Multiple-Exemplar Discriminant Analysis for Face Recognition

Shaohua Kevin Zhou and Rama Chellappa
Center for Automation Research and
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
{shaohua,rama}@cfar.umd.edu

Abstract

Face recognition is characteristically different from regular pattern recognition and, therefore, requires a different discriminant analysis other than linear discriminant analysis (LDA). LDA is a single-exemplar method in the sense that each class during classification is represented by a single exemplar, i.e. the sample mean of the class. In this paper, we present a multiple-exemplar discriminant analysis (MEDA) where each class is represented using several exemplars or even the whole available sample set. The proposed approach produces improved classification results when tested on a subset of FERET database where LDA is ineffective.

1. Introduction

Fisher linear discriminant analysis (LDA) is a standard pattern recognition tool. LDA is a single-exemplar method in the sense that each class during classification is represented by a single exemplar, i.e. the sample mean of the class. The single-exemplar property offers a simple classification mechanism, which is often very efficient in terms of classification results. The underlying assumption of LDA is that each class possesses a normal density with a different mean vector but a common covariance matrix. Under the above assumption, LDA coincides with the optimal Bayes classifier.

Even though LDA has been successfully applied to face recognition [1, 3, 9], its recognition effectiveness is limited to controlled scenarios, as documented in [7, 10]. For example, when the faces are in a frontal view, under a frontal illumination, and with a neutral expression, the recognition performance is quite accurate. However, when the image conditions of the training, gallery, and probe sets are different, the recognition performance drops quickly.

The inconsistency in recognition performance can be explained by the fact that face recognition is characteristically different from regular pattern recognition. Generally speaking, the main hurdle in face recognition is the sample-deficiency problem, i.e., there is only a small number of samples per class to represent a complex manifold. Therefore, all samples should be used as exemplars for a final classification. This is inconsistent with the single-exemplar property and causes LDA to be ineffective.

To overcome this drawback, we propose a multiple-exemplar discriminant analysis (MEDA) where each class is represented by several exemplars. Rather than minimizing the within-class distance while maximizing the between-class distance , the proposed MEDA find the projection directions along which the within-class exemplar distance (i.e. the distances between exemplars belonging to the same class) is minimized while the between-class exemplar distance (i.e. the distances between exemplars belonging to different classes) is maximized. To illustrate the effectiveness of the proposed approach, we test MEDA on a subset of the FERET database [7] where LDA has been ineffective.

The paper is structured as follows. In Section 2, we list some characteristics of face recognition that are different from regular pattern recognition. Then, in Section 3 we review the principle of LDA. In Section 4, we present the principle of MEDA, followed by several special examples of MEDA. We then in Section 5 present the experiment part, demonstrating the necessity of MEDA in cases where LDA may be ineffective. We conclude the paper in Section 6.

2. Characterization of Face Recognition

Face recognition possesses several characteristics different from regular pattern recognition (especially those appropriate for LDA).

[*C1*] Due to variations in such as pose, illumination, and facial expression, the face appearance of an object possess a complex density (or manifold), severely deviating from the normal assumption. In other words, the single-exemplar property of LDA is violated. Consequently, a very large

number of samples are required to sufficiently represent the complex density (or manifold).

[*C2*] Because of limitations of image acquisition, practical face recognition systems store only a small number of samples per subject. This aggregates the 'curse of dimensionality' problem. Typically, each sample represents one type of variations. For instance, we might have one sample under a frontal illumination and with a neutral expression, one sample under a different illumination and with a neutral expression, and one sample under a frontal illumination and with a different expression. Even so, it is far from sufficient to represent the complex density we are dealing with. Therefore, every sample matters and should be used as an exemplar, i.e., we should by all means to use all available samples during classification rather than using their mean as in LDA.

[*C3*] To remedy the sample-deficiency problem to some extent, one can exploit the strong visual similarity among face images of different subjects. It is this similarity that inspires the popularity of the 'Eigenface' approach [8]. However, again the 'Eigenface' approach actually only works for controlled scenarios. In our context, this similarity can be interpreted as a similarity among the 'shapes' of the face appearance manifolds belonging to different subjects. To understand the manifold 'shape', we use the analogy of derivative. The manifold is considered as a multidimensional function and its 'shape' as the 'derivatives' of the appearance manifolds.

