Nonsingular Discriminant Feature Extraction for Face Recognition

Chih-Pin Liao and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan 70101, ROC
jtchien@mail.ncku.edu.tw

ABSTRACT

It is popular to extract discriminant features using Fisher linear discriminant analysis (LDA) for general pattern recognition. LDA aims to find an optimal discriminant transformation matrix, which maximizes the ratio of between-class scatter to within-class scatter. However, in case of small sample size and high dimensional data, LDA is prone to be unrealizable due to the singularity of scatter matrices. In this paper, we present a nonsingular transformation prior to performing LDA. This method is to transform general features using *all* eigenvectors of scatter matrix with nonzero eigenvalues. As a result, the scatter matrix of transformed features is *nonsingular*. Subsequently, the discriminant transformation is applied according to LDA using the new scatter matrices. The superiority of nonsingular discriminant analysis of *between-class matrix* comes from the shrinkage of within-class scatters and accordingly the *enhancement of Fisher class separability*. From the experiments on facial databases, we find that the nonsingular discriminant feature extraction achieves significant face recognition performance compared to other LDA-related methods for a wide range of sample sizes and class numbers.

1. INTRODUCTION

There is no doubt that face recognition is an amicable approach for biometrics because the authentication can be completed in hands-free way with no touch and interruption of user activities. In general, face recognition system suffers from the problems of high-dimensional data and small sample size. Owing to high-dimensional data, it is indispensable to reducing the feature dimension and alleviating the computational load. Also, the face recognition performance is degraded by the variations of viewpoint, pose, illumination and expression, which are unseen in small training data. Accordingly, how to develop a discriminant feature extractor using small training samples becomes very challenging for face and general pattern recognition. To reinforce the classification performance, the linear discriminant analysis (LDA) extracts the most discriminant features according to Fisher criterion [4] where the ratios of between-class to within-class scatter matrices are maximized [6]. However, in case of small sample size, the scatter matrix is singular and the LDA has no solution to the generalized eigen-equation. In this paper, we present a new nonsingular discriminant analysis for facial feature extraction. A nonsingular transformation of between-class scatter matrix is merged in LDA so as to improve the Fisher class separability in nonsingular subspace.

LDA or PCA plus LDA algorithm has been attracting many researchers focusing on this topic. Fukunaga and Mantock [7] presented the nonparametric discriminant analysis where the nonparametric scatter matrix was determined using k nearest neighbor technique. The resulting matrix was of full rank. To prevent matrix singularity using high-dimensional data, the Fisherface features were extracted by sequentially performing PCA and LDA [1]. Similarly, LDA could be applied to low-dimensional wavelet features for extraction of discriminant waveletface [2]. In [1], the Fisher class separability was measured to see the significance of visual information in wavelet domain. Also, the LDA was improved by introducing a weighted pairwise Fisher criterion for multiclass pattern classification [11]. To relax the assumption of equal sample covariance in LDA framework, the heteroscedastic discriminant analysis (HDA) was proposed through a maximum likelihood procedure of Gaussian model [8]. Similarly, the geometric LDA was presented by maximizing a geometric mapping function, which was correlated to the minimum classification error [13].

This paper proposes the nonsingular discriminant feature extraction for face recognition under different sample sizes and class numbers. The general concept is motivated from solving the singular problem of scatter matrices for LDA procedure. We transform the face data using the eigenvectors of scatter matrix corresponding to all nonzero eigenvalues. This step assures that the scatter matrix of transformed data is nonsingular and positive definite. Even if the original scatter matrix is nonsingular, this transformation is able to optimize the between-class scatter, which is beneficial for pattern classification. At the second step, the Fisher criterion using the new nonsingular scatter matrices is maximized so as to improve the class discriminability. In this study, we systematically investigate different PCA plus LDA algorithms [10][14] by comparing their theoretical relations and experimental results. The experiments on face recognition show good performance of using proposed twofold linear transformation compared to other PCA plus LDA algorithms.

