REGULARIZED LOCAL DISCRIMIMANT EMBEDDING

Yanwei Pang[1,2], *Nenghai Yu*[1,2]

[1]Department of Electronic Engineering & Information Science, University of Science and Technology of China, Hefei 230027, China
[2]National Laboratory of Pattern Recognition, Information of Automation, Chinese Academy of Sciences, Beijing 100080, China
E-mail: pyw@ustc.edu.cn

ABSTRACT

Recently, Chen *et al*. (CVPR 2005) proposed a new manifold embedding method, *Local Discriminant Embedding* (*LDE*), which utilizes the neighbor and class relations of data to construct the embedding for classification. While having powerful classification ability, LDE suffers from *small size sample* problem, which leads to unstably numerical computation. To deal with this problem, we propose to a method of regularized LDE (RLDE) by imposing additional regularizing constraints on LDE. Experimental results show the effectiveness of the proposed method.

1. INTRODUCTION

Subspace-based face recognition method aims to find a low dimensional subspace of face appearance embedded in a high dimensional image space. The differences between different subspace-based methods lie in their different motivations and objective (or cost) functions. Eigenface, the underlying idea which is the *Principal Component Analysis* (PCA) [1], seeks a subspace that best represents the data in a least-squares sense. Therefore, the feature extracted by PCA is called the most expressive feature [2]. Fishferface, the underlying idea of which is the *Linear Discriminant Analysis* (LDA) [3], selects a linear transformation matrix in such a way that the ratio of the between-class scatter to with-class scatter is maximized. Therefore, the feature extracted by LDA is called the most discriminant feature [2].

Recently, several manifold learning algorithms were developed: *locally linear embedding* (LLE) [4], *Isomap* [5], *Laplacian Eigenmaps* [6]. They all utilize local neighborhood information to construct a global embedding of the manifold. Using Nyström formula, one can extent them to be able to map new test points [7]. Another way to apply these algorithms to new points is to introduce a linear transformation matrix to relate input with output. *Locality Preserving Projections* (LPP) [8] and *Neighborhood Preserving Projections* (NPP) [9] are the results of linear generalization of Laplacian Eigenmaps and LLE respectively. LPP and NPP can be categorized into subspace learning algorithm. Compared with PCA, LPP and NPP preserve the local structure instead of the global structure of the image space. However, this property does not necessarily mean optimal classification. Moreover, they are unsupervised learning methods, so the information carried by class labels is lost. More recently, *Local Disriminant Embedding* (LDE) [10] and *marginal fisher analysis* (MFA) [11] were proposed to overcome the drawbacks of LPP. LDE and MFA were developed by different researchers, but the underlying ideas of which are almost the same: the neighbor and class relations of data are utilized to construct the face space (subspace of the image space). Compared with LDE, MFA and LDE do not depend on the assumption that the data of each class is Gaussian distributed.

Despite its advantages, LDE suffers from the small sample size (SSS) problem. Small sample size problem occurs when there are few training samples compared to sample dimension, as often encountered in tasks such as face recognition. With this problem, LDE is involved in eigen-decomposition with singular matrix, which leads to unstably numerical computation. Though one can deal with this difficulty by employing PCA to perform dimension reduction for the data before conducting LDE, some useful information may be lost by discarding minor components and maintaining only principal components. This problem was addressed in various extensions of LDA, such as direct LDA [12] and dual-space LDA [13]. In this paper, we provided another way to avoid matrix being singular in LDE. In the method (RLDE), generalized eigenvalue problem is converted to standard eigenvalue problem without having to compute inverse matrix. The method is demonstrated with face recognition where several methods are compared. Results show that RLDE can not only handle the SSS problem, but also enhance the recognition accuracy.

The reminder of this paper is organized as follows. In Section 2, *Laplacian Eigenmaps*, which is the basis of LDE, is described. Then LDE is given in Section 3. The proposed method is presented in Section 4. The proposed method is evaluated on the AR face database [14] in Section 5. Finally, conclusion is drawn in Section 6.

2. LAPLACIAN EIGENMAPS

Because LDE is fundamentally based on Laplacian Eigenmaps, we will give a brief description of Laplcacian Eigenmaps first.

