalization is performed in the non masked facial pixels, (4) face data are further normalized to have zero mean and unit standard deviation. Fig.1 depicts some sample images after the preprocessing sequence is applied. For computational convenience, each image is finally represented as a column vector of length $D = 17154$ prior to the recognition stage.

The number of available training samples per subject, L, has a significant influence on the plug-in covariance matrix estimates used in all these discriminant analysis methods. To study the sensitivity of the performance, in terms of correct recognition rate (CRR), to L, 6 tests were performed with various L values ranging from $L = 2$ to $L = 7$. For a particular L, the FERET subset is randomly partitioned into two datasets: a training set and a test set. The training set is composed of $(L \times 49)$ samples: L images per person were randomly chosen. The remaining $(606 - L \times 49)$ images are used to form the test set. There is no overlapping between the two. To enhance the accuracy of the assessment, 5 runs of such a partition were executed, and all of the CRRs reported later have been averaged over the 5 runs.

Fig. 1. Some samples of six people from the normalized FERET subset.

3.2. The FR Performance Comparison

Besides RD-LDA and its special cases summarized in Table 1, the most well-known FR algorithm, the so-called Eigenfaces method [10], was also implemented to provide a performance baseline. The testing grid of (λ, γ) values was defined by the outer product of $\lambda = [1\mathbf{e}-4 : 0.01 : 1]$ and $\gamma = [1\mathbf{e}-4 : 0.01 : 1]$, where both of λ and γ started from $1e-4$ instead of zero in case $\mathbf{S}_i$ is singular. The CRRs obtained by RD-LDA in the grid are depicted in Fig.2. Since most peaks or valleys occur around the four corners, four 2D side faces of Fig.2 (only four representative cases $L = 2, 3, 4, 6$ are selected) are shown in Figs.3-4 for a clearer view. Also, a quantitative comparison of the best CRRs obtained by Eigenfaces, those methods depicted in Table 1, and RD-LDA with corresponding parameters, is summarized in Table 2.

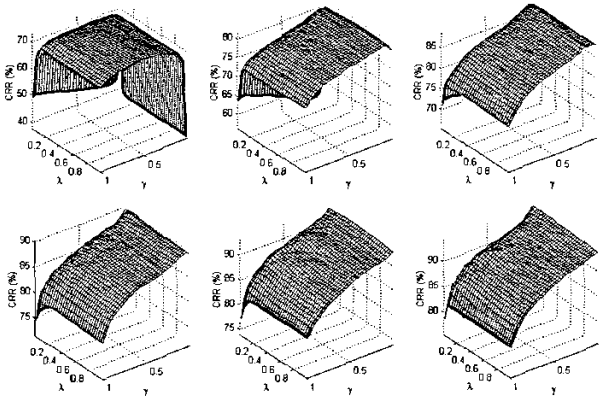

Fig. 2. CRRs obtained by RD-LDA as functions of (λ, γ). **Top**: $L = 2, 3, 4$; **Bottom**: $L = 5, 6, 7$.

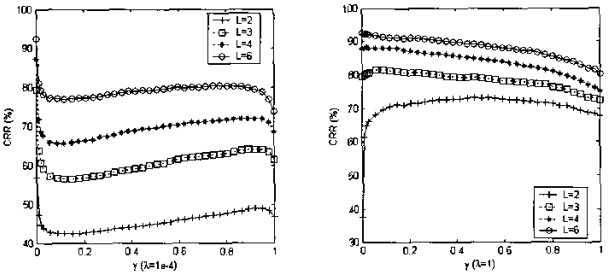

Fig. 3. CRRs as a function of γ with fixed λ.

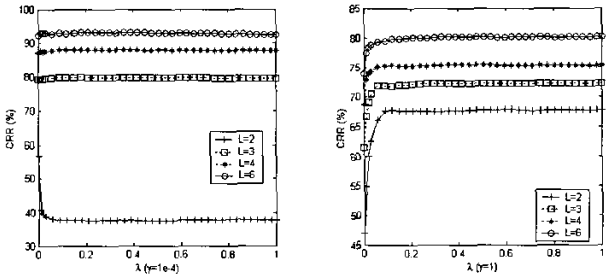

Fig. 4. CRRs as a function of λ with fixed γ.

The parameter λ controls the degree of shrinkage of the individual class covariance matrix estimates $\mathbf{S}_i$ toward the within-class scatter matrix of the whole training set $(\mathcal{H}^T \mathbf{S}_{WTH} \mathcal{H})$. Varying the values of λ within $[0, 1]$ leads

Table 2. Comparison of best CRRs (%).

