

---

# Manifold Learning and Applications in Recognition

Junping Zhang<sup>1</sup>, Stan Z. Li<sup>2</sup>, and Jue Wang<sup>3</sup>

<sup>1</sup> Intelligent Information Processing Laboratory, Fudan University, 200433, Shanghai, P.R.China

jpzhang@fudan.edu.cn

<sup>2</sup> Face Group, Microsoft Research Asia

100080, Beijing, P.R.China

szli@microsoft.com

<sup>3</sup> The Key Lab of Complex Systems and Intelligent Science, Institute of Automation, Chinese Academy of Sciences, 100080, Beijing, China.

jue.wang@mail.ia.ac.cn

## 1 Introduction

A large number of data such as images and characters under varying intrinsic principal features are thought of as constituting highly nonlinear manifolds in the high-dimensional observation space. Visualization and exploration of high-dimensional vector data are therefore the focus of much current machine learning research. However, most recognition systems using linear method are bound to ignore subtleties of manifolds such as concavities and protrusions, and this is a bottleneck for achieving highly accurate recognition. This problem has to be solved before we can make a high performance recognition system.

Recent years have seen progress in modeling nonlinear manifolds. Rich literature exists on manifold learning. On the basis of different representations of manifold learning, this can be roughly divided into four major classes: projection methods, generative methods, embedding methods, and mutual information methods.

1. The first is to find principal surfaces passing through the middle of data, such as the principal curves [1][2]. Though geometrically intuitive, the first one has difficulty on how to generalize the global variable-arc-length parameter- into higher-dimensional surface.
2. The second adopts generative topology models [3] [4] [5], and hypothesizes that observed data are generated from the evenly spaced low-dimensional latent nodes. And then the mapping relationship between the observation space and the latent space can be modeled. Resulting from the inherent insufficiency of the adopted EM (Expectation-Maximization) algorithms,

nevertheless, the generative models fall into local minimum easily and have slow convergence rates.

3. The third is generally divided into global and local embedding algorithms. ISOMAP [6], as a global algorithm, presumes that isometric properties should be preserved in both the observation space and the intrinsic embedding space in the affine sense. And extensions to conformal mappings is also discussed in [7]. On the other hand, Locally Linear Embedding (LLE) [8] and Laplacian Eigenmap [9] focus on the preservation of local neighbor structure.
4. In the fourth category, it is assumed that the mutual information is a measurement of the differences of probability distribution between the observed space and the embedded space, as in stochastic nearest neighborhood (henceforth SNE) [10] and manifold charting [11].

While there are many impressive results about how to discover the intrinsic features of the manifold, there have been fewer reports on the practical applications in manifold-learning, especially on object recognition. Some literature even makes negative conclusion that LLE is only useful for small numbers of dimensions, whereas the classifiers performs better for large numbers of dimensions on PCA-mapped data [12]. A possible explanation is that the practical data includes a large number of intrinsic features and have high curvature both in the observation space and in the embedded space, whereas present manifold learning methods strongly depends on the selection of parameters. Furthermore, we also found that if data of different classes belong to similar category, for example, face images, recognition can be implemented under the same subspace with manifold learning approaches. Otherwise, data (for example, character) should be mapped into the different subspace for further recognition.

Assuming that data are drawn independent and identically distributed from the underlying unknown distribution, we propose two recognition algorithms for processing the above-mentioned two cases in section 2. Experiments on image and character data show the advantages of our proposed recognition approaches. Finally, we discuss potential problems and further researches.

## 2 Manifold Learning Algorithm

### 2.1 dimensionality reduction

To establish the mapping relationship between the observed data and the corresponding low-dimensional data, the locally linear embedding (LLE) algorithm [8] is used to obtain the corresponding low-dimensional data  $Y$  ( $Y \subset \mathbb{R}^d$ ) of the training set  $X$  ( $X \subset \mathbb{R}^N, N \gg d$ ). And then the dataset  $(X, Y)$  is used for modeling the subsequently mapping relationship.