[*C4*] Unlike regular pattern recognition where the class labels involved in training and testing are same, face recognition systems often have no overlap between the training set and the galley/probe set, according to the FERET protocol [7]. Thus, generalization from known subjects in the training set to unknown subjects in the gallery/probe set is needed. Fortunately, this generalization is possible due to the above-mentioned similarity. Once we learn the 'shape' characteristic of the face manifold, we can apply this knowledge to novel subjects since the 'shapes' of all face manifolds are similar.

3. Linear Discriminant Analysis (LDA)

Consider a C-class problem with each class i consisting of a set of N_i d-dimensional samples $\{\mathsf{x}_1^i, \mathsf{x}_2^i, \ldots, \mathsf{x}_{N_i}^i\}$ where the superscript $(.)^i$ represents the class label. For illustrative purpose, we introduce a grand class which deprives the class labels. Denote the total number of samples by $N = \sum_{i=1}^{C} N_i$, the frequency of occurrence of the i^{th} class by $\mathsf{p}^i = N_i/N$, the sample mean for the i^{th} class by μ^i, and the grand sample mean (regardless of class labels) by μ. We compute the above-defined quantities as follows:

$$\mu^i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathsf{x}_j^i; \;\; \mu = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\mathsf{x}_j^i = \sum_{i=1}^{C}\mathsf{p}^i\mu^i. \quad (1)$$

LDA first estimates the within-class and between-class scatter matrices of size $d \times d$, denoted by Σ_W and Σ_B, respectively, given by

$$\Sigma_W = \sum_{i=1}^{C}\mathsf{p}^i\Sigma_W^i = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\mathsf{x}_j^i - \mu^i)(\mathsf{x}_j^i - \mu^i)^\mathsf{T}, \quad (2)$$

$$\Sigma_B = \sum_{i=1}^{C}\mathsf{p}^i\Sigma_B^i = \frac{1}{N}\sum_{i=1}^{C}N_i\sum_{j=1}^{N_i}(\mu^i - \mu)(\mu^i - \mu)^\mathsf{T}, \quad (3)$$

where Σ_W^i is the covariance matrix estimate for class i given by

$$\Sigma_W^i = \frac{1}{N_i}\sum_{j=1}^{N_i}(\mathsf{x}_j^i - \mu^i)(\mathsf{x}_j^i - \mu^i)^\mathsf{T}, \quad (4)$$

and Σ_B^i is the scatter matrix between the class i and the 'grand class' given by

$$\Sigma_B^i = (\mu^i - \mu)(\mu^i - \mu)^\mathsf{T}. \quad (5)$$

In other words, Σ_W is estimated by 'pooling' together $\{\Sigma_W^i;\ i = 1, ..., C\}$. Similarly, this holds for Σ_B.

Then, LDA finds a projection matrix W, say of size $r \times d$, that maximizes the criterion function

$$J_\mathsf{W} = \frac{det\{\mathsf{W}^\mathsf{T}\Sigma_B\mathsf{W}\}}{det\{\mathsf{W}^\mathsf{T}\Sigma_W\mathsf{W}\}}, \quad (6)$$

where $det\{.\}$ denotes matrix determinant. The value of r can not exceed $d - 1$. Given a test pattern y, its class label C_y is determined as

$$C_\mathsf{y} = \arg\min_{i=1,2,...,C}\{|\mathsf{W}^\mathsf{T}(\mathsf{y} - \mu^i)|^2 + \mathsf{D}_i\}, \quad (7)$$

where D_i is used to incorporate prior information[1].

4. Multiple-Exemplar Discriminant Analysis (MEDA) for Face Recognition

The basic principle of LDA is to minimize the within-class distance while maximizing the between-class distance, with each class represented by a single exemplar. Since MEDA uses all the available exemplars per class, the

[1] In practice, one often ignores the term D_i and use the L_1 norm instead of L_2 norm. We did this in the experiment reported herein.

within-class distance in LDA becomes the within-class exemplar distance (i.e. the distances between exemplars belonging to the same class).

