

Face Recognition by Regularized Discriminant Analysis

Dao-Qing Dai and Pong C. Yuen

Abstract—When the feature dimension is larger than the number of samples the small sample-size problem occurs. There is great concern about it within the face recognition community. We point out that optimizing the Fisher index in linear discriminant analysis does not necessarily give the best performance for a face recognition system. We propose a new regularization scheme. The proposed method is evaluated using the Olivetti Research Laboratory database, the Yale database, and the Feret database.

Index Terms—Face recognition, optimization, regularized discriminant analysis (RDA), small sample-size problem.

I. INTRODUCTION

Identification of persons with automatic computer interfaces has aroused increasing interest in the computer science community in recent years [2], [16], [44]. Linear discriminant analysis (LDA) [8], [9], [30] is a well-known and popular statistical method in pattern recognition and classification. Its basic idea is to optimize the Fisher discriminant index \mathcal{F} [8], [9], [30] defined by

$$\mathcal{F} = \max_W \text{tr}((W^T C_w W)^{-1} (W^T C_b W))$$

where $\text{tr}(X)$ is the trace of the matrix X , i.e., the sum of diagonal elements of X . C_b is the between-class scatter matrix, and C_w is the pooled within-class scatter matrix. The optimal Fisher transform is determined from eigenvectors of the matrix $C_w^{-1} C_b$.

Many algorithms based on LDA have been employed in face recognition technology. However, it suffers from a well-known small sample-size problem [11], [20], [24], [28], [29], [34], [36], [38], i.e., the number of samples is small compared with the size of the feature vector. To avoid this difficulty, techniques are employed to reduce the dimension from d to d' , where $d' < d$, so that in $\mathbb{R}^{d'}$ the within-class scatter matrix is not singular. The FisherFace [1], [19], the most discriminant features [32], QR-decomposition [42], the subspace methods [31], [44] and recursive LDA [39] have been developed.

These approaches are straightforward, but some of them suffer from two limitations. First, the FisherFace method might fail [45]. Selections of features are important issues also [26], [37]. Second, feature vectors in the null space of C_w can still have discriminative power as shown in [40].

Another direction is to modify the optimization criteria. The Fisher index is a combination of two measures C_w and C_b . We need to maxi-

Manuscript received December 9, 2005; revised June 14, 2006. This work was supported in part by the National Science Foundation (NSF) of China under Grants 60175031, 60575004, and 10231040, by the NSF of Guangdong under Grant 05101817, by the Ministry of Education of China under Grant NCET-04-0791, and by the Research Grant Council (RGC) Earmarked Research Grant HKBU-2119/03E. This paper was recommended by Associate Editor N. K. Ratha.

D.-Q. Dai is with the Center for Computer Vision and Department of Mathematics, Faculty of Mathematics and Computing, Sun Yat-Sen (Zhongshan) University, Guangzhou 510275, China (e-mail: stsddq@mail.sysu.edu.cn).

P. C. Yuen is with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: pcyuen@comp.hkbu.edu.hk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSMCB.2007.895363

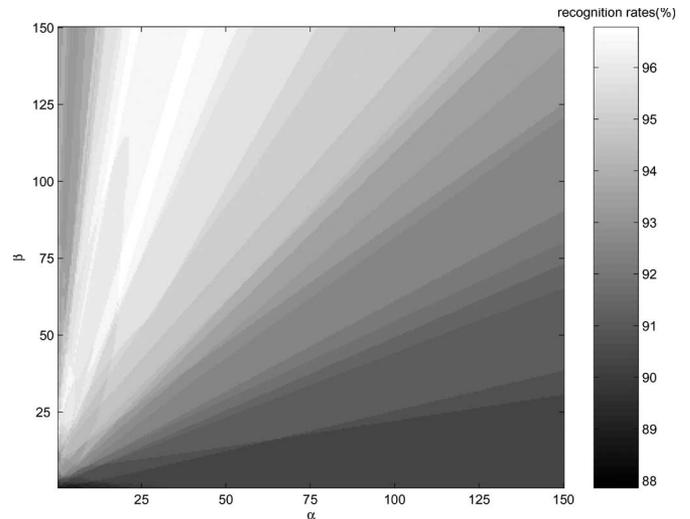


Fig. 1. Recognition rates via the parameters α, β ranging from 0.5 to 150. Larger recognition rates are not on the α -axis (small β).