Laplacian Eigenmaps is a geometrically motivated algorithm for constructing a representation for data sampled from a low dimensional manifold embedded in a higher dimensional space [6]. Let $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$ be a data set of D-dimensional vectors. Dimension reduction is conducted to map these points (vectors) to be new points $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$ in a d-dimensional space where $d<<D$. The objective function of Laplacian Eigenmaps is to maximize

$$J(\mathbf{Y}) = \sum_{i,j} \| \mathbf{y}_i - \mathbf{y}_j \|^2 \, w_{ij} \tag{1}$$

under appropriate constraints. Weight w_{ij} are defined as follows. If $\mathbf{x}_j$ is among k nearest neighbors of $\mathbf{x}_i$, then $w_{ij}=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/t)$, otherwise $w_{ij}=0$.

Weights w_{ij} give a heavy penalty if neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Therefore, minimizing J ensures that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close then $\mathbf{y}_i$ and $\mathbf{y}_j$ are close as well [8].

3. LOCALLY DISCRIMINANT EMBEDDING (LDE)

Laplacian Eigenmaps is an unsupervised manifold learning algorithm, whereas LDE [10] is a supervised subspace learning algorithm. Therefore, class label l_i of $\mathbf{x}_i$ ($i=1,\ldots N$) are used in LDE to determine a linear transformation matrix $\mathbf{U}$ such that

$$\mathbf{y}_i = \mathbf{U}^T \mathbf{x}_i \,. \tag{2}$$

The column vectors of $\mathbf{U}=[\mathbf{u}_1 \ \mathbf{u}_2 \ \ldots \ \mathbf{u}_d]$ span a d-dimensional subspace. The aim of LDE is to, in the low subspace, keep neighboring points close if they have the same class label, whereas prevent points of other classes from entering the neighborhood [10]. Its objective is to maximize the function

$$J_{LDE}(\mathbf{U}) = \sum_{i,j} \| \mathbf{U}^T \mathbf{x}_i - \mathbf{U}^T \mathbf{x}_j \|^2 \, w'_{ij} \tag{3}$$

subject to

$$\sum_{i,j} \| \mathbf{U}^T \mathbf{x}_i - \mathbf{U}^T \mathbf{x}_j \|^2 \, w_{ij} = 1 \,, \tag{4}$$

where weight w_{ij} and w'_{ij} are defined as follows. If $\mathbf{x}_j$ is among k nearest neighbors of $\mathbf{x}_i$ and $l_i \neq l_j$ then $w'_{ij}=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/t)$, otherwise $w'_{ij}=0$. If $\mathbf{x}_j$ is among k nearest neighbors of $\mathbf{x}_i$ and $l_i=l_j$ then $w_{ij}=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/t)$, otherwise $w_{ij}=0$.

Because

$$\begin{aligned}
&\sum_{i,j} \| \mathbf{U}^T \mathbf{x}_i - \mathbf{U}^T \mathbf{x}_j \|^2 w_{ij} \\
&= \sum_{i,j} tr[(\mathbf{U}^T \mathbf{x}_i - \mathbf{U}^T \mathbf{x}_j)(\mathbf{U}^T \mathbf{x}_i - \mathbf{U}^T \mathbf{x}_j)^T] w_{ij} \\
&= \sum_{i,j} tr[(\mathbf{U}^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{U}] w_{ij} \\
&= tr\{\mathbf{U}^T [\sum_{i,j} tr(\mathbf{x}_i - \mathbf{x}_j) w_{ij} (\mathbf{x}_i - \mathbf{x}_j)^T] \mathbf{U}\} \\
&= 2tr[\mathbf{U}^T \mathbf{X}(D-W)\mathbf{X}^T \mathbf{U}] \,,
\end{aligned}$$

The objective function (3) and the constraint in (4) can be reformulated as

$$\text{maximizing } J_{LDE}(\mathbf{U}) = 2tr[\mathbf{U}^T \mathbf{X}(D'-W')\mathbf{X}^T \mathbf{U}] \tag{5}$$

$$\text{subject to } 2tr[\mathbf{U}^T \mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T \mathbf{U}] = 1. \tag{6}$$

The optimization can be reduced to the following generalized eigenvalue problem:

$$\mathbf{X}(\mathbf{D}'-\mathbf{W}')\mathbf{X}^T \mathbf{u} = \lambda \mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T \mathbf{u} \,, \tag{7}$$

where the elements of the matrix $\mathbf{W}'$ are w'_{ij}, the elements of the matrix $\mathbf{W}$ are w_{ij}. The elements of diagonal matrices $\mathbf{D}$ and $\mathbf{D}'$ are defined as $d_{ii} = \sum_j w_{ij}$ and $d'_{ii} = \sum_j w'_{ij}$ respectively.