2. LINEAR DISCRIMINANT ANALYSIS

The purpose of LDA procedure is to calculate a transformation matrix A, which transforms the original feature vector $\mathbf{x} \in \Re^n$ into a reduced feature vector $\mathbf{z} \in \Re^p$, $p < n$, as $\mathbf{z} = A^T \mathbf{x}$. The optimal matrix is estimated by maximizing the Fisher's class separability criterion $F(A)$ defined by the ratio of the traces of between-class scatter matrix to within-class scatter matrix using the transformed features

$$F(A) = \frac{\mathrm{tr}(A^T S_b A)}{\mathrm{tr}(A^T S_w A)}, \tag{1}$$

where S_b and S_w are between-class and within-class scatter matrices, respectively [6]. The matrices S_b and S_w are positive semi-definite matrices. Then, the columns of optimal transformation matrix $A = \{\mathbf{a}_i\}$ correspond to the generalized eigenvectors for p leading eigenvalues in $S_b \mathbf{a}_i = \lambda_i S_w \mathbf{a}_i$. If S_w is nonsingular, this can be converted to a conventional eigenvalue

problem of $S_w^{-1}S_b$. Foley and Sammon [5] maximized the Fisher ratio with the constraint for deriving the orthonormal discriminant vectors. Also, Fukunaga [6] presented several variants of Fisher's criterion using scatter matrices S_b, S_w and S_t. In this paper, we concern the inverse criterion

$$F^{-1}(A) = \frac{\mathrm{tr}(A^T S_w A)}{\mathrm{tr}(A^T S_b A)}. \quad (2)$$

With this variant, the discriminant vectors are derived via minimizing $F^{-1}(A)$ or correspondingly finding the eigenvectors of $S_b^{-1}S_w$. But, in real world, the training face images of each person are usually insufficient. When the number of training samples is small or the feature dimension is high, LDA will become unrealizable because of the property $rank(S_w) = \min(n, C\times(M-1))$ and $rank(S_b) = \min(n, C-1)$ [12] where C is class numbers and M represents the sample numbers of each class. We are unable to realize the calculation for the eigenvectors for $S_w^{-1}S_b$ or $S_b^{-1}S_w$. To avoid total sample size smaller than feature dimension, we can reduce the feature dimension prior to performing LDA.

3. NONSINGULAR DISCRIMINANT ANALYSIS

In this paper, a nonsingular transformation is applied to between-class scatter matrix S_b. We are minimizing the inverse Fisher criterion of (2). It is necessary to deal with the eigenvalues of $S_b^{-1}S_w$ or the singular problem of S_b.

3.1 Nonsingular Transformation

The nonsingular transformation of S_b aims to optimize the trace of between-class scatter matrix through

$$A_b = \arg\max_A \mathrm{tr}(A^T S_b A). \quad (3)$$

The optimal solution A_b satisfies the diagonalization process $A_b^T S_b A_b = D_b$ with $n\times n$ diagonal matrix D_b. In case of singular S_b, there existing zero eigenvalues. We may construct a $n\times m$ nonsingular transformation matrix $\hat{A}_b$ consisted of m eigenvectors with nonzero eigenvalues of S_b. Applying the nonsingular transformation $\mathbf{y}_b = \hat{A}_b^T\mathbf{x} \in \Re^m$, we can maximize the distance between the class mean $\overline{\mathbf{y}}^c$ and the total mean $\overline{\mathbf{y}}$. *This discriminant process is significantly beneficial for pattern classification.* The resulting between-class scatter matrix $\hat{S}_b = \hat{A}_b^T S_b \hat{A}_b = \hat{D}_b$ is a $m\times m$ diagonal and nonsingular matrix. The $m\times m$ diagonal matrix $\hat{D}_b$ is related to $n\times n$ diagonal matrix D_b by

$$D_b = \left[\begin{array}{c|c} \hat{D}_b & 0 \\ \hline 0 & 0 \end{array}\right]. \quad (4)$$

Notably, the transformed $\hat{S}_b$ is *positive definite*, *nonsingular* and *invertible*. At the same time, the new within-class scatter matrix due to $\mathbf{y}_b = \hat{A}_b^T\mathbf{x}$ is derived by $\hat{S}_w = \hat{A}_b^T S_w \hat{A}_b$, which is not diagonal. With the new $\hat{S}_w$ and nonsingular $\hat{S}_b$, we may fulfill LDA for the transformed features $\mathbf{y}_b$. This resolves the small sample size problem of LDA. Besides, for the case of nonsingular S_b, it is easy to find that the conditions $m = n$, $\hat{A}_b = A_b$ and $\hat{D}_b = D_b = \hat{S}_b$ hold. Here, we would like to highlight that the eigen-analysis of between-class scatter matrix is different from the eigenface method [1] applied for the total scatter matrix S_t. In this study, we are interested in evaluating the properties of Fisher class separability $F(\hat{A}_b)$ when applying the nonsingular transformation matrix $\hat{A}_b$. The following two theorems are illustrated for the cases that the original between-class scatter matrix S_b is nonsingular as well as singular.