The main principle of LLE algorithm is to preserve local order relation of data in both the embedding space and the intrinsic space. Each sample in the

observation space is a linearly weighted average of its neighbors. The basic LLE algorithm based on local covering numbers can be described as follows:

Step 1: define

$$\psi(W) = \left\| x_i - \sum_{j=1}^K W_{ij} x_{ij} \right\|^2 \quad (1)$$

Consider the constraint  $\sum_{j=1}^K W_{ij} = 1$ , and if  $x_i$  and  $x_j$  are not in the same neighbor,  $W_{ij} = 0$ , compute the weighted matrix  $W$  according to the least square.

Step 2: define

$$\varphi(Y) = \left\| y_i - \sum_{j=1}^K W_{ij}^* y_{ij} \right\|^2 \quad (2)$$

where  $W^* = \arg \min_w \psi(W)$ . Consider the constraint  $\sum_i y_i = 0$  and  $\sum_i y_i y_i^T / n = I$ , where  $m$  is the number of local covering set. Calculate  $Y^* = \arg \min_Y \varphi(Y)$ .

Step 2 of the algorithm is equivalent to approximate the nonlinear manifold around point  $x_i$  by the linear hyperplane that passes through its neighbors  $\{x_{i1}, \dots, x_{ik}\}$ . Considering that the objective  $\varphi(Y)$  is invariant to translation in  $Y$ , constraint term  $\sum_i y_i = 0$  is added in the step 2. Moreover, the other term  $\sum_i y_i y_i^T / n = I$  is to avoid the degenerate solution of  $Y = 0$ . Hence, step 2 reduces to an eigenvector decomposition problem as follows:

$$\begin{aligned} Y^* &= \arg \min_Y \phi(Y) \\ &= \left\| y_i - \sum_{j=1}^K W_{ij}^* y_{ij} \right\|^2 \\ &= \arg \min_Y \left\| (I - W)Y \right\|^2 \\ &= \arg \min_Y Y^T (I - W)^T (I - W) Y \end{aligned} \quad (3)$$

The optimal solution of  $Y^*$  in Formula (3) is the smallest eigenvectors of matrix  $(I - W)^T (I - W)$ . Obviously, those eigenvalues which are zero need be removed. So we need to compute the bottom  $(d + 1)$  eigenvectors of the matrix and discard the smallest eigenvector considering constraint term.

Thus, we obtain the corresponding low-dimensional dataset  $Y$  in embedding space. And the completed set  $(X, Y)$  is used for the subsequent modeling of the mapping relationship.

A disadvantage of LLE algorithm is that it is difficult to compute the mapping of test samples due to the computational cost of eigenmatrix. With respect to weiestrass approximation theorem, we use the following gaussian RBF kernel to approximate the relationship:

$$y' = \sum_{i=1}^n \alpha_i K(x_i, x') \quad (4)$$

where  $K(x_i, x')$  is:

$$K(x_i, x') = \exp\left(-\frac{\|x_i - x'\|^2}{2\sigma^2}\right) \quad (5)$$

The parameter  $\sigma$  depends on the dimensionality and is usually predefined by user, and  $\alpha_i$  can be directly computed with the complete data  $(X, Y)$ . We name the procedure manifold learning algorithm (MLA) which means the most manifold learning approaches can be employed for reducing high-dimensional data into low-dimensional space.

### 3 Linear Discriminant Analysis

Assuming the data of different classes have the same or similarity categories, for instance, facial images sampled from different persons can be viewed as having the same cognitive concept. So data of different classes can be reduced into the same subspace with manifold learning approaches. While MLA is capable of recovering the intrinsic low-dimensional space, however, it may not be optimal for recognition. When the two highly nonlinear manifolds are mapped into the same low-dimensional subspace through MLA, for example, there is no reason to believe that the optimal classification hyperplane also exists between the two unravelled manifolds. If the principal axes of the two low-dimensional mapping classes of manifolds have an acute angle, the classification ability may be impaired [13]. Therefore, linear Discriminant analysis (LDA) is introduced to maximize the separability of data between different classes.

