Comparative Study: Face Recognition on Unspecific Persons using Linear Subspace Methods

Dahua Lin
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: dhlin4@ie.cuhk.edu.hk

Shuicheng Yan
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: scyan@ie.cuhk.edu.hk

Xiaoou Tang
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: xitang@microsoft.com

***Abstract*— Recently many Automatic Face Recognition (AFR) systems were developed for applications with unspecific persons, which is different from conventional pattern recognition problems where all classes are known in the training stage. In this paper, we present a systematic and comprehensive study on linear subspace methods for face recognition on unspecific persons. Over 6700 experiments using different algorithms with different training parameters and testing conditions are conducted on a large scale database (4550 samples) to investigate the compound effect of various influential factors. The observations based on these experiments are expected to provide widely applicable guidelines for designing practical AFR systems.**

I. Introduction

Owing to the wide range of application demands, face recognition techniques have been extensively studied and numerous experimental results were reported. Most of the experiments were designed to recognize persons among a known set of subjects. However, many practical systems are employed for unspecific persons which are unknown in training stage. Classification problems with specific classes and unspecific classes are two different problems: the former estimates the distribution of classes based on the training samples of given classes while the latter learns some discrimination rules from some classes which should be applicable to other classes.

In the past decade, statistical approaches were introduced into face recognition and great successes have been achieved by using a variety of linear subspace methods including PCA [1], LDA [2] and a series of its improved versions [3] [4] [5] [6] [7] [8] [9].

In addition to the selection of the algorithms, other factors including the size of training set, the similarity metrics and the number of features also notably affect the performance of the system. In this paper, we systematically and comprehensively investigate their interrelation and their compound effect on the the performance of the AFR system with a large quantity of experiments (over 6700 experiments in different parameters and conditions) on a large scale database (with 4550 samples from about 1400 subjects). we carefully examine the results and present our observations with theoretical analysis to justify the observations.

II. Linear Subspace Methods

In this section, we briefly review the linear subspace methods. In the following context, N denotes the total number of the training samples, C is the number of classes (subjects), and n_i means the number of samples in the i-th class. For each method, a projection matrix $\mathbf{W}$ is derived. Each sample $\mathbf{x}$ is transformed to the learned subspace as $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, and the similarity of any two samples can be measured with some metric.

A. Principle Component Analysis

Principle Component Analysis (PCA) [1] finds a set of bases that maximize the variation of all samples:

$$\mathbf{S}_t = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T \tag{1}$$

$$\mathbf{W} = \arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \operatorname{tr}(\mathbf{W}^T\mathbf{S}_t\mathbf{W}) \tag{2}$$

where m is the mean of all the samples. The solution can be obtained by taking the eigenvectors of $\mathbf{S}_t$ corresponding to the leading eigenvalues.

B. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [2] seeks a projection matrix that maximizes the trace-ratio of between-class scatter matrix $\mathbf{S}_b$ and within-class scatter matrix $\mathbf{S}_w$:

$$\mathbf{S}_b = \frac{1}{N}\sum_{i=1}^{C} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{3}$$

$$\mathbf{S}_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^T \tag{4}$$

$$\mathbf{W} = \arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\operatorname{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\operatorname{tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})} \tag{5}$$

where m_i is the mean of the samples in the i-th class and x_{ij} is the j-th sample in the i-th class. A series of LDA implementations have been proposed to obtain the solution in a robust and efficient way under the small-sample-size condition, and the most representative ones are described as

0-7803-9134-9/05/$20.00

below.

1) **Orthogonal Centroid Method (OCM)**
OCM [3] maximizes the scattering of class centroids:

$$\mathbf{W} = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{\operatorname{argmax}} \ \operatorname{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W}).$$

This method does not take the within class scattering matrix into account.

2) **Pseudo-Inverse LDA**
Pseudo-Inverse LDA (PILDA) [4] replaces $\mathbf{S}_w^{-1}$ by pseudo-inverse $\mathbf{S}_w^{\dagger}$, and performs eigenvalue decomposition on $\mathbf{S}_w^{\dagger}\mathbf{S}_b$ to obtain the solution.

3) **Generalized EVD**
As stated in [5], the solution of (5) satisfies $\mathbf{S}_b\mathbf{W} = \mathbf{S}_w\mathbf{W}\mathbf{\Lambda}$, which can be solved by Generalized Eigenvalue Decomposition (GEVD) despite the singularity of $\mathbf{S}_w$.

