FACE RECOGNITION BASED ON DISCRIMINANT EVALUATION IN THE WHOLE SPACE

Xudong Jiang, Bappaditya Mandal and Alex Kot

{exdjiang, bapp0001, eackot} @ ntu.edu.sg
School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

This paper proposes a face recognition approach that performs linear discriminant analysis in the whole eigenspace. It decomposes the eigenspace into two subspaces: a reliable subspace spanned mainly by the facial variation and an unstable subspace due to finite number of training samples. Eigenvalues in the unstable subspace are replaced by a constant. This alleviates the over-fitting problem and enables the discriminant evaluation in the whole space. Feature extraction or dimensionality reduction occurs only at the final stage after the discriminant assessment. These efforts facilitate a discriminative and stable low-dimensional feature representation of the face image. Experimental results comparing some popular subspace methods on FERET and ORL databases show that our approach consistently outperforms others.

***Index Terms*—** Face recognition, feature extraction, image recognition.

1. INTRODUCTION

Face recognition has gained significant attention in the last two decades due to its immense potential applications. After the introduction of the linear discriminant analysis (LDA) [1] in face recognition, numerous subspace based methodologies have been proposed. However, these methods discard a subspace before the discriminant evaluation to solve the singularity problem of the within-class scatter matrix. Fisherface (FLDA) [1] applies the principal component analysis (PCA) for dimensionality reduction before the application of LDA. The direct-LDA (DLDA) approach [2] first removes the null space of the between-class scatter matrix and then extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. Open question is how to scale the extracted features properly as the smallest eigenvalues are very sensitive to noise. Null space based approach (NLDA) [3] assumes that the null space of the within-class scatter matrix contains the most discriminative information and hence discards the range space. Interestingly, this appears to contradict the FLDA that only uses the range space and discards the null space. The unified framework of subspace methods (UFS) [4] discards even more dimensions and evaluates the discriminant value in a very small principal subspace. Its good recognition performance shows that the small eigenvalues of the within-class scatter matrix are unreliable and hence the removal of such dimensions from the discriminant evaluation alleviates the over-fitting problem. Open question is how to choose the number of principal dimensions for the first two stages of subspace decompositions before selecting the final number of features in the third stage.

A common problem of all these approaches is that they all lose some discriminative information, either in the principal or in the null space. As addressed in the literature, dimension reduction before the discriminant evaluation may result in the loss of crucial discriminative information [5, 6]. In fact, the discriminative information resides in the whole space. This is evidenced by the good recognition performance achieved by the dual-space approach (DSL) [6], which combines features extracted from the two complementary subspaces. Open questions of this approach are how to divide the space into the principal and the complementary subspaces and how to apportion a given number of features to the two subspaces. Furthermore, as the discriminative information resides in the both subspaces, it is inefficient or only suboptimal to extract features separately from the two subspaces.

In this paper, we present a new approach that performs discriminant evaluation in the whole space (DEWS). It decomposes the eigenspace of the within-class scatter matrix into a reliable and an unstable subspaces by finding the minimal value of the eigenratios. Eigenvalues in the unstable subspace are replaced by a constant to alleviate the over-fitting problem. This also enables the discriminant evaluation in the whole space and hence extracts the most discriminative and stable features for subsequent classification.

2. SUBSPACE DECOMPOSITION

Given a set of properly normalized face images, we can form a training set of column vectors $\{X_{ij}\}$, $X_{ij} \in \mathbb{R}^n$, for image j of person i. Let the training set contain p persons and q_i sample images for person i. The number of total training sample is $l = \sum_{i=1}^{p} q_i$. If c_i is the prior probability of person i, the within-class scatter matrix is defined by

$$\mathbf{S}^w = \sum_{i=1}^{p} \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)^T, \qquad (1)$$

where $\overline{X}_i = \frac{1}{q_i}\sum_{j=1}^{q_i} X_{ij}$. Letting $\overline{X} = \sum_{i=1}^{p} c_i \overline{X}_i$, the between-class scatter matrix $\mathbf{S}^b$ is defined by

$$\mathbf{S}^b = \sum_{i=1}^{p} c_i(\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})^T. \tag{2}$$

If all classes have equal prior probability, then $c_i = 1/p$.

It is well known that LDA is to solve the following eigen-decomposition problem:

$$(\mathbf{S}^{w^{-1}}\mathbf{S}^b)\mathbf{\Phi} = \mathbf{\Phi\Lambda}, \tag{3}$$

where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix and $\mathbf{\Phi}$ is the eigenvector matrix. Due to the singularity of $\mathbf{S}^w$, various approaches such as FLDA, DLDA, NLDA, UFS and DSL discard some dimensions and evaluate the discriminant value only in a subspace, which results in the loss of discriminative information [5, 6].