mize the between-class scatter matrix and to minimize the within-class scatter matrix. Substitutions of these two scatters are very natural [42]. Liu *et al.* [23] used the total scatter matrix C_t to replace the within-class covariance C_w . The rank of the matrix C_t is, in general, greater than that of the matrix C_w . But it can still be singular. Hence, they sought transform matrix W such that $\text{tr}(W^T C_b W) \neq 0$ under the constraint $\text{tr}(W^T C_t W) = 0$. Solutions of this optimization problem are not unique in general. Chen *et al.* [3] and Belhumeur *et al.* [1] further proposed to maximize $\text{tr}(W^T C_b W)$ in the null space of C_w . The final transformation matrix will lead the Fisher index to be infinite. Yu and Yang [43] pointed out that the vectors outside the null space still have discriminative power, and therefore they proposed direct LDA. Lu *et al.* [25] applied regularized LDA in the range of C_b ; see also [40].

We find out that when small sample-size problem occurs, optimizing the Fisher index \mathcal{F} does not necessarily lead to the best system performance.

In this correspondence, motivated by the above limitations we propose to use a regularized discriminate scheme. Our contribution consists of two parts.

- 1) Maximizing the Fisher index is not always good.
- 2) Use of regularization instead of optimizing the Fisher index.

We report the experiment results to justify the effectiveness of the proposed algorithm in Section III.

II. REGULARIZED DISCRIMINANT ANALYSIS (RDA)

The regularization method was originally proposed by Tikhonov to solve the ill-posed operator equation; its application in machine learning is addressed in [35]. We employ the idea of regularization [5], [25], [28] but not restrict ourselves to small perturbation. Along this line, we propose a two-parameter regularization scheme which will be reported in Section II-B. Moreover, from Fig. 1, we find that when the small sample-size problem occurs, maximizing the Fisher index does not give the best performance. Therefore, we propose a new optimization criterion based on posterior error rate; moreover we introduce the method of robust cross-validation (RCV) to solve the optimal problem. Details are discussed in Section II-D.

A. Deficiency of the Fisherface Method

When applying the principal component analysis (PCA) plus LDA [1] approach the following remarks should be considered.

- 1) LDA can still fail even after a PCA procedure. For the PCA projected data we get the matrix C'_w , C'_b , and C'_t , where $C'_t = C'_b + C'_w$ is the total scatter matrix. Then, there might exist a direction α such that $\alpha^T C'_t \alpha = \alpha^T C'_b \alpha$ so that $\alpha^T C'_w \alpha = 0$. Hence, the matrix C'_w is still singular.
- 2) The null space of the within-class scatter matrix C_w contains discriminative information for classification. For a projection direct β , if $\beta^T C_w \beta = 0$ and $\beta^T C_b \beta \neq 0$, obviously, the Fisher index is maximized [40].

These are illustrated by the following calculations. Suppose that the three scatter matrices are decomposed as

$$C'_t = C'_b + C'_w$$

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}_{d' \times d'} = \begin{pmatrix} \lambda_1^b & & & \\ & \lambda_2^b & & \\ & & \ddots & \\ & & & \lambda_{d'}^b \end{pmatrix} + \begin{pmatrix} \lambda_1^w & & & \\ & \lambda_2^w & & \\ & & \ddots & \\ & & & \lambda_{d'}^w \end{pmatrix}.$$

This is achieved by first diagonalizing the matrix C_t , in its range choosing d' ($d' \leq \text{rank}(C_t)$) eigenvectors for projection and normalizing to the $d' \times d'$ identity matrix. Then, diagonalizing the reduced within-class matrix to $C'_w = \text{diag}\{\lambda_1^w, \lambda_2^w, \dots, \lambda_{d'}^w\}$. Finally, the between-class matrix is diagonal also since $C'_b = C'_t - C'_w$ is the difference of two diagonal matrices.

We always have the identities

$$\lambda_i^b + \lambda_i^w = 1, \quad i = 1, 2, \dots, d'.$$

Without loss of generality, we suppose that

$$\lambda_1^b \geq \lambda_2^b \geq \dots \geq \lambda_\kappa^b > \lambda_{\kappa+1}^b = \dots = \lambda_{d'}^b = 0$$

for some $1 \leq \kappa < d'$, i.e., λ_i^b ($i = 1, 2, \dots, \kappa$) are eigenvalues corresponding to eigenvectors in the range $\mathcal{R}(C'_b)$. For λ_i^w ($i = 1, 2, \dots, \kappa$) there is no guarantee that it is not zero except that the null space of C'_w has empty intersection with $\mathcal{R}(C'_b)$; hence the quotient λ_i^b/λ_i^w is not defined and LDA cannot be applied.