Suppose that each class is equal probability of event, Within-class scatter matrix is therefore defined as:  $S_w = \sum_{i=1}^L \sum_{j=1}^{n_i} (y_j - m_i)(y_j - m_i)^T$  for  $n_i$  samples from class  $i$  with class means  $m_i, i = 1, 2, \dots, L$ . For the overall mean  $m$  for all samples from all classes, meanwhile, the between-class scatter matrix is defined as  $S_b = \sum_{i=1}^L (m_i - m)(m_i - m)^T$  [13].

To maximize the between-class distances while minimizing the within-class distances of manifolds, the column vectors of discriminant matrix  $W$  are the eigenvectors of  $S_b^{-1}S_w$  associated with the largest eigenvalues. Projection matrix  $W$  play a role that which projects a vector in the low-dimensional face subspace into discriminatory feature space.

With the combination of LLE and LDA (MLA+LDA), we therefore avoid the problem of dimensionality curse and recognition task can be realized on the basis of reduced dimensions.

### 4 Nonlinear Auto-Associative Model

If data to be classified have remarkable different categories (for example, characters), MLA+LDA will be inefficient for recognition as these data can't be

commonly mapped into a single subspace with MLA. A corresponding strategy is to extract the intrinsic principal features of these manifolds with some dimensionality reduction methods separately, and then the unknown sample can be auto-associated in light of the intrinsic principal features.

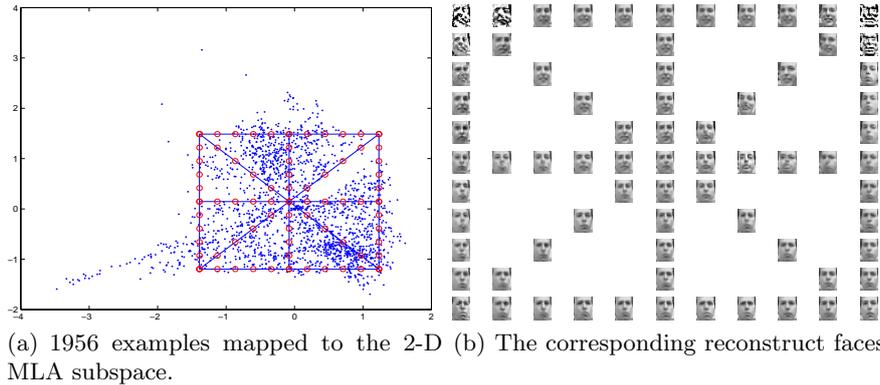
With light of Eq. (4), we can also construct the reconstructed formula of manifold learning as follows:

$$x' = \sum_{j=1}^{n_i} \beta_j k'(y_j, y') \quad y_j \in Z, y' \in R^d \quad (6)$$

$$k'(y_j, y') = \exp(-\|y_j - y'\|^2 / 2(\sigma')^2) \quad (7)$$

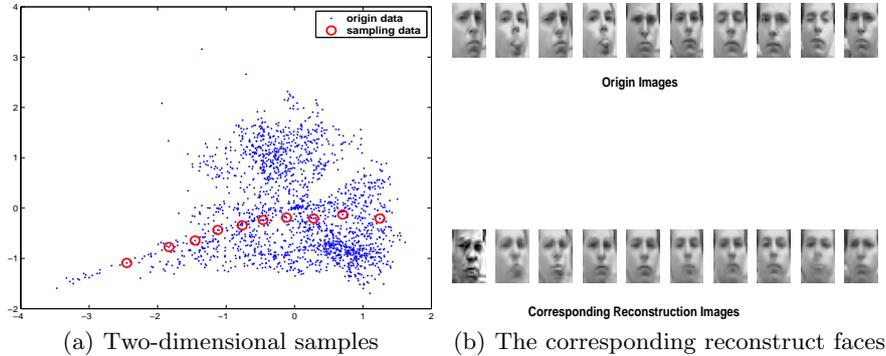
where  $\mathcal{B} = \{\beta_j\}$  is the  $N \times n$  weighted inverse mapping matrix or reconstruction matrix.