4) **Enhanced LDA**
[10] [6] derives the projection matrix by two stage's diagonalization: First, $\mathbf{S}_b$ is transformed to $\mathbf{K}_b = \mathbf{\Lambda}_w^{-1/2}\mathbf{\Phi}_w^T\mathbf{S}_b\mathbf{\Phi}_w\mathbf{\Lambda}_w^{-1/2}$ with $\mathbf{S}_w$ whitened by $\mathbf{\Phi}_w\mathbf{\Lambda}_w^{-1/2}$. Then the projection matrix is computed as $\mathbf{W} = \mathbf{\Phi}_w\mathbf{\Lambda}_w^{-1/2}\mathbf{\Phi}$, where $\mathbf{\Phi}$ is obtained from eigenvalue decomposition on $\mathbf{K}_b$ to maximize the scattering of the whitened class centers.

5) **Direct LDA**
[7] is based on the assumption that all discriminative information is within the principal subspace of $\mathbf{S}_b$. First, $\mathbf{S}_b$ is whitened by $\mathbf{\Phi}_b\mathbf{\Lambda}_b^{-1/2}$, with $\mathbf{S}_w$ transformed to $\mathbf{K}_w = \mathbf{\Lambda}_b^{-1/2}\mathbf{\Phi}_b^T\mathbf{S}_b\mathbf{\Phi}_b\mathbf{\Lambda}_b^{-1/2}$, eigenvector matrix $\mathbf{\Psi}$ that minimize $\mathbf{K}_w$ is used to get the final projection matrix is $\mathbf{W} = \mathbf{W}_1\mathbf{\Psi} = \mathbf{\Phi}_b\mathbf{\Lambda}_b^{-1/2}\mathbf{\Psi}$.

6) **Null-space LDA**
[8] is based on the assumption that the null space of $\mathbf{S}_w$ involves the most discriminative information. It first derives projection vectors $\mathbf{V}$ satisfying $\mathbf{V}^T\mathbf{S}_w\mathbf{V} = 0$. Then $\mathbf{S}_b$ is transformed to null space by $\mathbf{S}_{bn} = \mathbf{V}^T\mathbf{S}_b\mathbf{V}$. The projection matrix is $\mathbf{W} = \mathbf{V}\mathbf{\Psi}$, where $\mathbf{\Psi}$ is the eigenvector matrix that maximize $\mathbf{S}_{bn}$.

7) **Dual-space LDA**
[9] attempts to utilize the information in both principal subspace and null space of $\mathbf{S}_w$. After decomposing the whole space into principal subspace of $\mathbf{S}_w$ and its orthogonal complementary one, Enhanced LDA and nullspace LDA are integrated with their contributions scaled by eigenvalues.

III. Experiments

A. Experiment Configuration

Our experiments are conducted on the public databases: FERET [11], XM2VTS [12] and YALE database [13]. The training set consists of 2704 samples from 631 subjects selected from all the 3 databases, and each subject has 4 to 6 samples with different expressions. The testing set consists of 823 subjects. For each subject, one sample with neutral expression is selected as gallery sample, and other samples captured in different sessions are taken as probe samples. In total, there are 823 gallery samples and 1023 probe samples.

A variety of algorithms with different parameters and testing conditions are tested, here is a brief list:

- **8 Algorithms based on PCA or LDA**.
  PCA, OCM, Pseudo-Inverse LDA, GEVD-LDA, Enhanced LDA, Direct LDA, Null-space LDA and Dual-space LDA.

- **10 Different feature lengths**.
  Projected feature vectors are truncated by retaining the first 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 components.

- **2 Different metrics**.
  Two different metrics are tested: 1. L2 Distance: $d = ||\mathbf{x}_1 - \mathbf{x}||^2$, 2. Normalized Correlation: $s = \frac{\mathbf{x}_1^T\mathbf{x}_2}{\sqrt{||\mathbf{x}_1||^2||\mathbf{x}_2||^2}}$.

- **3 Different constructions of $\mathbf{S}_b$**.
  Schemes that select different sample pairs and weighting functions were proposed in [14] [10] to construct $\mathbf{S}_b$. Here we test 3 different constructions: using differences between class centers and total center (standard), using difference between every sample to nearest class center of different class (KNC), using difference between every sample to nearest sample of different class (KNS).

- **7 Different training set sizes**.
  We construct 6 subsets of the whole training set, and they have 100, 200, 300, 400, 500, 600 subjects respectively. Together with the whole training set, we train models in 7 different training set sizes.

- **2 Applications**.
  1. Face identification with performance measured in correct rate. 2. Face verification with performance measured in equal error rate.