B. Estimation of Covariance Matrix

When the small sample problem occurs, the covariance matrix C_w is singular, and it is required to provide an estimate for its inverse [28], [34]. Suppose that we have the singular value decomposition $C_w = \sum_{k=1}^d \lambda_k v_k v_k^T$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_g > \lambda_{g+1} = \dots = \lambda_d = 0$ are the eigenvalues of C_w and v_k ($k = 1, 2, \dots, d$) are the corresponding normalized eigenvectors; g is the rank of C_w , $1 \leq g < d$ because C_w is singular. To formulate the regularization scheme, there are two possible strategies:

- 1) regulate on eigenvalues $\{\lambda_k\}$;
- 2) regulate on orientation $\{v_k\}$.

From matrix theory, $\{\lambda_k\}$ depends on C_w continuously, but $\{v_k\}$ is very sensitive to the variation of C_w . Therefore, we are going to regulate the eigenvalues and propose to use a two-parameter family regularized inverse R of C_w with parameters $\alpha, \beta > 0$ by

$$R = R(\alpha, \beta) = \sum_{k=1}^g \frac{1}{\lambda_k + \alpha} v_k v_k^T + \frac{1}{\beta} \sum_{k=g+1}^d v_k v_k^T \quad (1)$$

where α controls the nonzero λ_k while β controls the zero λ_k .

The inverse of R is therefore given by $R^{-1} = \sum_{k=1}^g (\lambda_k + \alpha) v_k v_k^T + \beta \sum_{k=g+1}^d v_k v_k^T$ and satisfies the following.

- 1) R^{-1} is an approximation to the matrix C_w as α and β approach zero.
- 2) R^{-1} is symmetric.
- 3) R^{-1} keeps the first g principal eigenvectors the same as those of C_w , and they are in the same order. The last $(d - g)$ eigenvectors are subject only to orthogonality constraints, i.e., the corresponding eigenvalues are identical. If sufficient samples are given, all the eigenvalues should not be equal to zero. Because only limited samples are available, the last $d - g$ eigenvalues $\lambda_{g+1}, \dots, \lambda_d$ are estimated as zero.

Hence, instead of seeking eigenvectors of the matrix $C_w^{-1} C_b$, we solve the eigenproblem

$$(RC_b)W_{\text{rda}} = W_{\text{rda}}D_{\text{rda}} \quad (2)$$

where $D_{\text{rda}} = D_{\text{rda}}(\alpha, \beta)$ is a diagonal matrix with nonnegative entries, $W_{\text{rda}} = W_{\text{rda}}(\alpha, \beta)$ is the regularized Fisher transform. In [4], a one-parameter scheme with kernel is developed, as we can see from Fig. 1 that the optimal domain cannot be described by one parameter. In [28], a perturbation of quadratic discriminant analysis is developed.

C. Special Model: Infinity Fisher Index (Inf-F)

From the definition of the Fisher index, one needs to find an optimal solution W such that $C^T C_b W$ is as large as possible and meanwhile keeps $W^T C_w W$ as small as possible. When the matrix C_w is singular and the matrix W is sought in the null space of C_w the Fisher index will be maximum and reach infinity. It leads to the optimization problem

$$\arg \max_{W \in \mathcal{N}(C_w)} \text{tr}(W^T C_b W).$$

This scheme ignores the within-class variation information, and the data tend to be oversmoothed. Moreover, it should be noted that when the small sample-size problem occurs, **Inf-F** index can always be obtained in a null-space approach. In [3], [10], and [23], this idea was used.

D. Posterior Error Rate and the Optimal Parameters

The purpose of this section is to estimate the two parameters α and β . We proceed in the following steps.

- 1) Large Fisher index does not necessarily correspond to bigger recognition rate (Fig. 1).
- 2) Optimizing of the posterior error rate instead of the Fisher index.
- 3) Use of the RCV method to solve the optimization problem.

In order to find out the effect of parameters α and β for recognition, we carry out an experiment with the Olivetti Research Laboratory (ORL) database; please refer to Section III for detailed experiment setting. The recognition performance with different values of α and β are recorded and plotted in Fig. 1. The gray values represent the recognition rate of one experiment.