By choosing appropriate Eq. (5) and (7), the data can be reasonably reconstructed with the model representing the same category. For example, the Frey face database (20\*28 pixels, 1956 examples) [8] is also used to explain our proposed method. Firstly, in the MLA learning stage, the 491 cluster centers are extracted using vector quantization and mapped into 2-D space using LLE. Then all the 1956 face examples are mapped into the 2-D space based on the mapping learned in the first stage where  $\sigma^2 = 100$ , as shown in Figure 1(a). Thirdly, we randomly sample two points and use them as the upper-left and lower-right corner points for a rectangle, and then sample 11 evenly spaced points along each of the boundary and diagonal lines of the rectangle, and these points are reconstructed with Eq.(6) and (7), as displayed in Figure 1(b). We observe that a continuous expressional change in the vertical axes and pose change from the right side to the left side. Therefore, we have approximately recovered 2 intrinsic principal features, those of expression and pose, for the FREY database using our proposed method. To compare the dif-



**Fig. 1.** 2-dimensional mapping and reconstruction of Frey Face data

ference between the original images and the reconstructed images, 10 points from 2-dim reduced data as in Figure 2(a) are randomly sampled and then reconstructed via our proposed method where  $(\sigma')^2 = 1$ . The original facial images are shown on the top of Figure 2(b), while the corresponding reconstructed facial images at the bottom of Figure 2(b). It can be seen that our proposed method effectively reconstructs these images.



**Fig. 2.** 2-dimensional mapping and reconstruction of Frey Face data

This procedure has the foundation of cognitive sciences, namely, auto-association, which argues that recalling objects or concepts is achieved through preserving the underlying principal features of objects or concepts. Therefore, we call our proposed model "Nonlinear Auto-Associative modeling" (NAM).

It is obvious that the Frey data belong to one class. To implement recognition with NAM, we assume that each  $NAM_i$  represents the  $i$ th NAM of the  $i$ th class so that we can model a total of  $L$  different NAMs, for example, character 'a' corresponds to the 1-th NAM, and 'b' 2-th NAM, and so on. Under the assumption, the data of different classes can be represented as

$$(X(1), Y(1)), (X(2), Y(2)), \dots, (X(L), Y(L))$$

Different from the MLD+LDA, each completed data  $(X(i), Y(i))$ ,  $i = 1, \dots, L$  is realized with MLA separately. Consider auto-associative properties of NAMs, corresponding auto-associative sample through the NAM of the same class than through those of the different class would have higher similarity with the original sample. It is obvious that a variety of similarity measure techniques can be adopted. In this paper, recognition can be achieved by comparing the probability metric between each unknown sample and corresponding auto-associative sample with different NAMs. Without loss of generality, the probability metric (in this paper, we use Gaussian function) between each sample  $x'$  and auto-associative sample  $x'(i)$  of the  $i$ th NAM is given by:

$$P(x'(i)|x', NAM_i) = \exp(-\|x' - x'(i)\|^2), x', x'(i) \in R^N, i = 1, \dots, L \quad (8)$$

where  $x'_i$  means the auto-associative sample through the  $i$ th NAM given unknown sample  $x'$ . It is no difficult to see that when the reconstructed sample is the same as the original sample,  $P(x'(i)|x', NAM(i))$  is equal to 1, whereas if the reconstructed sample is far away from the original sample,  $P(x'(i)|x', NAM_i)$  will decreased to zero rapidly with respect to the properties of gaussian function.