Mathematically, we re-define the matrices Σ_W and Σ_B as follows:

$$\Sigma_W = \sum_{i=1}^{C} \frac{1}{N_i^2} \sum_{j=1}^{N_i} \sum_{k=1}^{N_i} (\mathsf{x}_j^i - \mathsf{x}_k^i)(\mathsf{x}_j^i - \mathsf{x}_k^i)^{\mathsf{T}}; \quad (8)$$

The basic element in (8) is a pairwise difference between any two exemplars belonging to the same class. Alternatively, we can view these basic elements as samples of a new space. This construction of such a space is validated by the property *C3* to capture the common 'shape' of the face appearance manifold. This space is called the intra-personal space (IPS) in [5].

Similarly, the between-class distance in LDA becomes the between-class exemplar distance (i.e. the distances between exemplars belonging to different classes),

$$\Sigma_B = \sum_{i=1}^{C} \sum_{j=1;j\neq i}^{C} \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} (\mathsf{x}_k^i - \mathsf{x}_l^j)(\mathsf{x}_k^i - \mathsf{x}_l^j)^{\mathsf{T}}, \quad (9)$$

and a so-called extra-personal space (EPS) can be constructed.

The proposed MEDA approach is find the projection matrix $\mathsf{W}_{d\times r}$ such as the same cost function J_{W} defined in (6) is maximized. But, here the number of projection directions r can exceed $C-1$. Given a test pattern y, its class label C_{y} is determined as

$$C_{\mathsf{y}} = \arg \min_{i=1,2,\ldots,C} \{ \min_{j=1,2,\ldots,N_i} \{|\mathsf{W}^{\mathsf{T}}(\mathsf{y} - \mathsf{x}_j^i)|^2 + \mathsf{D}_i\}\}. \quad (10)$$

Without much difficulty, our MEDA analysis can be extended to handle the cases where not all samples are used in classification and only several exemplars are extracted from the sample set to represent the class. Mathematically, we represent a class i by M_i exemplars associated with weights, i.e., $\{(\mu_k^i, \mathsf{p}_k^i);\ k = 1, 2, \ldots, M_i\}$ where the sample mean μ^i for class i is given by $\mu^i = \sum_{k=1}^{M_k} \mathsf{p}_k^i \mu_k^i$. Note that if we take $\mathsf{p}_k^i = N_i^{-1}$, it reduces to the proposed analysis.

Learning exemplar can be achieved by the K-means algorithm [2]. The K-means algorithm is a hard-clustering technique. A soft-clustering method such as mixture modeling can also be incorporated in the extended MEDA analysis.

4.1. Smoothing

As pointed out in the LDA literature [2] and its applications to face recognition [1, 3, 9], the projection directions themselves are very noisy and wiggly, which is an indication of over-fitting. Fig. 2 (a) also shows this phenomenon.

The over-fitting can be remedied by adding a penalty matrix Ω to the matrix Σ_W, as suggested by [4]. This penalty matrix penalizes the roughness of the projection vectors (they are actually images as shown in Fig. 2) to encourage smooth solutions. Alternatively, a pre-smoothing step which filters out the high-frequency components from the original images can be used. For example, used in [9] is an eigen-smoothing technique which is essentially the PCA approach (by retaining the top q components). In this paper, we adopt the former approach, i.e., adding an matrix $\Omega = \rho \mathtt{I}$ to Σ_W, with $\mathtt{I}$ being an identity matrix. The typical range of ρ is [10, 50].

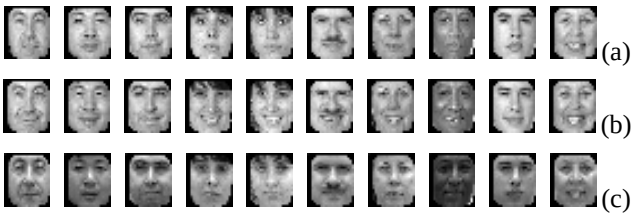

Figure 1. (a) Neutral faces. (b) Faces with facial expressions. (c) Faces under a different illumination. The image size is 24 by 21 in pixels.

4.2. Discussion

Our analysis is different from the Bayesian face recognition approach [5]. In [5], after constructing the intra-personal space (IPS) and extra-personal space (EPS), multivariate densities are fitted on top of them. The probabilistic subspace [6], which possesses some smoothing capability, is used. However, fitting the probabilistic subspace density on the IPS/EPS is not guaranteed to be optimal. Our discriminant analysis is based only on second-order statistics and no density fitting is needed.