From Fig. 1, we first examine two special parameters, and the corresponding recognition rates are not high.

- 1) Tikhonov regularization (T-Reg) [44], the parameter (α, β) is around the origin. The regularization matrix R is

$$R = (C_w + \epsilon I_{d \times d})^{-1} = \sum_{k=1}^g \frac{1}{\lambda_k + \epsilon} u_k u_k^T + \frac{1}{\epsilon} \sum_{k=g+1}^d u_k u_k^T.$$

- 2) Pseudoinverse [33], the parameter (α, β) is on the β -axis. The regularization matrix is $R = \sum_{k=1}^g (1/\lambda_k) u_k u_k^T$. This scheme discards completely the information in the null space of the covariance matrix C_w , which are estimated as zero because of insufficient data.

Moreover, it can be seen that the recognition accuracy with zero β value [i.e., on the α -axis (small β)] does not give the best result, and this corresponds to the **Inf-F** model. Fig. 1 also suggests the following.

- 1) For approximately the same recognition rate, its parameter space is not a single point; it forms a subspace. Any regularization parameter sets in the optimal subspace are good enough for classification.
- 2) Maximizing the Fisher index does not necessarily lead to better performance. Our experiments show that the Inf-F solution **Inf-F** does not have superior performance.
- 3) Optimal recognition rate is not on the two axes, which corresponding, respectively, the null space and range space of C_w . Better results are obtained through nontrivial parameter, that is using feature combination both in the null space and range space. In [40], a different combination scheme was used.

To get an alternate optimization criteria, we thus propose to use the posterior error rate as a criterion instead. Minimizing the error rate is equivalent to maximizing the recognition rate, which is determined by the parameters α, β . Therefore, the recognition rate r is a function of these parameters, $r = r(\alpha, \beta)$. In turn, the optimization problem is formulated as follows:

$$[\alpha_{\text{opt}}, \beta_{\text{opt}}] = \arg \max_{\alpha, \beta} r(\alpha, \beta). \quad (3)$$

To get the optimal parameter $\alpha_{\text{opt}}, \beta_{\text{opt}}$, we propose to use a method based on cross-validation.

1) *Robust Cross-Validation (RCV)*: The recognition rate $r(\alpha, \beta)$ depends on the training set and the probe set also. To increase the stability of the system, we introduce the method of RCV.

Suppose that image set ω is available; this implies that each class has at least three images; two are used for training, the rest for probing. Given an integer N , we choose randomly image subset ω_t from ω so that each class of images in ω_t has at least two images. The set ω_t is used as the training set, and the rest $\omega \setminus \omega_t$ is the probe set. This process is carried out N times as follows.

For fixed integers I, J , we use the grids (α_i, β_j) , $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$, where I and J stand for the number of grids in two directions to be used. For the k th experiment, the array $D_k(i, j)$



Fig. 2. Example images of two subjects (the first row) and the cropped images (the second row) with the Feret database.

records the recognition rate at grid α_i, β_j . The pseudocode reads as follows.

```

for i = 1, 2, ..., I
  for j = 1, 2, ..., J
    for k = 1, 2, ..., N
      Dk(i, j) = r(αi, βj), for each ωt
    end
  end
end

```

The maximal $[\alpha_{\text{opt}}, \beta_{\text{opt}}]$ is determined by

$$[\alpha_{\text{opt}}, \beta_{\text{opt}}] = \arg \max_{i, j} \sum_{k=1}^N D_k(i, j) / N.$$

The computational complexity for searching α_{opt} and β_{opt} is in proportion to $O(IJNd^3)$, which can be further reduced by using a coarse to fine search strategy. This work load is for the training stage. For the recognition stage, for each probe image, it needs to carry out a matrix to vector multiplication, which is about the same as those in the FisherFace method.

III. EXPERIMENT RESULTS

This section reports the evaluation results of the proposed RDA algorithm on face recognition. Three standard publicly available databases, namely ORL, Yale, and Feret, are used for evaluation. The ORL database contains 400 images of 40 subjects taken at ORL in Cambridge University, U.K. The Yale face database contains 165 gray scale images of 15 individuals. This set has considerable variations in facial expressions and illuminations. The Feret database [27] is more challenging. We shall use 432 front-view images of 72 subjects, each having six images. Fig. 2 shows images of two subjects. The images are manually aligned according to the positions of eyes and cropped as shown on the second row of Fig. 2. By means of a low-pass filter and row-by-row reshaping, we reduce image size as follows. The sizes are, respectively, 644 (= 23 × 28) in the ORL database and the Feret database and 667 (= 29 × 23) in the Yale database. Hence, the matrix C_w is singular.

