ELectronics
EXpress

Two-fold regularization for kernel Fisher discriminant analysis in face recognition

Sang-Ki Kim, Kar-Ann Toh, and Sangyoun Lee[a)]

Department of Electrical and Electronic Engineering,
Biometrics Engineering Research Center, Yonsei University,
134 Shinchon-dong, Seodaemun-gu, Seoul, Korea
a) *syleee@yonsei.ac.kr*

Abstract: Due to the inherent nature of kernel implementation, the kernel Fisher discriminant suffers from the small sample size problem. In this paper, we introduce a novel variant of the kernel Fisher discriminant formulation to circumvent this problem. By adopting a two-fold regularization scheme on the scatter matrices, we show both effectiveness and reliability of the proposed method particularly regarding the small sample size and the lack of dimensionality issues.

Keywords: face recognition, subspace learning, kernel discriminant analysis

Classification: Science and engineering for electronics

References

[1] G. Baudat and F. Anouar, “Generalized Discriminant Analysis Using a Kernel Approach,” *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[2] J. Lu, K. N. Plataniotis, and A. Venetsanopoulos, “Face Recognition Using Kernel Direct Discriminant Analysis Algorithms,” *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

[3] 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, pp. 711–720, May 1997.

[4] X. Jiang, B. Mandal, and A. Kot, “Eigenfeature Regularization and Extraction in Face Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, March 2008.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

[6] A. R. Martinez and R. Benavente, “The AR face database,” Technical Report 24, Computer Vision Center (CVC), June 1998.

1 Introduction

The linear discriminant analysis (LDA) has been widely applied as an effective feature extraction method in the area of face recognition. Essentially, LDA calculates the discriminating subspace based on the Fisher’s criterion.

DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009

Owing to the direct discriminant and supervised nature of this criterion, the LDA has shown remarkable performance in terms of recognition accuracy. To enhance the performance for nonlinear decision boundaries, several nonlinear discriminant subspaces have been proposed [1, 2]. Via a kernel trick, these methods offer an effective subspace in the high dimensional feature space.

The Fisher's criterion is defined on a ratio of the between-class scatter matrix and the within-class scatter matrix which needs to be inverted. In consequence, conventional LDA suffers from the so-called small sample size (SSS) problem [5] which is caused by a singular within-class scatter matrix. A common way to resolve SSS problem is to collect training images more than the matrix size. However, for kernel-based implementation of the Fisher's criterion, due to the inherent characteristic of the kernel trick, the within-class scatter matrix becomes always singular, and thus the SSS problem is unavoidable. The SSS problem also results in sampling noise such that most of the small eigenvalues of the within-class scatter matrix become very unreliable during matrix inversion [4].

Another problem of the Fisher criterion is the small dimensionality available for extracting features. The discriminating subspace is calculated from an intersection of the subspaces spanned by the within-class scatter and the between class scatter. Since the rank of between-class scatter is bounded by $C - 1$, which is one less than the number of classes/identities (C), the intersection happens to possess an even smaller dimensionality. This shortage of dimensionality eventually leads to loss of useful discriminative information from the training data.

To compensate for these problems in kernel discriminant analysis (KDA), a regularization scheme would be necessary. Recently Jiang et al. suggested an eigen-spectrum regularization (ER) technique for linear feature extraction [4]. ER modifies small and zero eigenvalues via a modeling function. This relaxes those sample noises residing in the zone of small eigenvalues and allows utilization of the null-space for computation of the inverse of the within-class scatter.

Motivated by the reliability and effectiveness of ER in linear feature extraction, in this work we propose a new variant of KDA, a two-fold regularized kernel discriminant analysis (R-KDA), for face recognition. By applying the regularization scheme of eigenfeature regularization and extraction (ERE) [3], this R-KDA attempts to alleviate the SSS problem and the dimensionality problem.