2.1. Problems in Feature Scaling And Extraction

The discriminant evaluation (3) can be performed by two separate eigen-decompositions: one for $\mathbf{S}^w$ and the other for $\mathbf{S}^b$ projected to the whitened eigenvectors of $\mathbf{S}^w$. Let $\mathbf{\Phi}^w = [\phi_1^w, ..., \phi_n^w]$ be the eigenvector matrix of $\mathbf{S}^w$, and $\mathbf{\Lambda}^w$ be the diagonal matrix of eigenvalues $\lambda_1^w, ..., \lambda_n^w$ corresponding to the eigenvectors. We assume that the eigenvalues are sorted in descending order $\lambda_1^w \geq, ..., \geq \lambda_n^w$. The plot of eigenvalues λ_k^w against the index k is called eigenspectrum. It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features. The whitened eigenvector matrix $\bar{\mathbf{\Phi}}^w = [\phi_1^w/\sigma_1^w, ..., \phi_n^w/\sigma_n^w]$, $\sigma_k^w = \sqrt{\lambda_k^w}$, is used to project the image vector X_{ij} before constructing the between-class scatter matrix for the second eigen-decomposition. Thus, image vector X_{ij} is first transformed by eigenvector, $Y_{ij} = \mathbf{\Phi}^{wT} X_{ij}$, and then multiplied by a weighting function $w_k^w = 1/\sqrt{\lambda_k^w}$. Discarding dimensions that have zero eigenvalues is equivalent to set $w_k^w = 0$ for these dimensions. The weighting function is thus

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r_w \\ 0, & r_w < k \leq n \end{cases}, \tag{4}$$

where r_w is the rank of $\mathbf{S}^w$. Fig. 1 shows a typical real eigenspectrum and the resulting weighting function. We see an undue sudden decrease of weighting function from the maximal value to zero. Furthermore, using the inverse of the square root of the eigenvalue to weight the eigenfeature amplifies noise and tends to over-fit the training samples. The small and zero eigenvalues are training-set-specific and very sensitive to different training sets. Adding new samples to the training set or using different training set may easily change some zero eigenvalues to nonzero and make some very small eigenvalues several times larger. Therefore, these small and zero eigenvalues are unreliable.

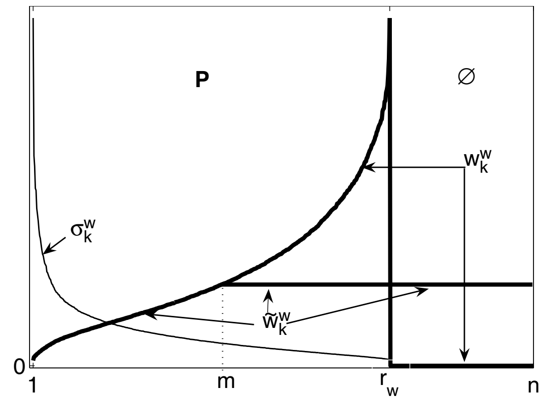

Fig. 1. Real Eigenspectrum and weighting functions (4), (8).

2.2. Subspace Decomposition Using Eigenratio-spectrum

The limited number of training samples and the high dimensionality of the image result in unreliable small and zero eigenvalues that may not well represent the true variance in the corresponding dimensions. Therefore, to improve the recognition accuracy, it is imperative to alleviate the problems of unreliable small and zero eigenvalues. It is well known that the eigenspectrum of the face image decreases rapidly and then stabilizes because the face variation resides in an intrinsic low dimension. Therefore, the phenomenon that the eigenspectrum accelerates its decrease is caused by the limited number of training samples. To study this, we define eigenratios as

$$\gamma_k^w = \frac{\lambda_k^w}{\lambda_{k+1}^w}, \quad 1 \leq k < r_w. \tag{5}$$

The plot of eigenratios γ_k^w against index k is called eigenratio-spectrum. Fig. 2 shows a typical eigenratio-spectrum of a real face training database. From the graph it is evident that the eigenratios first decreases very rapidly, then stabilizes and finally increases. The limited number of the training samples causes the increase of the eigenratios. The corresponding eigenvalues are thus unreliable. Therefore, the start point of the unreliable region $m+1$ is estimated by

$$\gamma_{m+1}^w = \min\{\forall \gamma_k^w, \quad 1 \leq k < r_w\}. \tag{6}$$

3. EIGENFEATURE SCALING AND EXTRACTION

From Fig. 1 it is evident that the inverses of the eigenvalues may cause undue over scaling of the features or eigenvectors for $k > m$. Bayesian Maximum Likelihood (BML) [7] algorithm uses a constant to replace the unreliable eigenvalues. The average eigenvalue in the unreliable subspace is used as the constant. The problem of this average eigenvalue is discussed in detail and an upper bound of eigenvalues in the unreliable subspace is proposed in [8] that enhances the recognition performance. However, there is no discriminant evaluation in the BML approach and hence the full image dimension