Our recognition rate is the rank-one rate. It is estimated by N_c/N_t , where N_t is the number of test images, N_c is the number of correctly recognized test images. For multiple runs, the average rate is defined by $r_{\text{av}} = \sum_{k=1}^m r_k / m$, where m is the number of runs, r_k is the recognition rate of each run. The classifier we shall use is the nearest neighbor rule.

A. Performance of RDA

We use a cross-validation method to estimate the optimal parameters $\alpha_{\text{opt}}, \beta_{\text{opt}}$ with two training images for each subject. The estimated parameters α_{opt} and β_{opt} will be used in the testing phase with three to nine training images. Each test set is randomly run 50 times.

In Tables I–III, we report the rank-one recognition rates based on our RDA algorithm on the three databases. When two images are used,

TABLE I
RANK-ONE RECOGNITION RATES OF 50 RANDOM RUNS WITH THE ORL DATABASE

# of training images	2	3	4	5	6	7	8	9
recognition rate(%)	85.75	92.28	94.87	96.65	97.34	97.73	98.55	98.55

TABLE II
RANK-ONE RECOGNITION RATES OF 50 RANDOM RUNS WITH THE YALE DATABASE

# of training images	2	3	4	5	6	7	8	9	10
recognition rate(%)	78.04	85.92	89.85	92.11	94.53	95.73	96.53	97.20	97.20

TABLE III
RANK-ONE RECOGNITION RATES OF 50 RANDOM RUNS WITH THE FERET DATABASE

# of training images	2	3	4	5
recognition rate(%)	82.4236	88.8611	91.9722	93.4722

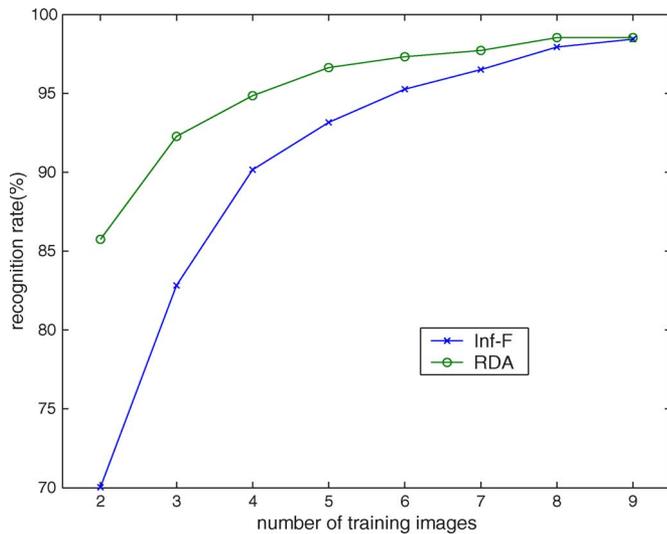


Fig. 3. Recognition rates of 50 random runs with the ORL database.

the average recognition rates are 85.75%, 78.04%, and 82.42% for the ORL, Yale, and Feret databases, respectively. Moreover, it is found that the recognition rates increase when the number of training images increases.

B. Infinite Fisher Index is Not Necessarily Good

Normally, the LDA-based algorithm would like to maximize the Fisher index. When the small sample-size problem occurs, the Inf-F index can be obtained. This intuition idea has been used in [3], [10], and [23]. The objective of this section is to demonstrate that Inf-F is not necessarily a good choice. In Figs. 3–5, we report the performance comparison of RDA with that of Inf-F. All the experiments are repeated 50 times with different training image sets. The rank-one recognition rates with two to eight training images are plotted. Although its Fisher index is infinity, the performance of the Inf-F method is not as good as that of the RDA method. Due to larger image variations in the Yale and Feret database, the difference between the Inf-F and RDA is larger than that in the ORL database. This suggests that when image variations in a database are considerably large, the use of Inf-F is not a good choice.