B. Experiment Observations

1) **Comparison on Algorithms.**
The performances of 8 algorithms trained on the whole training set are illustrated in Fig.1 and Fig.2. The results show that the algorithms (PCA and OCM) without

utilizing $\mathbf{S}_w$ are significantly inferior to LDA-based algorithms, which indicates the important role of $\mathbf{S}_w$ for discriminating. In the following discussion, we will focus on the 6 LDA-based algorithms.

The performance differences between LDA-based methods are conspicuous when feature length is small. But when the feature length continues increasing, the performances become stable, and gradually converge to nearly the same level. From Fig.1 and Fig.2 and Table.1, Dualspace LDA is the best performer, which is slightly better than Enhanced LDA. Nullspace LDA is the weakest.

2) **Comparison on Metrics**

We observed that the selection of metrics on final feature subspace has more significant impact on the performance. The comparison of two metrics is illustrated in Fig.3 and Fig.4, from which we can see that, the Normalize Correlation metric consistently outperforms L2 Distance in all LDA-based algorithms. The ratio of rising in recognition accuracies ranges from 9% to 16%. In face verification, the improvement is more remarkable, and the error rate drops down by 70% to 80%. The results convincingly demonstrate that the phase information plays a dominant role in classification instead of magnitude information.

3) **Comparison on Different Construction of $\mathbf{S}_b$**

A series of improved construction of $\mathbf{S}_b$ have been proposed in [14] and [10], based on the rationale that more emphasis should be placed on the samples near boundary, which are assumed to convey more discriminative information. Though the improvement of weighted schemes or nonparametric construction is encouraging in problems with specific subjects, as shown in Fig.5 and Fig.6, only slight improvements are obtained from two nonparametric constructions for unspecific persons. It is because that when a new probe subject comes, its boundary information may be far different from those in the trained model, and in such cases, the nonparametric construction is prone to overfitting.

4) **Issues of Training Set Size**

The effect of the training set size is also investigated in our experiments. Fig. 7 and Fig. 8 show that when the number of subjects in the training set is small, the LDA algorithms are sensitive to the training set size, and their performances ascend fast when the training set size is increased. However, when the size of training set is relatively large, their performances become stable.

5) **Sensitivity of GEVD**

We observed that GEVD-LDA is very sensitive to the size of training set. When the number of subjects in the training set is below 400, the performance of GEVD-LDA degrades drastically, which indicates that the training process of GEVD is sensitive to input.

IV. Conclusion

We presented a comprehensive study on linear subspace methods for face recognition on unspecific persons based on a large number of experiments. The observations, including the converging trend of algorithm performance with parameters and training size change, and the significant influence of metric selection, are important for designing real AFR systems on unspecific persons.

ACKNOWLEDGEMENTS

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong SAR. The work was done while all the authors are with the Chinese University of Hong Kong.

References

[1] M.Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI.*, vol. 19, no. 7, pp. 711–720, 1997.

[3] H. Park, M. Jeon, and J. Rosen, "Lower dimensional representation of text data based on centroids and least squares," *BIT*, 2003.

[4] Q. Tian, Y. Fainman, and S.H. Lee, "Comparison of statistical pattern-recognition algorithms for hybrid processing. ii. eigenvector-based algorithm," *J. Opt. Soc. Am*, 1988.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Ed)*, Wiley-Interscience, 2001.

[6] C. Liu and H. Wechsler, "Enhanced fisher linear discriminant models for face recognition," *Proceedings of ICPR'98*, vol. 2, pp. 2, 1998.

[7] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data - with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.

[8] L. Chen, J. Lin H. Liao, M. Ko, and G. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Journal of Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[9] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," *Proceedings of CVPR'04*, vol. 2, pp. 564–569, 2004.

[10] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, 1990.

[11] P. J. Phillips, H. Moon, S. A. Ryzvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, no. 10, pp. 1090–1104, 2000.

[12] J. Luettin and G. Maitre, "Evaluation protocol for the extended m2vts database (xm2vts).," *DMI for Perceptual Perceptual Artificial Intelligence*, 1998.

[13] "Yale univ. face database," *http:// cvc.yale.edu / projects / yalefaces / yalefaces.html*, 2002.

[14] M.Loog, R.P.W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise fisher criteria," *IEEE Trans. PAMI*, vol. 23, no. 7, pp. 762–766, 2001.