2 A two-fold regularization procedure for kernel Fisher discriminant

Suppose we have a nonlinear mapping function ϕ which implicitly maps the input space ($x_i \in \mathbb{S}$) into a high-dimensional feature space ($\phi_i \in \mathbb{F}$):

$$\phi : \mathbb{S} \rightarrow \mathbb{F}.$$

The KDA finds the discriminating subspace within the feature space ($\mathbb{F}$) where the complex classification structure is hopefully linearized. The bases

DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009

(v) of the demanded subspace can be driven by solving the following eigenvalue problem:

$$v = \arg\max_{v} \left(\left| \frac{v^{\mathrm{T}}\mathbf{T}v}{v^{\mathrm{T}}\mathbf{W}v} \right| \right), \tag{1}$$

where $\mathbf{T}$ and $\mathbf{W}$ are the total scatter matrix and the within-class scatter matrix, respectively. Different from the conventional Fisher's criterion, we used $\mathbf{T}$ instead of the between-class scatter ($\mathbf{B}$) following [4]. Using $\mathbf{T}$ has more advantageous characteristic than $\mathbf{B}$ in the aspect of regularization. It is also empirically supported by [4] where Jiang et al. tested both $\mathbf{B}$ and $\mathbf{T}$. Since $\mathbf{T}$ can be decomposed into a sum of $\mathbf{W}$ and $\mathbf{B}$ [5], maximizing (1) is equivalent to maximizing the conventional Fisher's criterion with a regularization term:

$$\lambda = \frac{v^{\mathrm{T}}\left(\mathbf{B}+\mathbf{W}\right)v}{v^{\mathrm{T}}\mathbf{W}v} = \frac{v^{\mathrm{T}}\mathbf{B}v}{v^{\mathrm{T}}\mathbf{W}v} + v^{\mathrm{T}}\mathbf{I}v, \tag{2}$$

where λ is the eigenvalue and $\mathbf{I}$ is the identity matrix. By utilizing $\mathbf{T}$, we obtain a higher rank $(\mathrm{N}-1)$ in the numerator, where N is the number of training samples, instead of $\mathbf{B}$ with rank $\mathrm{C}-1$. Here, the dimensionality limit by $\mathbf{B}$ is much loosened so that the entire sample space is covered.

Since any eigenvector (v) with $\lambda \neq 0$ must lie within the span of training samples, it can be represented as a linear combination of $\mathbf{\Phi}$ [1, 2]:

$$v = \sum\nolimits_{i=1}^{\mathrm{N}} \alpha_i \phi_i = \mathbf{\Phi}\alpha, \tag{3}$$

where $\mathbf{\Phi} = [\phi_i, \ldots, \phi_{\mathrm{N}}]$ correspond to the training samples in feature space, and $\alpha = [\alpha_1, \ldots, \alpha_{\mathrm{N}}]$ correspond to a weighting factor in linear combination. $\mathbf{T}$ and $\mathbf{W}$ can be expressed in terms of $\mathbf{\Phi}$ as follows:

$$\begin{aligned} \mathbf{T} &= \frac{1}{\mathrm{N}} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{N}} (\phi_i - \bar{\phi})(\phi_i - \bar{\phi})^{\mathrm{T}} = \mathbf{\Phi}(\mathbf{I} - 2\mathbf{A} + \mathbf{A}\mathbf{A})\mathbf{\Phi}^{\mathrm{T}}, \\ \mathbf{W} &= \frac{1}{\mathrm{N}} \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\phi_{ij} - \bar{\phi}_i)(\phi_{ij} - \bar{\phi}_i)^{\mathrm{T}} = \mathbf{\Phi}(\mathbf{I} - 2\mathbf{A}_C + \mathbf{A}_C\mathbf{A}_C)\mathbf{\Phi}^{\mathrm{T}} \end{aligned} \tag{4}$$

where $\bar{\phi}$ is the mean of training samples, $\bar{\phi}_i$ is the mean of ith class, $\mathbf{A}$ is an (NxN) matrix with all terms equal to 1/N, $\mathbf{A}_C$ is an (NxN) block diagonal matrix where the ith diagonal term is an ($\mathrm{N_i}$x$\mathrm{N_i}$) matrix being filled with $1/\mathrm{N_i}$, and $\mathrm{N_i}$ is the number of the i^{th} class samples. These representations of $\mathbf{T}$ and $\mathbf{W}$ are simpler than those in [2] via omitting the redundant centering procedure in feature space, but yield exactly the same results. Then we can rewrite the above criterion in (2) as:

$$\begin{aligned} \lambda &= \frac{\alpha^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}\mathbf{\Phi}\alpha}{\alpha^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{W}\mathbf{\Phi}\alpha} = \frac{\alpha^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{I}-2\mathbf{A}+\mathbf{A}\mathbf{A})\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\alpha}{\alpha^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{I}-2\mathbf{A}_C+\mathbf{A}_C\mathbf{A}_C)\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\alpha} \\ &= \frac{\alpha^{\mathrm{T}}\mathbf{K}(\mathbf{I}-2\mathbf{A}+\mathbf{A}\mathbf{A})\mathbf{K}\alpha}{\alpha^{\mathrm{T}}\mathbf{K}(\mathbf{I}-2\mathbf{A}_C+\mathbf{A}_C\mathbf{A}_C)\mathbf{K}\alpha} = \frac{\alpha^{\mathrm{T}}\mathbf{T}'\alpha}{\alpha^{\mathrm{T}}\mathbf{W}'\alpha} \end{aligned} \tag{5}$$

where $\mathbf{K}$ is an (NxN) kernel matrix consisting of the inner products of samples in the feature space wherein each matrix element is given by: $k_{i,j} = \phi_i \cdot \phi_j = k(x_i, x_j)$. Now, the criterion has been transformed into another eigenvalue

© IEICE 2009
DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009

problem of the newly defined tractable matrices $\mathbf{T}'$ and $\mathbf{W}'$ with eigenvector α.

To solve the eigenvalue problem, $\mathbf{W}'$ needs to be inversed. Here, we adopt the ER scheme [4] to treat the inverse problem reliably and without any loss of dimension. First, we diagonalize $\mathbf{W}'$ via a singular value decomposition (SVD), and then sort its eigenvalues $(\lambda_i^{\mathrm{W}})_{i=1,\ldots,\mathrm{N}}$ and the corresponding eigenvectors $(v_i^{\mathrm{W}})_{i=1,\ldots,\mathrm{N}}$ in a decreasing order. By thresholding the eigenvalues based on λ_m^{W} and λ_r^{W} (the minimum reliable eigenvalue and the smallest nonzero eigenvalue), ER separates the space spanned by $\mathbf{W}'$ into three subspaces: a reliable subspace (face space), an unstable subspace (noise space), and a null subspace [4]. r is set to the maximum integer satisfying $\lambda_r^{\mathrm{W}} > e$ where e is a very small value compared to λ_1^{W}, and m is set to an integer satisfying $\lambda_{m+1}^{\mathrm{W}} = \max\left\{\forall\lambda_i^{\mathrm{W}} \mid \lambda_i^{\mathrm{W}} < (\lambda_{med}^{\mathrm{W}} + \mu(\lambda_{med}^{\mathrm{W}} - \lambda_r^{\mathrm{W}}))\right\}$ where $\lambda_{med}^{\mathrm{W}} = median\left([\lambda_i^{\mathrm{W}}]_{i=1,\ldots,r}\right)$. The transition of the eigenvalues of the face space is then modeled by a reciprocal function $\alpha_{er}/(i+\beta_{er})$. The model parameters α_{er} and β_{er} are calculated substituting the first and the last eigenvalues of the face space, λ_1^{W} and λ_m^{W}:

$$\alpha_{er} = \frac{\lambda_1^{\mathrm{W}}\lambda_m^{\mathrm{W}}(m-1)}{\lambda_1^{\mathrm{W}} - \lambda_m^{\mathrm{W}}} \quad \text{and} \quad \beta_{er} = \frac{m\lambda_m^{\mathrm{W}} - \lambda_1^{\mathrm{W}}}{\lambda_1^{\mathrm{W}} - \lambda_m^{\mathrm{W}}}.$$

Then the eigenvalues are regularized separately for each subspace:

$$\tilde{\lambda}_i = \begin{cases} \lambda_k, & i \le m & \text{(face space)} \\ \alpha_{er}/(i+\beta_{er}) & m < i \le r & \text{(noise space)}\ . \\ \alpha_{er}/(r+1+\beta_{er}) & r < i \le N & \text{(null space)} \end{cases} \tag{6}$$