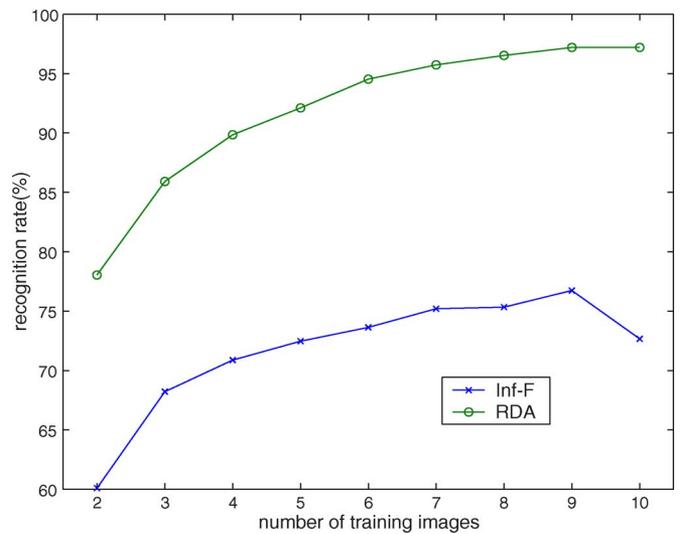


Fig. 4. Recognition rates of 50 random runs with the Yale database.

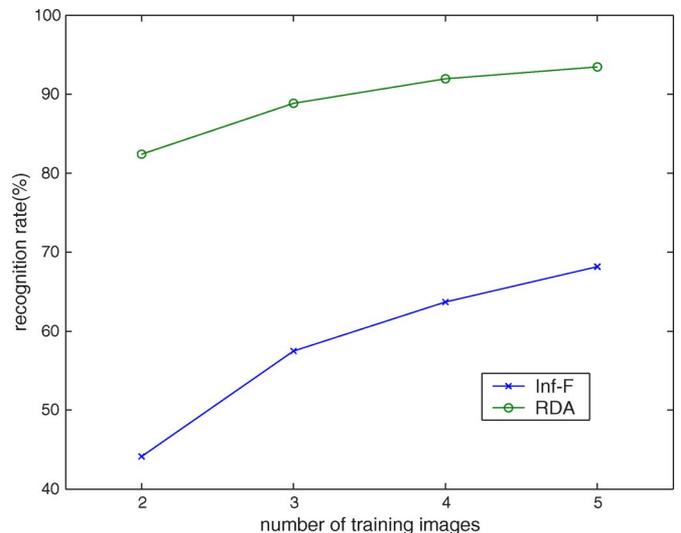


Fig. 5. Recognition rates of 50 random runs with the Feret database.

C. Comparison With Other Methods

In Tables IV and V and Fig. 6, we shall compare the performance of the proposed system with some of the existing methods published in the literature, including FisherFace, kernel-FisherFace, Direct LDA, and ICA. The recognition rates are extracted from the corresponding references directly.

TABLE IV
COMPARISON ON THE PERFORMANCE WITH THE ORL DATABASE

method	training set	recognition rate(%)
CNN[22]	five, 1 run	96.2
HMM+DCT[6]	five, 1 run	100
DCT[14]	five, 1 run	91
Direct LDA[43]	five, ≥ 10 runs	90.8
UDF[17]	five, 1 run	97.5
RBF+NN[7]	five, 6 runs	98.08
FHLA[13]	five, 4 runs	99.55
UODV[18]	two, 1 run	81.25
ILDA[19]	two, 1 run	87.19
PQDA[28]	five, na	≤ 90
RDA	two, 50 runs	85.75
	five, 50 runs	96.65

TABLE V
“LEAVE-ONE-OUT” PERFORMANCE WITH THE YALE DATABASE

method	recognition rate(%)
Eigenface (w/o 1 st three eigenvectors)[1]	75.6 (84.7)
Linear Subspace[1]	79.4
FisherFace[1], [12], [41]	93.7
ICA[41]	70.91
ICA-FX[21]	96.36
Kernel-Fisher[41]	93.94
Edge map[12]	74.94
LEM[12]	85.45
RDA	97.6

With the ORL database: The comparison on the ORL database is summarized in Table IV. The selected algorithms/systems used two or five images of one subject as training and the rest images as probe set. The training images are different from one algorithm to another. Also, some algorithms only reported the best result. In our RDA system, we run the program 50 times.