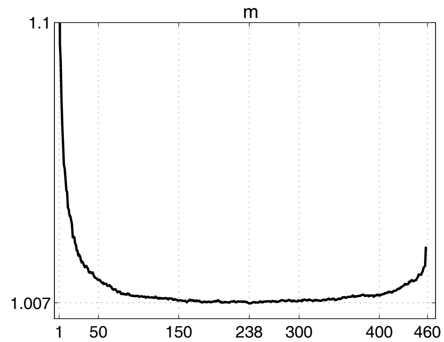

Fig. 2. Eigenratio-spectrum (5) from a real eigenspectrum.

is used in the classification, which is time consuming. In this work, we replace the unreliable eigenvalues $\{\lambda_k^w\}_{k=m+1}^n$ by a constant obtained by

$$\lambda_{const}^w = \max\{\forall\lambda_k^w, \quad m \leq k \leq r_w\}. \tag{7}$$

Thus, the final weighting function can be written as

$$\tilde{w}_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq m \\ 1/\sqrt{\lambda_{const}^w}, & m < k \leq n \end{cases}. \tag{8}$$

Fig. 1 shows the proposed feature weighting function $\tilde{w}_k^w$ calculated by (5), (6), (7) and (8) comparing with that w_k^w of (4). Obviously, the new weighting function $\tilde{w}_k^w$ is identical to w_k^w in the principal subspace and remains constant in the unreliable subspace and null space.

Using this weighting function and the eigenvectors ϕ_k^w, training samples are transformed to

$$\tilde{Y}_{ij} = \tilde{\mathbf{\Phi}}_n^{w^T} X_{ij}, \tag{9}$$

where

$$\tilde{\mathbf{\Phi}}_n^w = [\tilde{w}_k^w \phi_k^w]_{k=1}^n = [\tilde{w}_1^w \phi_1^w, ..., \tilde{w}_n^w \phi_n^w] \tag{10}$$

is a full rank matrix. There is no dimension reduction in this transformation. A new between-class scatter matrix is then formed by vectors $\tilde{Y}_{ij}$ of the training data as

$$\tilde{\mathbf{S}}^b = \sum_{i=1}^{p} c_i(\overline{\tilde{Y}}_i - \overline{\tilde{Y}})(\overline{\tilde{Y}}_i - \overline{\tilde{Y}})^T, \tag{11}$$

where $\overline{\tilde{Y}}_i = \frac{1}{q_i}\sum_{j=1}^{q_i} \tilde{Y}_{ij}$ and $\overline{\tilde{Y}} = \sum_{i=1}^{p} \frac{c_i}{q_i}\sum_{j=1}^{q_i} \tilde{Y}_{ij}$. The discriminant evaluation in the whole space is performed by solving the eigenvalue problem of $\tilde{\mathbf{S}}^b$. Suppose that the eigenvectors in the eigenvector matrix $\tilde{\mathbf{\Phi}}_n^b = [\tilde{\phi}_1^b, ..., \tilde{\phi}_n^b]$ are sorted in descending order of the corresponding eigenvalues. The dimensionality reduction is performed here by keeping the eigenvectors with the d largest eigenvalues

$$\tilde{\mathbf{\Phi}}_d^b = [\tilde{\phi}_k^b]_{k=1}^d = [\tilde{\phi}_1^b, ..., \tilde{\phi}_d^b], \tag{12}$$

where d is the number of features usually selected by a specific application. Thus, the proposed feature scaling and extraction matrix $\mathbf{U}$ is given by $\mathbf{U} = \tilde{\mathbf{\Phi}}_n^w \tilde{\mathbf{\Phi}}_d^b$. It transforms a face image vector X, $X \in \mathbb{R}^n$, into a feature vector F, $F \in \mathbb{R}^d$, by $F = \mathbf{U}^T X$. In the experiments of this work, the cosine distance measure and the first nearest neighborhood classifier (1-NNK) is applied to test the proposed approach of discriminant evaluation in the whole space (DEWS).

4. EXPERIMENTS AND DISCUSSIONS

We evaluate our proposed algorithm on FERET and ORL databases. In all the experiments, images are preprocessed following the CSU Face Identification Evaluation System [9]. The proposed DEWS method is tested and compared with PCA with Euclidian distance (PCAE), PCA with Mahalanobis distance (PCAM), FLDA, BML, DSL and UFS approaches.