5. Experimental Results

We perform face recognition using a subset of the FERET database [7] with 200 subjects only. Each subject has 3 images: (a) one taken under controlled lighting condition with a neutral expression; (b) one taken under the same lighting condition as above but with different facial expressions (mostly smiling); and (c) one taken under different lighting condition and mostly with a neutral expression. Fig. 1 shows some face examples in this database. All images are pre-processed using zero-mean-unit-variance operation and manually registered using the eye positions.

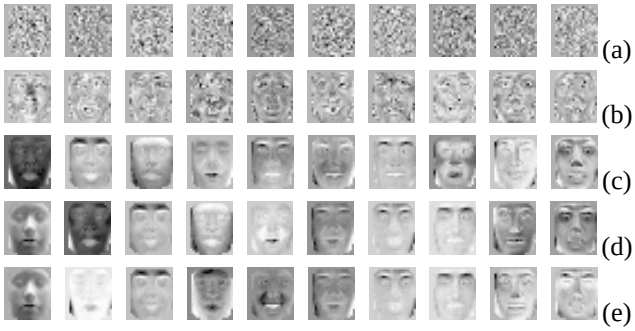

Figure 2. (a) The top 10 discriminant component [1, 3]. (b) The top 10 discriminant component obtain in [9] (with a pre-smoothing). (c) The top 10 principal components of the IPS [5]. (d) The top 10 principal components of the EPS [5]. (e) The top 10 discriminant components of the MEDA approach.

	Expression	Illumination
MEDA	66%	72%
IPS	64%	69%
BayesFR	50%	50%
subLDA	55%	59%
LDA	44%	43%

Table 1. A summary of recognition rates obtained by different approaches.

We randomly divide the 200 objects into two sets, with one set for training and the other one for testing. The training set consists of 300 images, with three images belonging to the 100 training subjects. Because we focus on the effects of two different variations in facial expression and illumination, for one particular variation, say expression variation, we use the remaining 200 images as the gallery and probe sets for testing, with the 100 images in the category (a) as the gallery set, and the 100 images in the category (b) as the probe set.

For comparison, we implement the following three discriminant methods besides the proposed MEDA approach: the LDA approach [1, 3], the 'subLDA' approach approach [9] (PCA followed by LDA), and the Bayesian face recognition ('BayesFR') approach [5]. In addition, we also implement the 'IPS' approach in which the projection vectors are eigenvectors of the IPS. For each of the tested approaches, we tune the parameters (e.g. the number of components) to maximize the recognition performance. Fig. 2 shows the projection directions obtained in the tested approaches. See the figure caption for detailed description.

Table 1 lists the recognition rates obtained by all tested approaches, using the top one match. It is not surprising that the LDA approach records the worst performance since the underlying assumptions of LDA are severely violated. The 'subLDA' approach overperforms the LDA approach which highlights the virtue of eigen-smoothing as a preprocessing method. The 'BayesFR' approach is also better than the LDA approach, however the improvement is not very significant possibly because the fitted density is misspecified. The 'IPS' approach is very competitive, which confirms the face characteristics *C3*, i.e., the IPS characterizes the 'shape' of the face manifold. The proposed MEDA approach yields the best performance since it performs a discriminant analysis of the IPS and EPS, with multiple-exemplar modeling embedded.

6. Conclusion

In this paper, we illustrated the characteristics of face recognition other than those of regular pattern recognition. These characteristics inspires the propose multiple-exemplar discriminant analysis in lieu of regular linear discriminant analysis. The preliminary results are very promising and we still need to investigate the recognition performance on a large-scale database. Finally, even though we use face recognition as an application, our analysis is quite general and is applicable to other recognition tasks, especially those involving very high dimensional patterns.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19, 1997.

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

[3] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of Optical Society of America A*, pages 1724–1733, 1997.

[4] T. Hastie and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102, 1995.

[5] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian modeling of facial similarity. *Advances in Neural Information Processing System*, 1998.

[6] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696–710, 1997.

[7] P. J. Phillips, H. Moon, S. Rivzi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Tran. PAMI*, 22:1090–1104, 2000.

[8] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:72–86, 1991.

[9] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, pages 14–16, 1998.

[10] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, 12:399–458, 2003.