L	2	3	4	5	6	7
PCA	59.8	67.8	73.0	75.8	81.3	83.7
D-NC	67.8	72.3	75.3	77.3	80.2	80.5
D-WNC	46.9	61.7	68.7	72.1	73.9	75.6
D-QDA	57.0	79.3	87.2	89.2	92.4	93.8
YD-LDA	37.8	79.5	87.8	89.5	92.4	93.5
JD-LDA	70.7	77.4	82.8	85.7	88.1	89.4
(γ)	0.84	0.75	0.69	0.65	0.61	0.59
RD-LDA	73.2	81.6	88.5	90.4	93.2	94.4
(λ)	0.93	0.93	0.35	0.11	0.26	0.07
(γ)	0.47	0.10	0.07	0.01	1e-4	1e-4

to a set of intermediate classifiers between D-QDA and YD-LDA. In theory, D-QDA should be the best performer among the methods evaluated here if sufficient training samples are available. It can be observed at this point from Fig.2 and Table 2 that the CRR peaks gradually moved toward the corner $(0, 0)$ that is the case of D-QDA from the central area as L increases. Small values of λ have been good enough for the regularization requirement in many cases ($L \geq 3$) as shown in Fig.4:Left.

However, it is also can be seen from Fig.3:Right and Table 2 that both of D-QDA and D-LDA performed poorly when $L = 2$. This should be attributed to the high variance in estimates of $\mathbf{S}_i$ and $\mathbf{S}$ due to insufficient training samples. In these cases, $\mathbf{S}_i$ and even $\mathbf{S}$ are singular or close to singular, and the resulting effect is to dramatically exaggerate the importance associated with the eigenvectors corresponding to the smallest eigenvalues. Against the effect, the introduction of another parameter γ helps to decrease the larger eigenvalues and increase the smaller ones, thereby counteracting for some extent the bias. This is also why JD-LDA outperforms YD-LDA when L is small. Although JD-LDA seems to be a little over-regularized compared with the optimal (λ^*, γ^*), the method almost guarantees a stable sub-optimal solution, 4.5% CRR difference in average over $L = 2 - 7$ from the best one found by RD-LDA. Therefore, JD-LDA could be the first choice when insufficient prior information about the training samples is available and a cost effective processing solution is sought. Although RD-LDA is the top performer amongst all methods compared here, the determination of its optimal parameter values is computationally demanding as it is based on exhaustive searches. A fast and cost effective RD-LDA parameter optimization method will be the focus of future research.

4. CONCLUSION

A new method for face recognition has been introduced in this paper. The proposed method combines the D-LDA technique with regularization strategies to effectively address the SSS problem commonly encountered in FR tasks. The D-LDA technique is utilized to map the original face patterns to a low-dimensional discriminant feature space, where the regularization strategy becomes applicable. The regularization strategy provides a balance between the variance and the bias in sample-based estimates addressing the SSS problem. It also has been shown that a series of traditional discriminant analysis methods including the recently introduced YD-LDA and JD-LDA can be derived from the proposed RD-LDA framework by adjusting the regularization parameters. Experimental results indicate that the RD-LDA method outperforms the commonly used Eigenfaces method as well as other discriminant analysis approaches across various SSS settings.

RD-LDA can be seen as a general pattern recognition method capable to address with nonlinear and SSS problems. We expect that in addition to FR, RD-LDA will provide excellent performance in applications, such as image/video indexing, retrieval, and classification.

5. REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, 1997.

[2] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu, "A new LDA-based face recognition system which can solve the small sample size problem," Pattern Recognition, vol. 33, pp. 1713–1726, 2000.

[3] Hua Yu and Jie Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," Pattern Recognition, vol. 34, pp. 2067–2070, 2001.

[4] Juwei Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using LDA based algorithms," IEEE Transactions on Neural Networks, vol. 14, no. 1, January 2003.

[5] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology," CVGIP: Image Understanding, vol. 59, pp. 254–261, 1994.

[6] Juwei Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," IEEE Transactions on Neural Networks, vol. 14, no. 1, January 2003.

[7] Jerome H. Friedman, "Regularized discriminant analysis," Journal of the American Statistical Association, vol. 84, pp. 165–175, 1989.

[8] R.A. Fisher, "The use of multiple measures in taxonomic problems," Ann. Eugenics, vol. 7, pp. 179–188, 1936.

[9] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1090–1104, 2000.

[10] Matthew A. Turk and Alex P. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86, 1991.