To guarantee the consistency of probability metric, normalization is performed. The corresponding equation is given by:

$$\begin{aligned} P(x'(i)|x') &= \frac{P(x'(i)|x', NAM_i)}{\sum_{j=1}^L P(x'(j)|x', NAM_j)} \\ &= \frac{\exp(-\|x' - x'(i)\|^2)}{\sum_{j=1}^L \exp(-\|x' - x'(j)\|^2)} \end{aligned} \quad (9)$$

Consider Formula (9), the NAM where probability metric between auto-associative samples and original sample is highest can be viewed as a criterion of recognition. Hence, the equation for recognition is given by:

$$Cl(x') = \arg \max_i P(x'(i)|x'), i = 1, \dots, L \quad (10)$$

Our proposed NAM has several obvious merits: Firstly, our proposed NAM avoids the problems of local minimum and convergence rate. Secondly, the proposed NAM is constructive, and geometrically intuitive. Thirdly, it can find unlabeled sample through predefined threshold. This suggests "semi-supervised learning" characteristics where only partially labelled data are needed for the learning. Finally, it can add new NAMs without redesigning original NAMs.

## 5 Experiments

Experiments are performed using three object (face) databases (namely the Olivetti database [14], UMIST database [15] and JAFFE database [16]) and two character databases ((UCI character database [17] and OCR (optical character recognition) database [18])), to evaluate the feasibility of our proposed nonlinear dimension reduction (MLA+LDA) method and NAM method, respectively.

### 5.1 Image Recognition

The first object database provided by AT&T Cambridge Laboratories (formerly known as the Olivetti database) consists of 10 different images for 40 people each (four female and 36 male subjects). The images are taken at different time, varying the lighting, facial expressions (open/closed eyes,

smiling/non-smiling), and facial details(glasses/no-glasses). All the images are taken against a dark homogeneous background and the people are in up-right, frontal position. There are unstructured intermediate changes ( $\pm 20$  degrees) in head pose. Examples of ORL database are shown in Figure 3. We crop images into  $112 \times 96$  pixels (namely 10304 dimensions).



**Fig. 3.** Examples of ORL Face Database

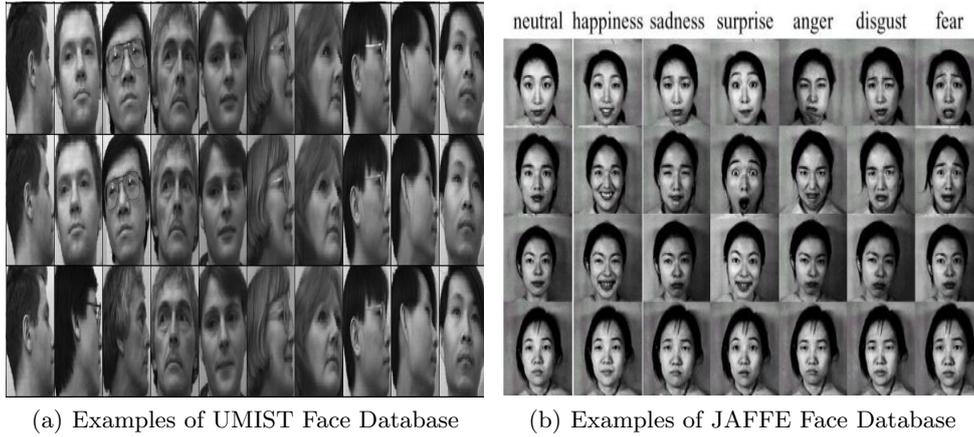
The training set and test set are divided in the same way as in [19]: The 10 images of each of the 40 persons are randomly partitioned into two sets, that is, 200 training images and 200 test images, without overlapping between the two sets.