Theorem 1 Assuming that the original between-class scatter matrix S_b is ***nonsingular***, we transform the original features $\mathbf{x}$ using $\mathbf{y}_b = \hat{A}_b^T\mathbf{x}$ where $\hat{A}_b = A_b$. The *between-class* and *within-class scatters* of the transformed features $\mathbf{y}_b$ are the *same* as those of original features $\mathbf{x}$. Equivalently, the *Fisher class separability* is *invariant* under the transformation.

Theorem 2 Assuming that the original between-class scatter matrix S_b is *singular* and has rank of m, we perform the transformation $\mathbf{y}_b = \hat{A}_b^T\mathbf{x}$ where $n\times m$ matrix $\hat{A}_b$ is consisted of m eigenvectors of S_b corresponding to nonzero eigenvalues. The new between-class scatter matrix $\hat{S}_b$ is nonsingular. After the transformation, the between-class scatter is *unchanged* and the within-class scatter is *shrunk*. Equivalently, the *Fisher class separability* is *enlarged*.

For the proof of Theorem 2, we explain the shrinkage of within-class scatter matrix as follows

$$\mathrm{tr}(\hat{S}_w) = \mathrm{tr}\left[\sum_{c=1}^{C}\sum_{k=1}^{M}(\hat{A}_b^T\mathbf{x}_k^c - \hat{A}_b^T\overline{\mathbf{x}}^c)(\hat{A}_b^T\mathbf{x}_k^c - \hat{A}_b^T\overline{\mathbf{x}}^c)^T\right]$$
$$= \sum_{c=1}^{C}\sum_{k=1}^{M}\left\|\hat{A}_b^T(\mathbf{x}_k^c - \overline{\mathbf{x}}^c)\right\|^2 \le \sum_{c=1}^{C}\sum_{k=1}^{M}\left\|\mathbf{x}_k^c - \overline{\mathbf{x}}^c\right\|^2 = \mathrm{tr}(S_w). \quad (5)$$

Then, the Fisher class separability is enlarged because

$$F(\hat{A}_b) = \frac{\mathrm{tr}(\hat{A}_b^T S_b \hat{A}_b)}{\mathrm{tr}(\hat{A}_b^T S_w \hat{A}_b)} = \frac{\mathrm{tr}(\hat{S}_b)}{\mathrm{tr}(\hat{S}_w)} \ge \frac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w)}. \quad (6)$$

3.2 Discriminant Transformation

After the original features $\mathbf{x} \in \Re^n$ are transformed to $\mathbf{y} \in \Re^m$, the second step of proposed nonsingular discriminant transformation (NDT) is to fulfill the standard LDA and transform $\mathbf{y}$ to $\mathbf{z} \in \Re^p$, $p \le m \le n$ using the transformed within-class scatter and between-class scatter matrices. Namely, we will estimate the transformation matrix $\hat{A}_d$ by minimizing the inverse

Fisher criterion where the new scatter matrices $\hat{S}_w$ and $\hat{S}_b$ are considered

$$\hat{A}_d = \arg\min_A \frac{\mathrm{tr}(A^T \hat{S}_w A)}{\mathrm{tr}(A^T \hat{S}_b A)}. \quad (7)$$

Similar to the general LDA procedure, the discriminant transformation matrix $\hat{A}_d$ is established using the eigenvectors of $\hat{S}_w \hat{S}_b^{-1}$. Because $\hat{S}_b$ is nonsingular after nonsingular transformation, $\hat{A}_d$ always exists. Overall, the dual transformation for NDT feature extraction is formed by

$$\mathbf{z}_b = \hat{A}_d^T \mathbf{y}_b = \hat{A}_d^T \hat{A}_b^T \mathbf{x}. \quad (8)$$

