# k-Means Discriminant Maps for Data Visualization and Classification

Vo Dinh Minh Nhat
Ubiquitous Computing Lab
Kyung Hee University
Suwon, Korea
vdmnhat@uclab.khu.ac.kr

SungYoung Lee
Ubiquitous Computing Lab
Kyung Hee University
Suwon, Korea
sylee@uclab.khu.ac.kr

## ABSTRACT

Over the years, many dimensionality reduction algorithms have been proposed for learning the structure of high dimensional data by linearly or non-linearly transforming it into a low-dimensional space. Some techniques can keep the local structure of data, while the others try to preserve the global structure. In this paper, we propose a linear dimensionality reduction technique that characterizes the local and global properties of data by firstly applying k-means algorithm on original data, and then finding the projection by simultaneously globally maximizing the between-cluster scatter matrix and locally minimizing the within-cluster scatter matrix, which actually keeps both local and global structure of data. Low complexity and structure preserving are two main advantages of the proposed technique. The experiments on both artificial and real data sets show the effectiveness and novelty of proposed algorithm in visualization and classification tasks.

## Categories and Subject Descriptors

I.5.2 [**Computing Methodologies**]: Pattern Recognition—*Design Methodology.*

## General Terms

Algorithms, Design, Experimentation, Performance, Theory.

## Keywords

Dimensionality Reduction, k-Means, Manifold Learning, Linear Discriminant Analysis.

## 1. INTRODUCTION

The purpose of dimensionality reduction is to transform high dimensional data into a low-dimensional space, while retaining most of the underlying structure in the data. The reason for using dimensionality reduction is based on the fact that some features may be irrelevant and "intrinsic" dimensionality of data may be smaller than the number of features. Dimensionality reduction can also be used to visualize high dimensional data by transforming the data into two or three dimensions, thereby giving additional insight into the problem at hand. With the rapidly increasing demand on dimensionality reduction techniques, it is not surprising to see an overwhelming amount of research publications on this topic in recent years. In general, there are linear and nonlinear dimensionality reduction techniques. Linear dimensionality reduction methods include Principal Component Analysis (PCA) developed by Pearson (1901) and Hotelling (1933) [6][3], and Multi-Dimensional Scaling (MDS) by Torgerson (1952) and Shepard (1962) [10][8]. While PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space, MDS finds an embedding that preserves the inter-point distances, which is equivalent to PCA when the distances are Euclidean. Besides linear methods, there are also non-linear dimensionality reduction techniques which have been developed up-to-date. Kernel PCA (KPCA) [4] map inputs nonlinearly to a new space, then perform PCA. Laplacian Eigenmaps (LE) [1] preserve nearness relations as encoded by graph Laplacian. ISOMAP [9] assumes that the data lie on a (Riemannian) manifold and maps data to its low-dimensional representation in such a way that the geodesic distance between two date points is as close to the Euclidean distance between two respectively points in low-dimensional space as possible. Diffusion Maps (DM) [5] is based on defining a Markov random walk on the graph of the data. In the low-dimensional representation of the data, the pairwise diffusion distances are retained as well as possible. Locally Linear Embedding (LLE) [7] maps its inputs into a single global coordinate system of lower dimensionality by computing low-dimensional, neighborhood preserving embedding of high-dimensional inputs, and its optimization does not involve local minima. It actually recovers global nonlinear structure from locally linear fits. Due to the length of paper some other techniques most of which are variants of above reviewed techniques are not covered here. In this paper, we propose a linear dimensionality reduction technique called k-Means Discriminant Maps (kDM). The algorithm firstly applies k-Means to cluster the original data, then for the purpose of keeping both local and global structure of data, it try to find a desirable projection that simultaneously minimizes the within-cluster scatter and maximizes the between-cluster scatter matrices. Some main contributions of proposed algorithm can be described as: low com-


*SAC'08* March 16-20, 2008, Fortaleza, Ceará, Brazil

plexity due to its linear property, and keeping both local and global structure of data. The outline of this paper is as follows. The proposed method is described in Section 2. In Section 3, experimental results are performed on both artificial and real data sets to demonstrate the effectiveness of our method. Finally, conclusions are presented in Section 4.