Defining $\mathbf{A} \triangleq \mathbf{X}(\mathbf{D}'-\mathbf{W}')\mathbf{X}^T$ and $\mathbf{B} \triangleq \mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T$, we have

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{B}\mathbf{u} \,, \tag{8}$$

or

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{u} = \lambda \mathbf{u} \,. \tag{9}$$

The above formulation can be problematic when the sample size is small. In this case, $\mathbf{B}$ becomes singular and the computation of (9) is unstable. One possible solution is to employ PCA to perform dimensionality reduction before conducting LDE; however, some useful information may be lost as a consequence. If considerable principal components are discarded, then information is lost not only in the sense of reconstruction but also in recognition. Therefore, we propose a method called *Regularized LDE* (RLDE) to solve this problem.

4. REGULARIZED LDE (RLDE)

Now we discuss Laplacian Eigenmaps and LDE, then describe our idea to improve LDE.

4.1. The basic idea

Fig. 1 (a) illustrates the local neighborhood relationship of the original data. The hollow circles ($\mathbf{x}_1^1$, $\mathbf{x}_2^1$, $\mathbf{x}_3^1$, and $\mathbf{x}_?^1$) belong to one class (class #1) and the solid circles ($\mathbf{x}_1^2$, $\mathbf{x}_2^2$, and $\mathbf{x}_3^2$) belong to another class (class #2). Suppose the distances from $\mathbf{x}_?^1$ to any other points are all equal. The points in Fig. 1 (a) are mapped by Laplacian Eigenmaps (or LPP), and the corresponding new points ($\mathbf{y}_1^1$, $\mathbf{y}_2^1$, $\mathbf{y}_3^1$, $\mathbf{y}_?^1$, $\mathbf{y}_1^2$, $\mathbf{y}_2^2$, and $\mathbf{y}_3^2$), are shown in Fig. 1(b). Note that, without loss of generalization, we let the dimensionality of the new points be equal to that of the original points. The local structure of the original data (Fig. 1(a)) is kept exactly in Fig. 1(b). This is the so-called locality-preserving property [6][8]. However, it contributes little for pattern classification in this case. Suppose $\mathbf{x}_?^1$ is a

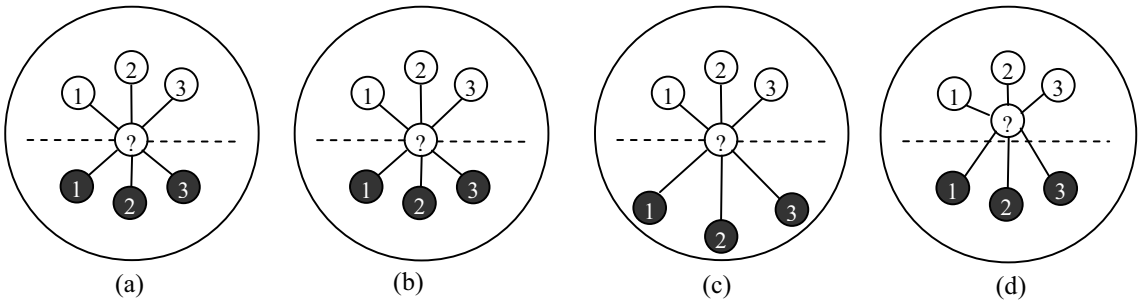

Fig. 1 Data embedding in a local structure: (a) original data; (b) embedded by Laplacian Eigenmaps; (c) embedded by LDE; (d) embedded by RLDE (the proposed method)

probe point, one can find that the distances from its corresponding point, $\mathbf{y}_?^1$, to any other points belonging to either class #1 or class #2 are equal as they are in Fig. 1(a). It is hard to say which class $\mathbf{y}_?^1$ should be classified into.

Fig 1 (c) shows the result of LDE. We find that the distances from $\mathbf{y}_?^1$ to the points of class #1 ($\mathbf{y}_1^1$, $\mathbf{y}_2^1$, and $\mathbf{y}_3^1$) are maintained, whereas the distances to the points of class #2 ($\mathbf{y}_1^2$, $\mathbf{y}_2^2$ and $\mathbf{y}_3^2$) are stretched. The probe point, $\mathbf{y}_?^1$ (corresponding to $\mathbf{x}_?^1$), can thus be well separated. This phenomena stems from the optimization process of LDE: maximizing the distances between the points of the same class while constraining the distances between different classes (see Eq. (3) and (4)).

Though LDE has powerfully discriminating ability in theory, its computation (Eq. (8) and (9)) is not stable when B is not full rank (*i.e.* singular). One can deal with this difficulty by employing PCA to perform dimension reduction for the data before conducting LDE, but some useful information may be lost by discarding minor components and maintaining only principal components.