The second one, UMIST database consists of 575 images of 20 people with varied poses. The images of each subject cover a range of poses from right profile (-90 degree) to frontal (0 degree)[15]. Examples of the UMIST database are shown in Figure 4(a). All these images of UMIST database are cropped to the size of  $112 \times 94$ , namely 10304 dimensions.

The main difficulty of UMIST database is that face data in the observation space may have higher curvature and stronger nonlinearity in multiple views than in frontal views. From the aspect of computer vision, meanwhile, "the variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to change in face identity"[20]. This makes multi-view face recognition a great challenge.

In this experiment, we randomly select 10 images of each person as the training set, and the remaining 375 images as the test set.

The JAFFE database, which has been used in facial expression recognition, consists of 213 images of 10 Japanese females. The head is almost in frontal pose. The number of each image represent one of the 7 categories of expressions (neutral, happiness, sadness, surprise, anger, disgust and fear). In our experiment, the database is used for both oriental face recognition and



(a) Examples of UMIST Face Database

(b) Examples of JAFFE Face Database

expression recognition. All these images of JAFFE database are cropped to the size of  $146 \times 111$  pixels.

When being used for face recognition, the JAFFE database is partitioned into two sets: 6 images of each of the 10 persons are randomly extracted to make 60 training set and the remaining 153 images are used as the test images. Meanwhile, in expression recognition, 24 images of each expressional categories are randomly extracted to make 168 training set and the remaining 45 images are used as the test set. Examples of JAFFE database are shown in Figure 4(b).

The dimension of LLE-reduced data is set to be 150 except for JAFFE face recognition (where the dimension is 50 as in Figure 4(e)). For the 2nd mapping, LDA based reduction, the reduced dimension cannot be more than  $L - 1$ . otherwise eigenvalues and eigenvectors will have complex values. Actually, we remain the real-value part of complex values when the 2th reduced dimensions are higher than  $L - 1$ .

In order to compare the performance of MLA+LDA in dimensionality reduction, we introduce a classical linear dimensionality reduction algorithm—PCA (principal component analysis) [21], and then design four combinational algorithms for face recognition: the combination of 1-nearest neighborhood classifiers with PCA+LDA (PCA+LDA+NN), the combination of means classifiers with PCA+LDA (PCA+LDA+M), the combination of 1-nearest neighborhood with MLA+LDA (MLA+LDA+NN) and the combination of means classifiers with MLA+LDA (MLA+LDA+M). All the experimental data have been normalized. The experimental results are the average of 100 runs.

In our experiments, two parameters (neighbor factor  $K'$  of LLE algorithm and  $\sigma^2$  of kernel function) need to be predefined. Without loss of generality, we set  $K'$  be 40 for ORL, UMIST and Jaffe expression database, 20 for JAFFE Face database, and set  $\sigma^2$  be 10000 for ORL and UMIST databases, 80000 for JAFFE expression and face database.

The experimental results are illustrated as in Figure 4(c), Figure 4(d), Figure 4(e), and Figure 4(f), respectively. The error rates (ER) of several face recognitions are also tabulated in Table 1. Where 1 means MLA+LDA+NN, 2 MLA+LDA+M, 3 PCA+LDA+NN, and 4 PCA+LDA+M in Table 1, respectively.