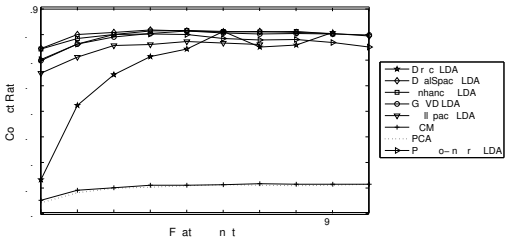

Fig. 1. Comparison of different algorithms for face identification

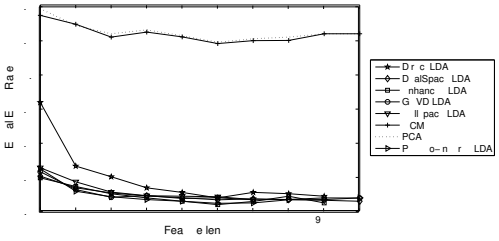

Fig. 2. Comparison of different algorithms for face verification

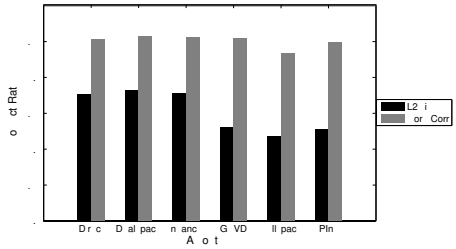

Fig. 3. Comparison of different metrics for face identification

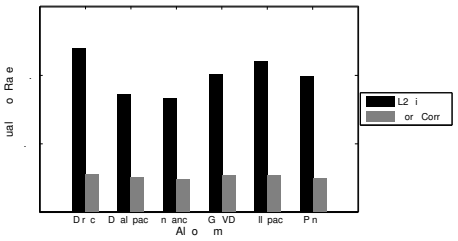

Fig. 4. Comparison of different metrics for face verification

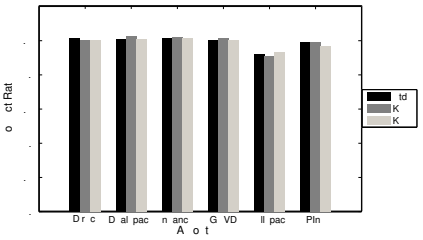

Fig. 5. Comparison of different methods for the construction of $\mathbf{S}_w$ for face identification

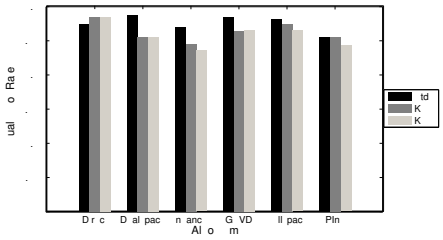

Fig. 6. Comparison of different methods for the construction of $\mathbf{S}_w$ for face verification

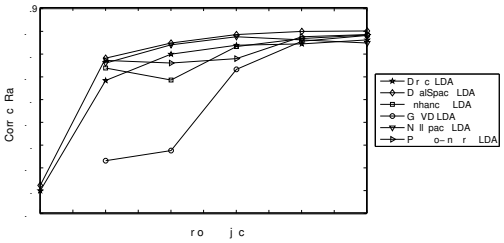

Fig. 7. Influence of training size to performance of face identification

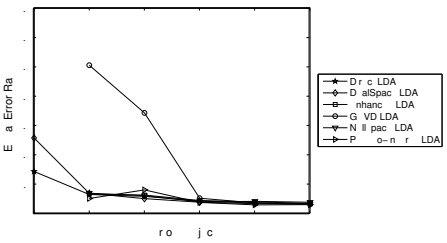

Fig. 8. Influence of training size to performance of face verification

		Direct	Dualspace	Enhanced	GEVD	Nullspace	PInverse	OCM	PCA	Best
Recognition Correct Rate	L2 Dist	0.7752	0.7810	0.7771	0.7292	0.7165	0.7263	0.5562	0.5552	0.7810
	NormCorr	0.8524	0.8553	0.8544	0.8534	0.8328	0.8475	0.5445	0.5445	0.8553
	Best	0.8524	0.8553	0.8544	0.8534	0.8328	0.8475	0.5562	0.5552	0.8553
Verification Error Rate	L2 Dist	0.1189	0.0860	0.0829	0.1003	0.1095	0.0987	0.1243	0.1254	0.0829
	NormCorr	0.0274	0.0254	0.0235	0.0263	0.0264	0.0242	0.1174	0.1182	0.0235
	Best	0.0274	0.0254	0.0235	0.0263	0.0264	0.0242	0.1174	0.1182	0.0235

TABLE I

THE PERFORMANCES FROM USING DIFFERENT ALGORITHMS AND METRICS FOR FACE IDENTIFICATION AND VERIFICATION