The convolution neural network in [22] achieved recognition rate 96.2% for one run. Using the pseudo-2-D hidden Markov models and discrete cosine transform (HMM+DCT) [6] the recognition rate reached 100% for one run. The recognition rate of the DCT-based system [14] is 91% for one run. The average recognition accuracy for direct LDA [43] for more than ten runs is 90.8%. The uncorrelated discriminant feature [17] with one run is 97.5%. For the radial basis function and neural network (RBF+NN) [7], the average recognition rate of six runs is 98.08%. With a fuzzy hybrid learning algorithm [13], the neural network system reaches 99.55% with four runs. With two training images, the uncorrelated optimal discrimination vectors method is 81.25% and the improved LDA

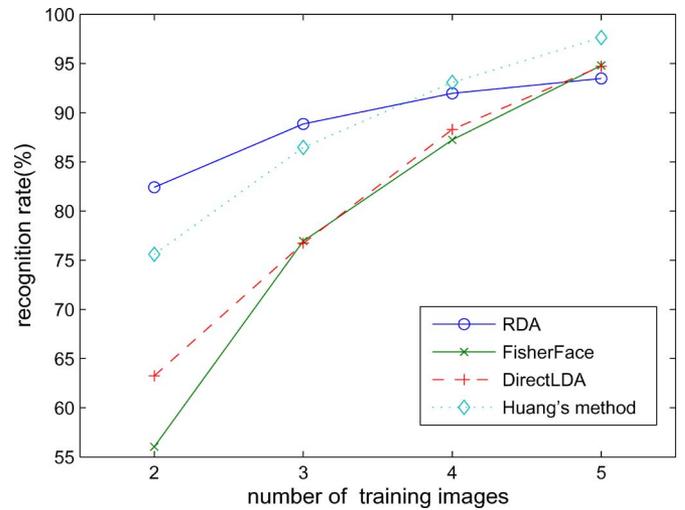


Fig. 6. Comparison of RDA with other methods on the Feret database.

(ILDA) is 87.19% for a single run; our method reaches 85.75% for 50 random runs. The standard deviation is 1.5 and the maximum rate is 100%. For the perturbation-based quadratic discriminant analysis its error rate is larger than 10% when five training images are used.

With the Yale database: Table V summarizes the results on some of the existing methods on the Yale database using the “leave-one-out” strategy. The recognition rate of the Eigenface method is 75.6%. If the first three eigenvectors are removed, it increases to 84.7%. The correlation method and linear subspace method are 76.1% and 79.4%, respectively. The FisherFace method [1], [12], [41], which uses PCA for dimension reduction and then applies LDA, is 93.7%. The kernel-Fisher method [41] is 93.94%. The independent component analysis (ICA) is 70.91%, if a feature extraction process is applied the ICA’s (ICA-FX) [21] rate increases to 96.36%. The edge map method and the line edge map method (LEM) [12] give the recognition rate 74.94% and 85.45%, respectively. Our system RDA is 97.6%.

With the Feret database: Fig. 6 is a comparison of RDA with FisherFace [1], Direct LDA [43], and Huang *et al.*’s method [15]. The data for the last three methods are extracted from [15] directly, where six images of 70 persons were used. When the number of training images are 2 and 3, RDA has the best performance. The overall performance of the four methods is comparable with four and five images.

IV. CONCLUSION

In this correspondence, we have proposed to use a RDA to solve the small sample-size problem and applied it to human face recognition. The algorithm has been evaluated using the ORL database, Yale database, and Feret database. We use the T-Reg method to solve the ill-posed eigenvalue problem and show that even Inf-F does not necessarily lead to optimal system performance.

A drawback of the proposed algorithm is that the computation load of the proposed method is higher than that of the FisherFace method in the training stage, although sometimes offline training is allowed. Therefore, our future direction is to reduce the number of regularized parameters so that the computation load can be reduced.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for valuable suggestions to improve the manuscript.