4. EXPERIMENTS

4.1 Experimental Setup

The proposed NDT feature extraction was evaluated using two popular face databases: ORL (http://www.cam-orl.co.uk/facedatabase.html) and FERET [14] provided by AT&T Lab Cambridge and NIST, respectively. Some examples of two persons selected from two databases are shown in Figure 1. ORL database contained ten different images of 92×112 pixels for each of 40 distinct persons ($C = 40$). These pictures were taken at different time, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). Also, we sampled ten different images of 256 ×384 pixels for each of 150 persons ($C = 150$) from FERET database. These images were collected in similar illumination condition but varying expressions and pose angles. We rescaled the images to 92×104 pixels. As displayed in Figure 1, we applied three-level wavelet decomposition [2] and reduced the size of all images to 12×13, i.e. $\mathbf{x} \in \Re^{156}$, before executing different LDA related methods. For each subject of ORL and FERET, we randomly selected M images as prototypes and the remaining $10 - M$ images as queries. All recognition rates were reported by employing 10-fold cross-validation. During recognition, the nearest feature line classifier [9] was applied. In this study, enhanced Fisher linear discriminant model (EFM) [10], direct LDA (D-LDA) 0 and proposed NDT were implemented. To investigate the cases of singular and nonsingular scatter matrices, we chose several class numbers C and training sample numbers per class M for face recognition. For objective comparison, we fixed the dimension of the extracted features $\mathbf{z} \in \Re^p$ to be $p = C - 1$ when using different methods. This dimension was considered because D-LDA and NDT dealt with eigen-analysis of S_b having the property $rank(S_b) = \min(n, C-1)$.

4.2 Evaluation of Different Methods on ORL Database

We compare the face recognition rates of D-LDA, EFM and proposed NDT methods for different class numbers C and training sample numbers M using ORL database. The baseline LDA achieves the recognition rate of 86% for the case of $C = 40$ and $M = 5$. The other cases of C and M are unrealizable because of singular S_w. Basically, the feature extraction using EFM is fulfilled via performing whitening transformation of S_w prior to LDA procedure. Similarly, D-LDA extracts facial features by performing whitening transformation of S_b and then diagonalization of the transformed within-class scatter matrix $\tilde{S}_w$. In Figure 2, we evaluate the recognition rates of D-LDA, EFM and NDT for different C and M. We can see that the recognition rates are decreased for smaller M and larger C. For different C and M, NDT achieves higher recognition rates than EFM and D-LDA. EFM is better than D-LDA for $M = 2$ but worse than D-LDA for $M = 3$. For example, NDT obtains the recognition rate of 92%, which is higher than 83.8% of EFM and 89.9% of D-LDA in the case of $C = 30$, $M = 3$. LDA is unrealizable for this case. Also, Figure 3 shows the recognition results with standard deviation bars for D-LDA and NDT for the case of $M = 3$. We find that the recognition rates of NDT are more stable than those of D-LDA.

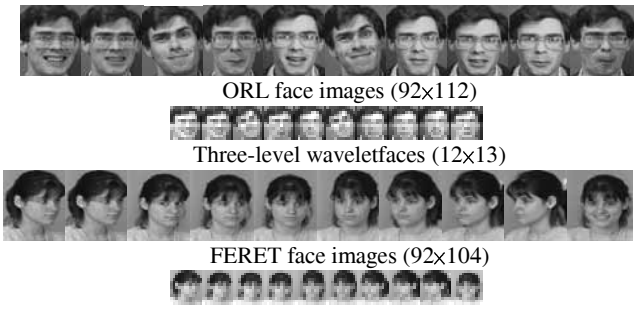

Figure 1: Some samples of two persons selected from ORL and FERET databases. Three-level waveletfaces are shown.

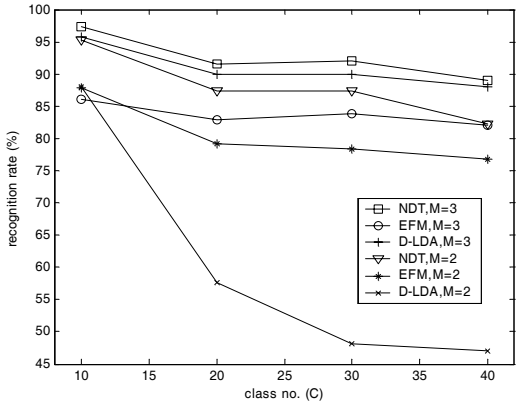

Figure 2: Comparison of recognition rates of D-LDA, EFM and NDT under different class numbers and training sample numbers. ORL database is used.

4.3 Evaluation of Baseline LDA and NDT on FERET Database

In the subsequent experiments, the conventional LDA is referred as the baseline for evaluation. Without any modification, LDA encounters small sample size problem for many cases of class number C and training sample number per class M. However, the proposed NDT performs a nonsingular transformation S_b prior to LDA procedure and try to maximize the between-class scatter, which is beneficial for discriminant pattern recognition.