Fig. 1(d) illustrates the idea of the proposed method, *RLDE*. Notice the position of the point, $\mathbf{y}_?^1$, we can found that its distances to class #1 are shrunk and, meanwhile, the distances to class #2 are stretched. From the classification point of view, the effect of RLDE is equivalent in ideal condition. However, in the subsection 4.2, one can found that small size sample problem does not exist in RLDE. All the information can be utilized and thus RLDE is superior to LDE.

4.2. Regularized LDE (*RLDE*)

To realize the idea of RLDE, we modified the objective function and the constraint of the LDE. The optimization problem becomes then

$$J_{RLDE}(\mathbf{U}) = \sum_{i,j} \| \mathbf{U}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{x}_j \|^2 w'_{ij} - \sum_{i,j} \| \mathbf{U}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{x}_j \|^2 w_{ij} \quad (10)$$

subject to

$$\mathbf{u}^T\mathbf{u} = 1 . \quad (11)$$

From (10), one can find that the larger the value of the first item (of the right part of (10)) is and, at the same time, the less of the second item is, the larger the J_{RLDE} is. The first item is the weighted squared distance between neighboring points belonging to different classes. In contrast to the first item, the second item is the weighted squared distance between neighboring points belonging to the same classes. Therefore, in a local manifold structure (take Fig. 1 (d) for example), points of the same class will move towards a compact cluster, and those of different classes can be separated more reliably.

The orthogonal constraint (Eq. (11)) is introduced to deal with the ill-posed problem when maximizing (10).

Eq. (10) can be rewritten as (refer to Section 3)

$$J_{RLDE}(\mathbf{U}) = 2tr[\mathbf{U}^T\mathbf{X}(\mathbf{D}'-\mathbf{W}')\mathbf{X}^T\mathbf{U}] - 2tr[\mathbf{U}^T\mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T\mathbf{U}]$$

The constrained maximization can then be done using the method of Largrange multipliers:

$$\begin{aligned} L(\mathbf{u}_i) &= \mathbf{u}_i^T[\mathbf{X}(\mathbf{D}'-\mathbf{W}')\mathbf{X}^T]\mathbf{u}_i \\ &\quad -\mathbf{u}_i^T[\mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T]\mathbf{u}_i + \lambda(1-\mathbf{u}_i^T\mathbf{u}_i) \end{aligned} \quad (12)$$

Compute the gradients with respect to $\mathbf{u}_i$ and set the gradients to zero, we have the following eigenvalue problem:

$$[\mathbf{X}(\mathbf{D}'-\mathbf{W}')\mathbf{X}^T - \mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T]\mathbf{u}_i = \lambda_i\mathbf{u}_i \quad (13)$$

with $\lambda_1<\lambda_2<\ldots<\lambda_d$. If we employ the definition of **A** and **B** in Section 2, (13) can be rewritten as

$$(\mathbf{A}-\mathbf{B})\mathbf{u}_i = \lambda_i\mathbf{u}_i \quad (14)$$

By defining **C**=**A**-**B**, (14) can be reformulated as

$$\mathbf{C}\mathbf{u}_i = \lambda_i\mathbf{u}_i \quad (15)$$

Note that (15) is a standard eigenvalue problem, instead of the generalized eigevalue problem in (9). Because (15) does not involve in inverse matrix, the proposed method, RLDE, the small size sample problem can be avoided.

5. EXPERIMENTAL RESULTS

The proposed RLDE method is demonstrated in comparison to several contrasted methods including PCA [1], LPP [8], LDA [3], and LDE [10]. The AR face database [14] was used to evaluate the proposed method. The nearest neighborhood classifier was employed in the experiments.

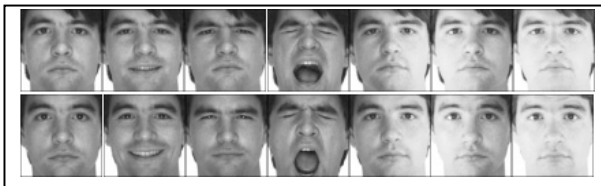

Fig.2 Example images of one subject used in the experiments. Top: taken in the 1st session; Bottom: taken in the 2nd session

Table 1: Performance comparison on AR database (G2/P5) (%)