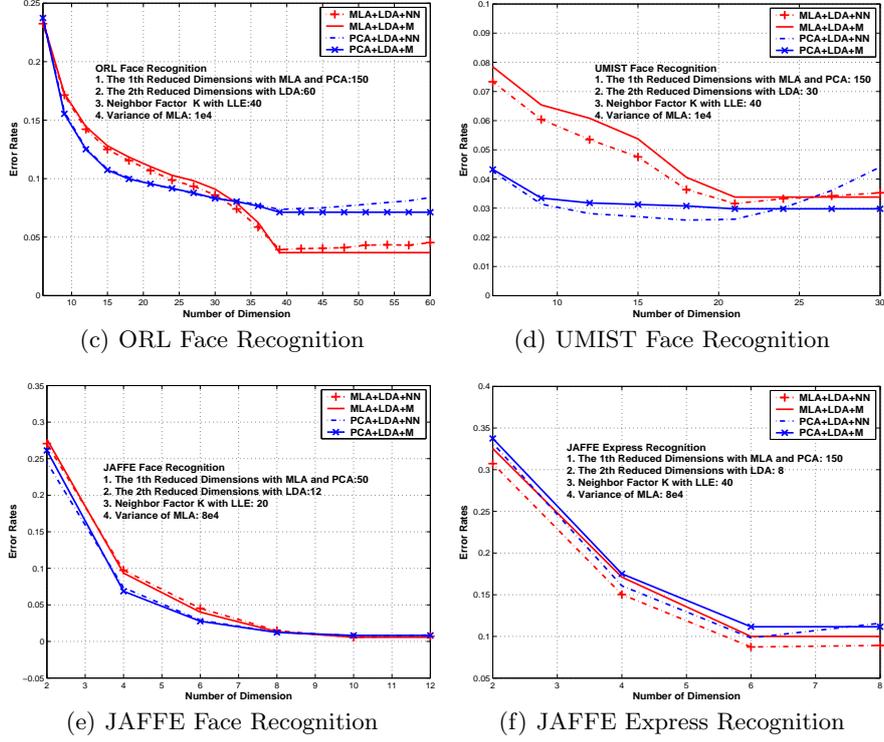


Fig. 4. Recognition Comparison

Table 1. Error rates with the MLA+LDA

Database(Dim)	MLA+LDA+NN	MLA+LDA+M	PCA+LDA+NN	PCA+LDA+M
ORL (39)	3.93%	3.68%	7.37%	7.13%
UMIST (21)	3.16%	3.38%	2.62%	2.98%
JAFFE (10)	0.57%	0.58%	0.78%	0.82%
EXPRESS(6)	8.73%	10%	9.82%	11.16%

From the Figures and Table 1, it can be seen that dimension of face manifolds has been remarkable reduced based on our proposed MLA+LDA methods. For example, the ratio between original dimensions and the 2th biggest reduced dimension of ORL database is about 264. Compared with PCA+LDA, recognition obtain better results in the reduced dimensions of three face databases except for UMIST database. For example, in 150 reduced dimensions, the error rates of MLA+LDA+M algorithm is about 93.6% of the MLA+LDA+NN, 49.9% of the PCA+LDA+NN, and 51.6% of the PCA+LDA+M on the ORL face database.

It is worthy noting that several parameters influence final experimental results. For example, the influences of variance  $\sigma^2$  of Gaussian RBF kernel, and training samples on ORL face recognition are illustrated as in Figure 5(a) and Figure 5(b), respectively. We observe that the curve of error rate about parameter  $\sigma^2$  will appear a 'valley' which is corresponding to the lowest error rates if adjusting parameter  $\sigma^2$ . Therefore, we assume that parameter  $\sigma^2$  may be selected automatically in further work. When training sample is 9 each person, meanwhile, our proposed MLA+LDA+M than other three methods has lowest error rates and is about 1.32%.