In the first experiment, 2388 images comprising of 1194 persons are selected from the FERET database [10]. Images are cropped into the size of 38×33. 497 people are randomly selected for training and the remaining 697 people are used for testing. There is no overlap in person between the training and testing sets. The recognition error rate is the percentage of the incorrect top 1 match on the testing set. Fig. 3 shows the recognition error rate on the testing set against the number of features d used in the matching.

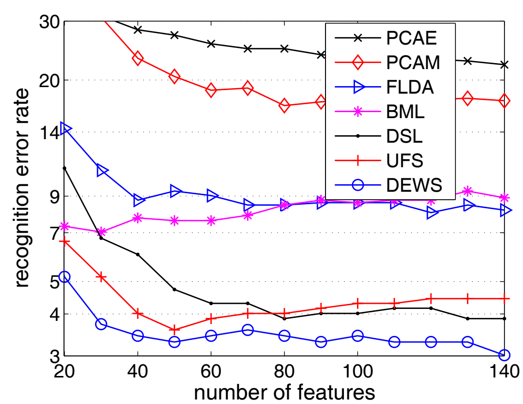

Fig. 3. Recognition error rate against the number of features on the FERET database of 994/1394 training/testing images.

Fig. 3 shows that DSL that uses information from two complementary subspaces performs better than PCAE, PCAM, FLDA and BML approaches. However, for small number of features, UFS outperforms DSL. This shows that it is inefficient or only suboptimal to extract features separately from the two subspaces. The proposed DEWS approach consistently outperforms all other approaches for all number of features.

In the second experiment, images of the ORL database [11] are cropped into the size of 57×50. The ORL database contains 400 images of 40 people (10 images per person). We use the first 5 samples per person for training and the remaining 5 samples per person for testing. Hence, there are 200 images in the training set and 200 images in the testing set. Fig. 4 shows the recognition error rate on the testing set against the number of features. As the training set has only 200 images, it may not well represent the variations of testing images. Therefore, the small principal space does not capture the discriminative information well. This results in poor performance of FLDA. UFS discards more dimensions before the discriminant evaluation and hence performs worse than FLDA. DSL that extracts features in two complementary subspaces is better than FLDA. BML that works in the whole space is better than DSL. Again, the proposed DEWS approach consistently outperforms all other approaches for all number of features.

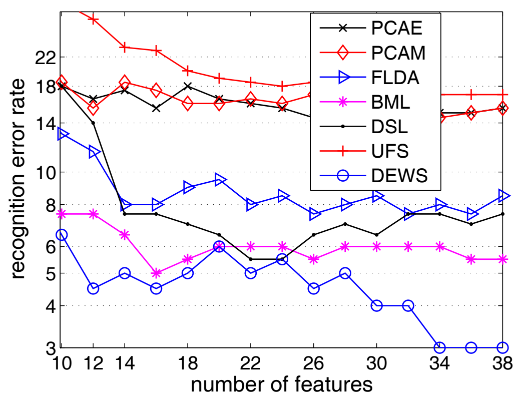

Fig. 4. Recognition error rate against the number of features on the ORL database of 200 training and 200 testing images.

5. CONCLUSIONS

This work addresses problems of eigenfeature scaling and extraction based on the linear discriminant analysis in face recognition. Dimension reduction before the discriminant evaluation may result in the loss of crucial discriminative information and the unreliable eigenvalues cause over-fitting problem. We define an eigenratio-spectrum to decompose the eigenspace into a reliable and an unstable subspaces. Eigenvalues in the unstable subspace are replaced by a constant determined by the largest eigenvalues in this subspace. This not only alleviates the over-fitting problem but also enables us to perform the discriminant evaluation in the whole space. Feature extraction or dimension reduction occurs only after the discriminant assessment. These facilitate a discriminative and stable low-dimensional feature representation of the face image, which boosts the face recognition accuracy. Experiments on the FERET and ORL databases demonstrate that the proposed approach consistently outperforms other approaches for all number of features.

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, July 1997.

[2] 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, October 2001.

[3] W. Liu, Y. Wang, S. Z. Li, and T. N. Tan, "Null space approach of fisher discriminant analysis for face recognition," in *ECCV workshop on Biometric Authentication*, Prague, Czech Republic, May 2004, pp. 32–44.

[4] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222–1228, September 2004.

[5] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, January 2005.

[6] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington DC, USA, June 2004, vol. 2, pp. 564–569.

[7] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, November 2000.

[8] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *IEE Electronics Letters*, vol. 42, no. 19, pp. 1089–1090, September 2006.

[9] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The csu face identification evaluation system users guide: Version 5.0," *Technical Report: http://www.cs.colostate.edu/evalfacerec/data/normalization.html*, 2003.

[10] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The feret evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[11] F. Samaria and A. Harter, "Parameterization of a stochastic model for human face identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA, December 1994, pp. 138–142.