Figure 4 displays the recognition rates of LDA and NDT when evaluating them using FERET face database. Herein, LDA is realizable for $C \geq 80$ when $M = 3$ and $C \geq 40$ when $M = 5$. We can see that NDT is better than LDA for different C and M. The recognition rates of NDT taper off to those of LDA as C increases. For this case, the Fisher class separabilities using LDA and NDT become unchanged.

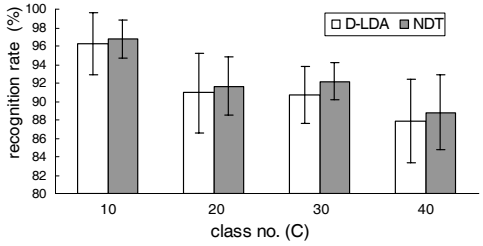

Figure 3: Recognition rates with standard deviation bars for D-LDA and NDT for $M = 3$ and different class numbers.

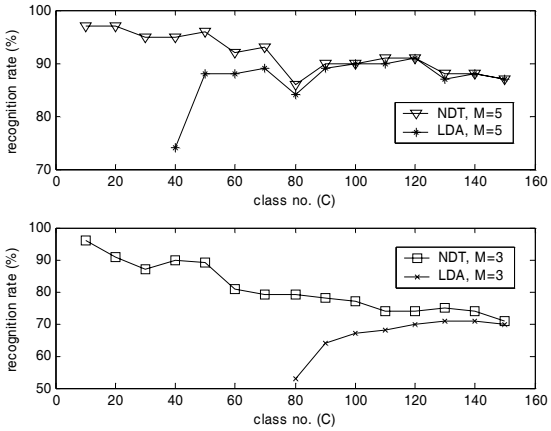

Figure 4: Comparison of recognition rates of baseline LDA and NDT under different class numbers and training sample numbers. FERET database is used.

5. CONCLUSION

We have presented a new NDT facial feature extraction for face recognition under different categories and training data numbers. NDT aimed to transform the original features into the nonsingular or principal space of between-class scatter matrix. The singularity problem of scatter matrices is resolved so as to fulfill the optimization of Fisher criterion. Our algorithm was developed and carried out by two transformations: nonsingular transformation and discriminant transformation. This scheme was applicable for LDA procedures in cases of singular as well as nonsingular scatter matrices. Using NDT, the transformed within-class and between-class scatters were unchanged when S_b was nonsingular. The Fisher class separability was unchanged as well. However, when S_b was singular, the transformed between-class scatter was unchanged and simultaneously the transformed within-class scatter was shrunk. The resulting Fisher class separability was increased. This method was different from principal component analysis (PCA) plus LDA, which transformed the features using the eigenvectors of total scatter matrix. In the experiments, we evaluated the feature extraction algorithms on ORL and FERET face databases. It is found that NDT solves the small sample size problem. NDT is superior to LDA. Recognition performance of NDT is substantially higher than that of EFM and D-LDA for various conditions.

6. REFERENCES

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

[2] J.-T. Chien and C.-C. Wu, "Discriminant waveletfaces and nearest feature decisions for face recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644-1649, December 2002.

[3] K. Etemad and R. Chellappa, "Face recognition using discriminant eigenvectors", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2150-2153, 1996.

[4] R. A. Fisher, "The statistical utilization of multiple measurements", *Annals of Eugenics*, vol. 8, pp. 376-386, 1938.

[5] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors", *IEEE Trans. on Computer*, vol. C-24, no. 3, pp. 281-289, March 1975.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition Second Edition*, Academic Press, Inc., 1990.

[7] K. Fukunaga and J. M. Mantock, "Nonparametric discriminant analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. PAMI-5, no. 6, pp. 671-678, November 1983.

[8] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", *Speech Communication*, vol. 26, no. 4, pp. 283-297, 1998.

[9] S. Z. Li and J. Lu. "Face recognition using the nearest feature line method". *IEEE Trans. Neural Networks,* vol. 10, no. 2, pp. 439-443, 1999.

[10] C. Liu and H. Wechsler, "Enhanced Fisher linear discriminant models for face recognition", *Proc. IEEE International Conference on Pattern Recognition,* vol. 2, pp. 1368 –1372, 1998.

[11] M. Loog, R. P. W. Duin and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 23, no. 7, pp. 762 –766, 2001.

[12] B. Noble and J. W. Daniel, *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[13] M. Ordowski and G. G. L. Meyer, "Geometric linear discriminant analysis", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3173 –3176, 2001.

[14] P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, 2000.

[15] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data - with application to face recognition", *Pattern Recognition*, vol. 34, no. 10, pp. 2067-2070, 2001.