# 2. K-MEANS DISCRIMINANT MAPS

The dimension reduction problem is, given a data set $\{x_1, x_2, ..., x_N\}$ where $x_i \in \Re^n$ , to find a set of points $\{y_1, y_2, ..., y_N\}$ where $y_i \in \Re^m$ and $(m \ll n)$, such that each $y_i$ "represents" its counterpart $x_i$. For the convenience of presentation, we denote the matrix $X = [x_1, x_2, ..., x_N]$ and correspondingly the matrix $Y = [y_1, y_2, ..., y_N]$. In this section, our emphasis is on the description of our proposed algorithms. Due to the paper length, those previous dimensionality reduction techniques can be referenced in respective literatures.

## 2.1 k-Means

The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function

$$f = \sum_{i=1}^{k} \sum_{x_j \in \Pi_i} \|x_j - \mu_i\|^2 \quad (1)$$

where there are $k$ clusters $\mu_i$, $i = 1, 2, ..., k$ and $\Pi_i$ is the centroid or mean point of all the points $x_j \in \Pi_i$.

## 2.2 k-Means Discriminant Maps - kDM

Assume that after clustering data by k-means algorithm each data sample belongs to one of $C$ cluster $\{\Pi_1, \Pi_2, ..., \Pi_C\}$. Let $N_i$ be the number of the samples in cluster $\Pi_i (i = 1, 2, ..., C)$, $\mu_i = \frac{1}{N_i} \sum_{x \in \Pi_i} x$ be the mean of the samples or the centroid in cluster $\Pi_i$. Then we define the between-cluster scatter matrix $S_b$ and the within-cluster scatter matrix $S_w$ as follow

$$S_b = \frac{1}{N} \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

$$S_w = \frac{1}{N} \sum_{i=1}^{C} \sum_{x_k \in \Pi_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (3)$$

For the purpose of keeping both local and global structure of data, we try is to find a projection which will draw the close samples (ones in the same cluster) closer together while simultaneously making the distant samples (ones from different clusters) even more distant from each other. From this point of view, a desirable projection should be the one that, at the same time, minimizes the within-cluster scatter and maximizes the between-cluster scatter matrices. So in kDM, the projection $W_{opt}$ is chosen to maximize the ratio of the determinant of the between-cluster scatter matrix of the projected samples to the determinant of the within-cluster scatter matrix of the projected samples, i.e.,

$$J(w) = \arg\max_w \frac{w^T S_b w}{w^T S_w w} \quad (4)$$

From the criterion in ( 4), we can find the projection by simultaneously globally maximizing the between-cluster scatter and locally minimizing the within-cluster scatter, which actually keep both local and global structure of data. It is also easy to realize that the criterion ( 4) is formally similar to the Fisher criterion since they are both Rayleigh quotients. However in kDM, we form the between-cluster scatter matrix $S_b$ and the within-cluster scatter matrix $S_w$ without knowing the class labels of samples. This means Fisher discriminant projection is supervised, while the projection determined by kDM can be obtained in an unsupervised manner. The optimal projection for kDM is $W = [w_1 w_2 ... w_m]$, where $\{w_i \,|i = 1, 2, ..., m\}$ is the set of generalized eigenvectors of $S_b$ and $S_w$ corresponding to the $m$ largest generalized eigenvalues $\{\lambda_i \,|i = 1, 2, ..., m\}$, i.e.,

$$\begin{array}{l} S_b w_i = \lambda_i S_w w_i \\ \Leftrightarrow S_w^{-1} S_b w_i = \lambda_i w_i \end{array} \quad i = 1, 2, ..., m \quad (5)$$

However in some cases, the dimension of the sample space is typically larger than the number of samples. As a consequence, $S_w$ is singular. This problem is known as the "small sample size (3S) problem"[2]. To solve this problem, we use the strategy of Direct-LDA[11] to implement our kDM algorithm in the case of 3S problem. The key idea of DLDA is to discard the null space of $S_b$, which contains no useful information, rather than discarding the null space of $S_w$, which contains the most discriminative information. $S_b$ is firstly diagonalized as $S_b = U \Lambda U^T$, where $U \in \Re^{n \times (C-1)}$ is a matrix whose columns are eigenvectors of $S_b$, $C$ is the number of classes (equivalent to the number of clusters $k$ in k-Means of kDM, we use $k$ and $C$ interchangeably) and $\Lambda$ is a diagonal matrix with eigenvalues. The new projected within scatter matrix is formed as

$$\tilde{S}_w = \Lambda^{-1/2} U^T S_w U \Lambda^{-1/2} \quad (6)$$

Let $\tilde{W}_w = [w_1 w_2 ... w_{C-1}]$, where $\{w_i \,|i = 1, 2, ..., C-1\}$ is the set of eigenvectors of $\tilde{S}_w$. Then, the optimal projection for DLDA is $W_{opt} = U \Lambda^{-1/2} \tilde{W}_w$.