Since available dimensionality is limited to the sample space in kernel implementation, the range of the null space of (6) is modified to be upper bounded by the number of samples N, and not by the size of training image as that in [4]. Using the regularized eigenvalues ($\tilde{\mathbf{\Lambda}}_{\mathrm{W}} = \mathrm{diag}[\tilde{\lambda}_1^{\mathrm{W}}, \ldots, \tilde{\lambda}_N^{\mathrm{W}}]$) and the eigenvectors ($\mathbf{V}_{\mathrm{W}} = [v_1^{\mathrm{W}}, \ldots, v_N^{\mathrm{W}}]$), we can compose a regularized within-class scatter matrix $\tilde{\mathbf{W}}' = \mathbf{V}_{\mathrm{W}}\tilde{\mathbf{\Lambda}}_{\mathrm{W}}\mathbf{V}_{\mathrm{W}}^{\mathrm{T}}$. This regularization substantially raises the small eigenvalues which are susceptible to sample noises. Consequently, the corresponding eigenvectors are de-emphasized for increasing overall stability. Also, the regularization on zero eigenvalues allows the usage of null-space.

The inverse problem of $\tilde{\mathbf{W}}'$ can be solved by a whitening procedure. The whitening transform matrix is defined as $\mathbf{P} = \mathbf{V}_{\mathrm{W}}\tilde{\mathbf{\Lambda}}_{\mathrm{W}}^{-1/2}$ where $\tilde{\mathbf{W}}'$ is whitened as: $\mathbf{P}^{\mathrm{T}}\tilde{\mathbf{W}}'\mathbf{P} = (\tilde{\mathbf{\Lambda}}_{\mathrm{W}}^{-1/2}\mathbf{V}_{\mathrm{W}}^{\mathrm{T}})(\mathbf{V}_{\mathrm{W}}\tilde{\mathbf{\Lambda}}_{\mathrm{W}}\mathbf{V}_{\mathrm{W}}^{\mathrm{T}})(\mathbf{V}_{\mathrm{W}}\tilde{\mathbf{\Lambda}}_{\mathrm{W}}^{-1/2}) = \mathbf{I}$. Since $\mathbf{P}$ is a full-rank matrix, there exists a unique solution α satisfying $\alpha = \mathbf{P}\alpha'$. Substituting these representations into (5), we can diagonalize $\tilde{\mathbf{W}}'$ and collapse the denominator by $\alpha'^{\mathrm{T}}\alpha' = 1$:

$$\frac{\alpha^{\mathrm{T}}\mathbf{T}'\alpha}{\alpha^{\mathrm{T}}\mathbf{W}'\alpha} = \frac{\alpha'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{T}'\mathbf{P}\alpha'}{\alpha'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\tilde{\mathbf{W}}'\mathbf{P}\alpha'} = \frac{\alpha'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{T}'\mathbf{P}\alpha'}{\alpha'^{\mathrm{T}}\mathbf{I}\alpha'} = \alpha'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{T}'\mathbf{P}\alpha'. \tag{7}$$

Here, we have a regular form of eigenvalue problem with the new eigenvector α' instead of α of (5). By solving the eigenvalue problem, we can find α', and the discriminant feature vector v is derived as:

$$v = \mathbf{\Phi}\alpha = \mathbf{\Phi}\mathbf{P}a'. \tag{8}$$

DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009

For an arbitrary input y, its projection onto the subspace spanned by the feature vectors, v, can be computed as:

$$z = v^{\mathrm{T}}\phi(y) = a'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}\phi(y)) = a'^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{K}_y \tag{9}$$

where $\mathbf{K}_y$ is a (Nx1) kernel matrix with its ith term given by $\phi_i \cdot \phi(y) = k(x_i, y)$.

To summarize, we introduced several detailed problems of KDA which are differentiated from those of LDA, and applied a two-fold regularization procedure to alleviate those problems. Meanwhile, we made several modifications both on the ER procedure to compute the scatter matrices for our implementation of R-KDA.

3 Experiments

In order to evaluate the proposed method, the AR database [6] is adopted in this study. A total of 1680 face images from 120 identities have been randomly divided into two subsets without any common identity. We consider only the verification scenario where the results are reported in terms of the average equal error rates (EERs) measured from the two-fold cross validation. The proposed R-KDA will be compared with three other variants of Fisher discriminant: generalized discriminant analysis (GDA) [1], kernel direct discriminant analysis (KDDA) [2], and ERE [4].