REFERENCES

- [1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
- [2] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.
- [3] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, 2000.
- [4] W. S. Chen, P. C. Yuen, J. Huang, and D. Q. Dai, "Kernel machine-based one-parameter regularized Fisher discriminant method for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 4, pp. 659–669, Aug. 2005.
- [5] D. Q. Dai and P. C. Yuen, "Regularized discriminant analysis and its applications to face recognition," *Pattern Recognit.*, vol. 36, no. 3, pp. 845–847, Mar. 2003.
- [6] E. Eickeler, S. Müller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image Vis. Comput.*, vol. 18, no. 4, pp. 279–287, Mar. 2000.
- [7] M. J. Er, S. Wu, J. Lu, and H. L. Toh, "Face recognition with radial basis function (RBF) neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 697–710, May 2002.
- [8] R. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.
- [9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
- [10] Z. Q. Hong and J. Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recognit.*, vol. 24, no. 4, pp. 317–324, 1991.
- [11] P. Howland, J. Wang, and H. Park, "Solving the small sample size problem in face recognition using generalized discriminant analysis," *Pattern Recognit.*, vol. 39, no. 2, pp. 277–287, Feb. 2006.
- [12] Y. Gao and M. K. Leung, "Face recognition using line edge map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 764–779, Jun. 2002.
- [13] J. Haddadniaa, K. Faeza, and M. Ahmadib, "A fuzzy hybrid learning algorithm for radial basis function neural network with application in human face recognition," *Pattern Recognit.*, vol. 36, no. 5, pp. 1187–1202, May 2003.
- [14] Z. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *Int. J. Comput. Vis.*, vol. 43, no. 3, pp. 167–188, 2001.
- [15] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," in *Proc. ICPR*, 2002, vol. 3, pp. 29–33.
- [16] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004.
- [17] Z. Jin, J. Y. Yang, Z. S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transform," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, Jul. 2001.
- [18] X. Y. Jing, D. Zhang, and Z. Jin, "UODV: Improved algorithm and generalized theory," *Pattern Recognit.*, vol. 36, no. 11, pp. 2593–2602, 2003.
- [19] X. Y. Jing, D. Zhang, and Y. Y. Tang, "An improved LDA approach," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 5, pp. 1942–1951, Oct. 2004.
- [20] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data," *Appl. Stat.*, vol. 44, no. 1, pp. 101–115, 1995.
- [21] N. Kwak, C. Choi, and N. Ahuja, "Face recognition using feature extraction based on independent component analysis," in *Proc. Int. Conf. Image Process.*, 2002, vol. 2, pp. 337–340.
- [22] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
- [23] K. Liu, Y. Cheng, and J. Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion," *Pattern Recognit.*, vol. 26, no. 6, pp. 903–911, 1993.
- [24] Q. Liu, H. Lu, and S. Ma, "Improving kernel Fisher discriminant analysis for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 42–49, Jan. 2004.
- [25] J. W. 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 Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.
- [26] A. M. Martinez and M. L. Zhu, "Where are linear feature extraction methods applicable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1934–1944, Dec. 2005.
- [27] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
- [28] I. Pima and M. Aladjem, "Regularized discriminant analysis for face recognition," *Pattern Recognit.*, vol. 37, no. 9, pp. 1945–1948, Sep. 2004.
- [29] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.
- [30] A. C. Rencher, *Multivariate Statistical Inference and Applications*. New York: Wiley, 1998.
- [31] J. Ruiz-Del-Solar and P. Navarrete, "Eigenspace-based face recognition: A comparative study of different approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 315–325, Aug. 2005.
- [32] D. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.
- [33] Q. Tian, M. Barbero, Z. H. Gu, and S. H. Lee, "Image classification by the Foley–Sammon transform," *Opt. Eng.*, vol. 25, no. 7, pp. 834–840, 1986.
- [34] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, "A new covariance estimate for Bayesian classifiers in biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 214–223, Feb. 2004.
- [35] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 2000.
- [36] M. Visani, C. Garcia, and J. M. Jolion, "Bilinear discriminant analysis for face recognition," in *Proc. Pattern Recog. and Image Anal.*, 2005, vol. 3687, pp. 247–256, part 2.
- [37] J. Wang, K. N. Plataniotis, and A. N. Venetsanopoulos, "Selecting discriminant eigenfaces for face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 10, pp. 1470–1482, 2005.
- [38] X. G. Wang and X. O. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.
- [39] C. Xiang, X. A. Fan, and T. H. Lee, "Face recognition using recursive Fisher linear discriminant," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2097–2105, Aug. 2006.
- [40] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA Plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.
- [41] M.-H. Yang, "Face recognition using kernel methods," in *Advances of Neural Information Processing Systems*, vol. 14. Cambridge, MA: MIT Press, 2002, pp. 215–220.
- [42] J. P. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.
- [43] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, Oct. 2001.
- [44] W. Zhao, R. Chellappa, and P. J. Phillips *et al.*, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–459, 2003.
- [45] X. S. Zhuang and D. Q. Dai, "Inverse Fisher discriminant criteria for small sample size problem and its application to face recognition," *Pattern Recognit.*, vol. 38, no. 11, pp. 2192–2194, 2005.