| Technique | Parameter Settings |
|---|---|
| PCA | None |
| Kernel PCA | $\kappa = (XX^T + 1)^3$ |
| Diffusion Maps | $\sigma = 1$ |
| LLE | $k = 12$ |
| LE | $k = 12, \sigma = 1$ |
| ISOMAP | $k = 12$ |
| kDM | $k = 3$ |

**Table 1: Parameter settings for the experiments**

# 3. EXPERIMENTS

In this section, a systematic empirical experiments of the performance of previous techniques and our proposed technique kDM are performed. We perform the evaluation on two types of datasets: (1) artificial datasets and (2) real datasets (ORL face database and PolyU Palmprint database).

## 3.1 Data Visualization on Artificial Datasets

The artificial datasets on which we performed experiments are: (1) the Swiss roll dataset and (2) the intersecting dataset. Some parameters used in this part of experiments can be seen in Table 1. We perform PCA, Kernel PCA, Diffusion Maps, LLE, LE, ISOMAP and kDM on 1000 data points of

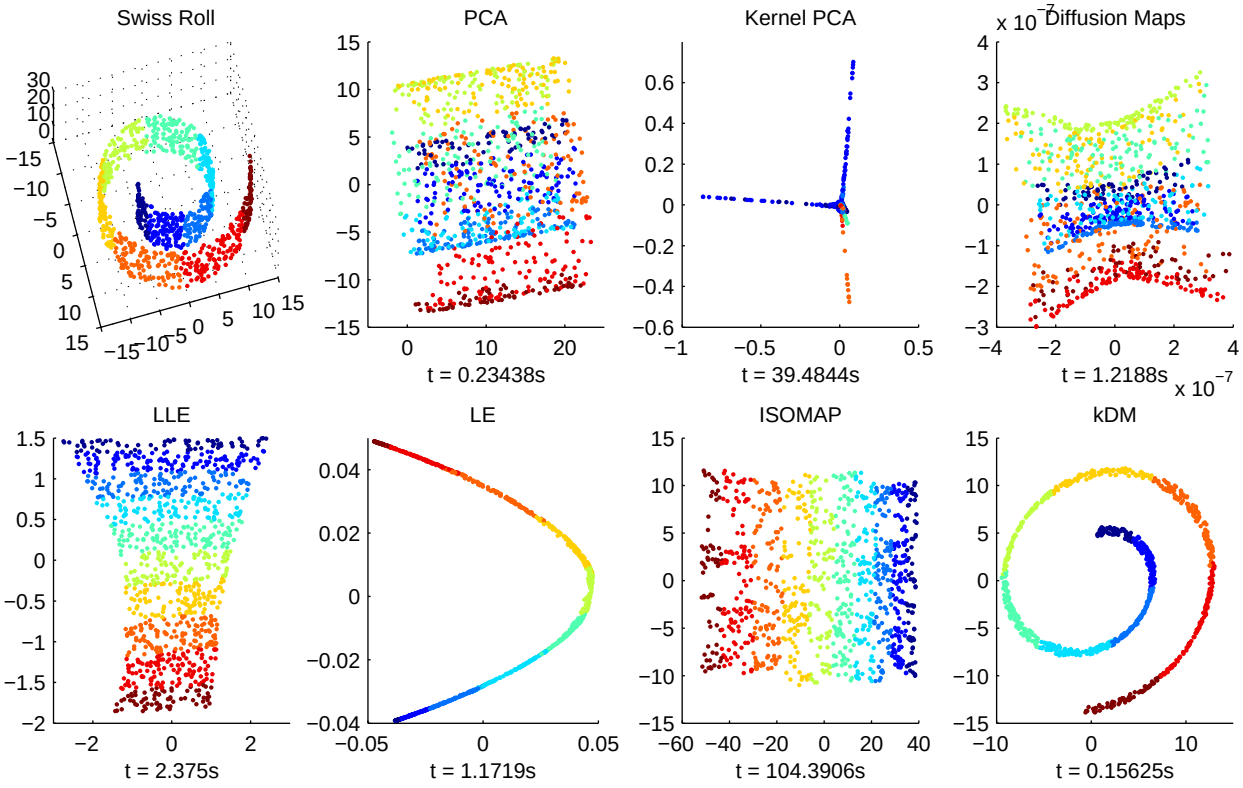


**Figure 1: Two-dimensionality Visualization of the Swiss roll dataset based on variety of techniques**

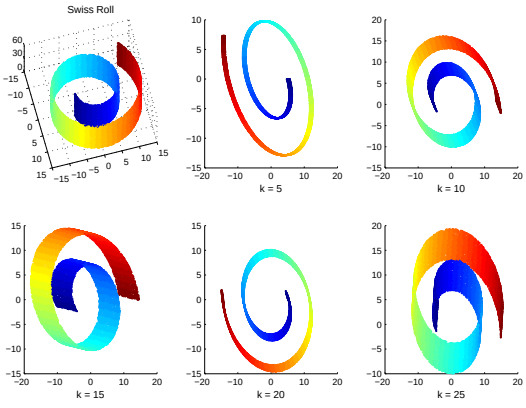


**Figure 2: Performance of kDM versus $k = 5, 10, 15, 20$ on Swiss roll dataset.**

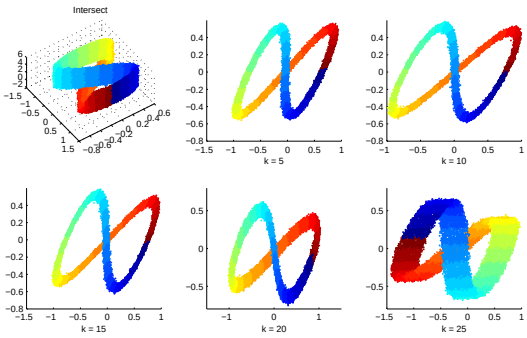


**Figure 3: Performance of kDM versus $k = 5, 10, 15, 20$ on Intersection dataset.**

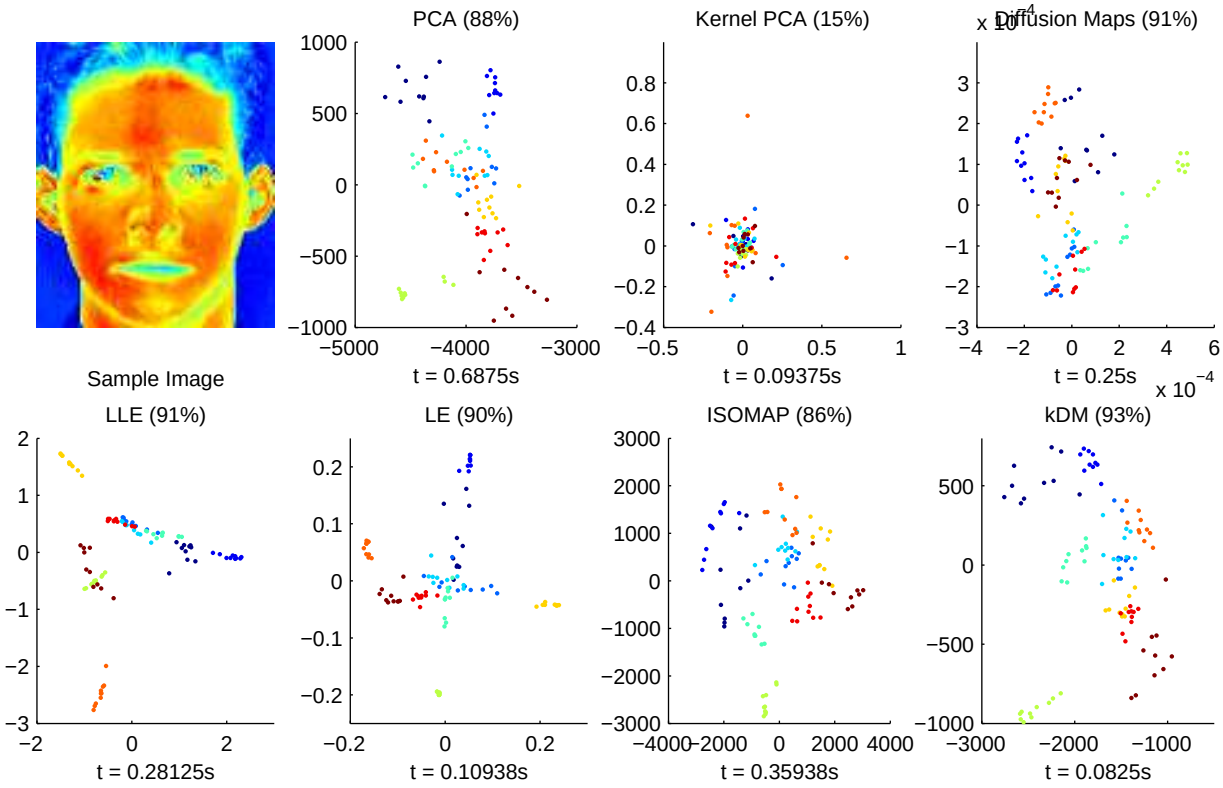


**Figure 4: Two-dimensionality Visualization of the ORL face database based on variety of techniques**

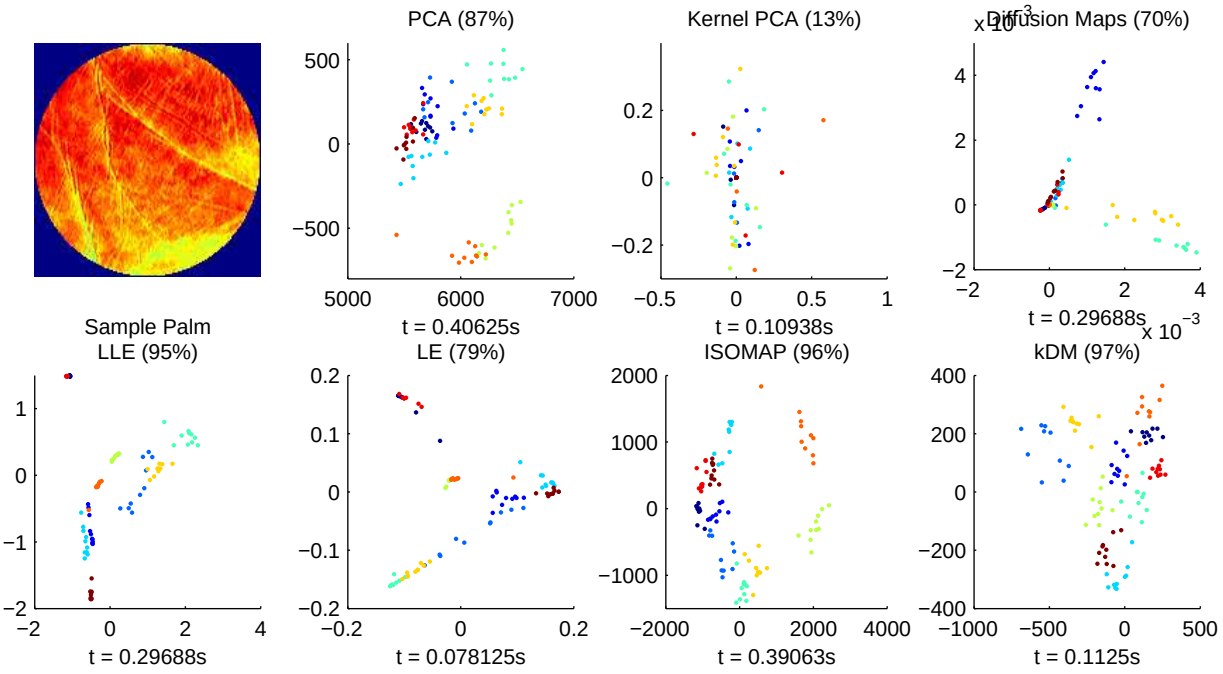