Two experiments will be performed in this evaluation. The first experiment is to compare among the above mentioned kernel-based methods. A polynomial kernel function was adopted and the comparison was performed by varying a crucial kernel parameter in:

$$k(x_i, x_j) = (w(x_i \cdot x_j) + b)^D \tag{10}$$

For simplicity, we fixed the bias parameter (b) to one and the degree parameter (D) to three, and changed only the weight parameter (w). To compare each algorithm with an individual value of D, the optimal number of feature vectors which show the best performance has been used. Fig. 1 (a) shows the error rates plotted over w varying ranging from 10^{-9} to 10^{3}. The ERE [4] cannot be included in this comparison because no kernel function has been adopted in its original form.

As seen from Fig. 1 (a), the performance trends of KDDA and GDA appear complementary at the two extremes of w values, whereas the proposed method shows the best performance over the entire range of kernel parameter setting. This stable performance of R-KDA can be attributed to the leveraging of advantageous strategies of both methods by appropriate regularization. Especially, R-KDA well suppresses the sample noises, its EERs fluctuate less over w values than those of KDDA and GDA.

In the second experiment, we test all four algorithms by varying the number of feature vectors. The kernel parameter w is set to its optimum value for each method. The results are shown in Fig. 1 (b). Since both

© IEICE 2009
DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009

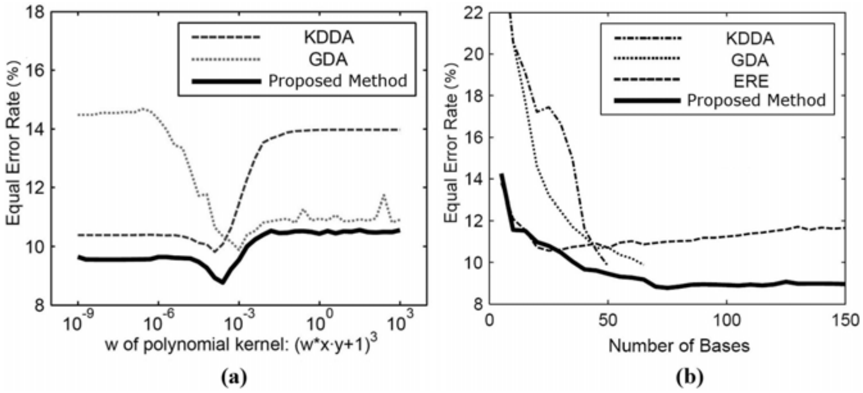

Fig. 1. Equal error rates (a) as functions of the kernel parameter, and (b) as functions of the number of feature vectors

KDDA and GDA discard those small eigenvalues and use only the conventional between-class scatter, the available dimensionality is much limited. As seen in Fig. 1 (b), this dimensionality limitation confines the performance of the algorithms. On the other hand, R-KDA utilizes the entire sample space to find the discriminant vectors. Although the maximum rank of between-class scatter is limited to 59, since data from 60 different identities have been used in our experiment, the EER of R-KDA decreases even when the number of basis vectors used is larger than 60. This result evidences the effectiveness of R-KDA which comes from the alleviation of the dimensionality limit allowing a better extraction of useful information from the training data. The effectiveness is also evidenced by the observation that R-KDA showing a much smaller recognition error rate in the range of small feature sizes. The original ERE shows a similar trend with the proposed method, but its best performance seems to be limited by its linearity nature.

4 Conclusion

Despite of its effectiveness in nonlinear classification, the kernel expansion of Fisher's criterion carries two major drawbacks: sample noise and lack of dimensionality. Attributed to a two-fold regularization on both the between-class scatter and the within-class scatter, the proposed algorithm successfully alleviated these problems with an improvement over the conventional kernel Fisher discriminant.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University (R112002105070030(2008)), and by the Basic Research Program of the Korea Science & Engineering Foundation (R01-2006-000-11016-0).

DOI: 10.1587/elex.6.540
Received February 18, 2009
Accepted April 07, 2009
Published May 10, 2009