DB	PCA	LPP	LDA	LDE	RLDE
#1	74.01	68.37	75.38	75.21	78.11
#2	75.50	70.94	77.43	74.35	80.68
#3	80.34	68.88	77.26	82.39	83.07
#4	76.92	68.71	78.46	80.34	80.85
#5	77.77	74.52	82.56	83.32	86.21

Table 2: Performance comparison on AR database (G7/P7) (%)

PCA	LDA	LDE	RLDE
77.53	92.87	93.27	93.62

In the first experiment, 117 subjects are selected from a total of 126 subjects. Only 7 nonoccluded images per person from the first session are used in our experiments (see top line of Fig. 2). The images were cropped based on the centers of eyes, and the cropped images were resized to 60×60 pixels. Then they were normalized to have zero mean and unit variance. Two images of each subject were randomly chosen for training, while the remaining five images were used for testing. In this way, we ran the system 5 times and obtained 5 different training and testing sets. The recognition rates were found by averaging the recognition rate of each run (table 1).

From table 1, we find the three supervised learning methods, LDA, LDE, and RLDE, outperforms the unsupervised learning methods, PCA and LPP. Among the supervised methods, the proposed method achieves the highest recognition rates. Furthermore, RLDE is not prone to overfitting. LDA is not always superior to PCA due to overfitting. The number of principal component, s, in LDA (PCA plus LDA) is at most N-C = 117 (C is class number). In LDE the maximum value of s is around 80. But in RLDE s can be maximum value N-1=233. So no information is lost in RLDE (80<<233) in the reconstruction sense. Note that we can perform RLDE on the original data, but the computational time is very large.

In the second experiment, 14 images per subject were used (see Fig. 2). Seven images of each subject were randomly chosen for training, while the remaining seven images were used for testing. In this way, we ran the system 5 times and obtained 5 different training and testing sets. The recognition rates (Table 2) were found by averaging the recognition rate of each run. We find that RLDE has the highest recognition rate. Comparing table 1 with 2, we might conclude that RLDE outperforms LDE significantly when the *small size sample* problem is critical.

6. CONCLUSIONS

We have presented a subspace learning method, called RLDE, to deal with the small sample problem in LDE. The proposed method utilizes the neighbor and class relations of data to compute the subspace for classification. In comparison to LDE, the computation of RLDE is more stable. Because more information of the data is reserved in RLDE, it outperforms LDE in terms of recognition rate. For future work, we will generalize RLDE into feature space by kernel trick.

REFERENCES

[1] M. Turk, A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, 1991.

[2] D.L. Swets and J Wen, "Using discriminant eigenfeatures for image retrieval," *IEEE PAMI*, Vol. 18, No. 8, pp. 831-836,

[3] P. N. Belhumeur, J.P. Hespanha, D.J. Kriengman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projections," *IEEE PAMI*, Vol. 19, No. 7, pp. 771-720, 1997.

[4] S. Roweis and K. Saul, "Nonlinear Dimension Reduction by Locally Linear Embedding," *Science*, Vol. 290, No. 5500, pp. 2323-2326, 2000.

[5] J. Tenenbaum, V. Silva, J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, Vol. 290, pp. 2319-2322, 2000.

[6] M. Belkin, P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, Vol. 14, pp. 1373-1396, 2003.

[7] Y. Bengio, O. Delalleau, N. Roux, J. Paiement, P. Vincent, M. Ouimet, "Learning eigenfunctions links spectral embedding and kernel PCA," *Neural Computation*, Vol. 6, No. 10, pp. 2197-2219, 2004.

[8] X. He, S. Yan, Y. Hu, H. Zhang, "Learning a locality preserving subspace for visual recognition, *ICCV*, 2003.

[9] Y. Pang, L. Zhang, Z. Liu, N. Yu, H. Li, "Neighborhood preserving projections (NPP): a novel linear dimension reduction method," *LNCS*, Vol. 3644, Ⅰ, pp. 117-125, 2005.

[10] H. Chen, H. Chang, T. Liu, "Local discriminant embedding and its variants," *CVPR*, 2005.

[11] S. Yan, D. Xu, B. Zhang, H. Zhang, "Graph embedding: a general framework for dimensionality reduction," *CVPR*, 2005.

[12] H. Yu, J. Yang, "A direct LDA algorithm for high-dimensional data – with application to face recognition," Pattern Recognition, Vol. 34, pp. 2067-2070, 2001.

[13] X. Wang, X. Tang, "Dual-space linear discriminant analysis for face recognition," CVPR, 2004.

[14] A. M. Martinez, R. Benavente, "The AR face database," CVC Technical Report #24, Computer Vision Center at the U.A.B. June 1998.