**Figure 5: Two-dimensionality Visualization of the PolyU Palmprint database based on variety of techniques**

Swiss roll dataset to show the two-dimensional representations of the Swiss roll dataset which can be seen in Fig. 1. From the depicted representations, we can see that PCA, Kernel PCA and Diffusion Maps techniques are not capable of successfully learning the 2-dimensional structure of the Swiss roll manifold. While LLE and Laplacian Eigenmaps are capable of learning the local structure of the manifold, ISOMAP can learn the global structure of data. Also, from the graph we can see advantages of new dimensionality reduction technique kDM that it can learn both local and global structure of Swiss roll dataset, i.e. "close" data points will be retained "close" and "far" data points will lie "far" in the embedding coordinates. Since kDM is a linear method, the running time is quite very low compared to other nonlinear techniques. We next vary the value of cluster number $k = 5, 10, 15, 25$ in kDM algorithm to see how kDM works (see Fig. 2, 3). It seems to us that , to some extent, kDM does not systematically depends on the number of clusters $k$, which actually is an issue under our investigation.

## 3.2 Experiment on Biometrics Databases

In this section, we do some experiment on real biometrics databases which are ORL face database and PolyU Palmprint database. We choose the value $k = 5$ in kDM algorithm to perform on both databases, while the parameters for the other methods are still same as in Table 1. Due to high dimensionality of biometrics data (3S problem), in this section the kDM algorithm is implemented based on the strategy of DLDA as discussed in previous section. In ORL face database, we randomly select 10 subjects, each of which contains 10 sample images. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement) and are manually cropped and resized to 50x50 pixel images. Two-dimensionality visualization of the ORL face database based on variety of techniques are presented in Fig. 4. It should be noted that the classification error rates are calculated and put on the title of each subplot in Fig. 4. We can see that kDM give the best accuracy rate (93%), LLE and Diffusion Maps are second best (91%), while Kernel PCA gives a very bad performance. The PolyU Palmprint Database[12] contains 7752 grayscale images corresponding to 386 different palms. We also select randomly 10 subjects and 10 palms for each subject to do the experiment. We use inscribed circle-based segmentation approach in [13] to extract palms and resize each palm to radius of 25 pixels. In the case of palmprint database, from Fig. 5, kDM still give good performance in terms of both data visualization and classification with 97% accuracy.

## 3.3 Discussion

The experiments on both artificial and real biometrics datasets have been systematically performed. These experiments reveal a number of interesting points as follow:

- kDM can be a good candidate for data visualization because it can learn the whole structure (both local and global structure) of data.
- Though kDM is a unsupervised technique, it still have the ability of finding discriminative features which is very helpful in classification taks.
- It is quite easy to implement and run fast compared to the other non-linear techniques.

# 4. CONCLUSIONS

In this paper, we propose a linear dimensionality reduction technique that can keep both local and global structure of data. The experiments on both artificial and real datasets show its potential in data visualization and classification tasks. The corner-stone of the idea is the usage of nice properties from k-means and Fisher criteria. In the first step of applying k-means, those "close" data samples will be tendentiously kept in the same cluster, those "distant" data samples will be grouped into different clusters. And this topology of data will be preserved by using Fisher criteria to embed data into low-dimensional representation. A future work is obviously the effect of k-means on the kDM algorithm.

# 5. ACKNOWLEDGMENTS

This research was supported by the MIC(Ministry of Information and Communication), Korea, Under the ITFSIP (IT Foreign Specialist Inviting Program) Supervised by the IITA (Institute of Information Technology Advancement).

# 6. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems 14*.

[2] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[3] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educational Psychology*, 27:417–441, 1933.

[4] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel pca and de-noising in feature spaces. pages 536–542, 1998.

[5] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21.

[6] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philoshophical Magazine*, 2:559–572, 1901.

[7] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[8] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125–140, 1962.

[9] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. 290(5500):2319–2323, December 2000.

[10] W. S. Torgerson. Multidimensional scaling. *Psychometrika*, 17:401–419, 1952.

[11] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data - with application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

[12] D. Zhang. Polyu palmprint palmprint database - http://www.comp.polyu.edu.hk/ biometrics/.

[13] D. Zhang. *Palmprint Authentication*. Kluwer Academic, 2004.