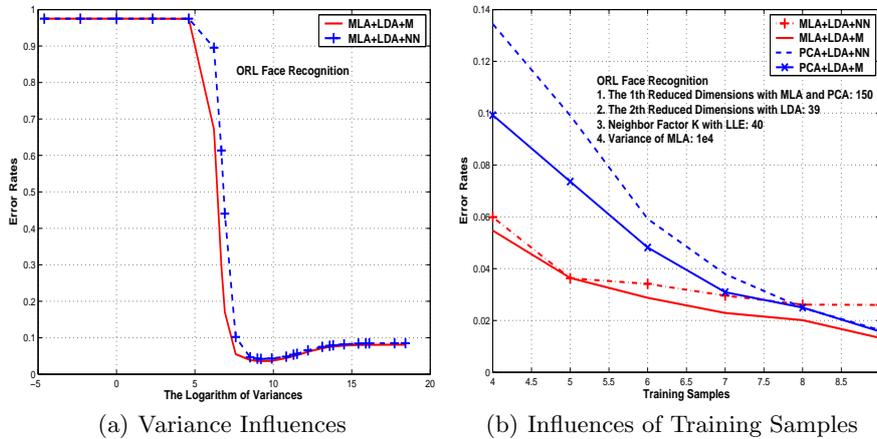


Fig. 5. Influences of Parameters

For comparing the recognition performance between the proposed MLA+LDA and MLA, we cite the recently experimental results from literature [22] as in Table 5.1:

Where NM means nearest manifold approach. The detail of NM approach can be seen in [22]. It is no difficult to see that our proposed MLA+LDA has greater reduced dimensions that MLA (Note: The reason is that the total number of classes of these mentioned data is less that the reduced dimensions of the data). And also with comparison of the recognition ability, the

**Table 2.** Error rates with MLA

Database(Dim)	MLA+NM	MLA+NN	PCA+NM	PCA+NN
UMIST(150)	3.73%	5.71%	4.83%	8.11%
ORL(120)	3.83%	7.75%	8.13%	9.86%
JAFFE (50)	2.99%	8.86%	8.76%	11.12%
EXPRESS(150)	12.39%	13.73%	16.63%	32.94%

recognition results of MLA+LDA is better than that of MLA in the average sense.

## 5.2 Character Recognition

The first character dataset from the UCI repository, comprises of a total of 20,000 labelled samples of, on an average, 770 examples per class. The total number of classes is 26. The character images were based on 20 different fonts and each character within these fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 numerical attributes (statistical moments and edge counts), which were then scaled to fit into a range of integer values from 0 through 15. Examples of the character images are illustrated in Figure 6. Because of the wide diversity among the different fonts and the primitive nature of the attributes, the recognition task was especially challenging. The database is randomly partitioned two disjoint-

**Fig. 6.** Examples of UCI character databases

ing sets, that is, 350 training samples of each of the 26 classes as the training set, and the remaining samples as the test set.

The second database was created by National Institute of Standards and Technology (NIST) and contains 16,280 handwritten characters. There are, on an average, 600 characters per class (26 classes). Each character is represented by a 30 dimensional feature vector of edge tangents. Each dimension in both datasets was linearly scaled to [0,1] interval. In this experiment, we randomly

select 300 samples of each character concept as the training set, and the remaining samples as the test set.

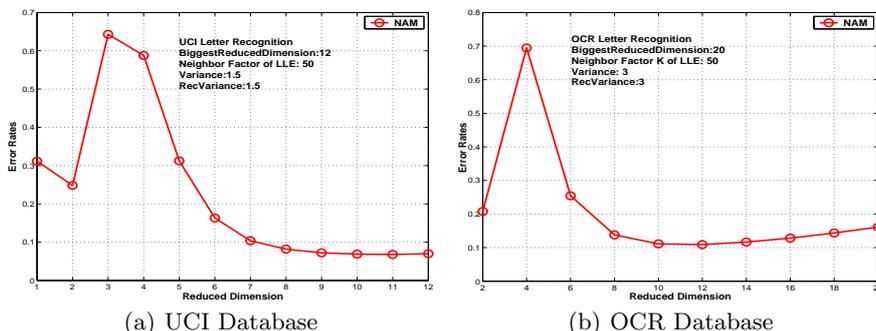


Fig. 7. Character Recognition

The biggest reduced dimensions are extracted based on the spectral properties of LLE algorithm [8](where the biggest reduced dimensions are 12 for UCI and 20 for OCR, respectively), and then the dimensions are gradually decreased according to the descending-order of eigenvalues used by LLE algorithm.

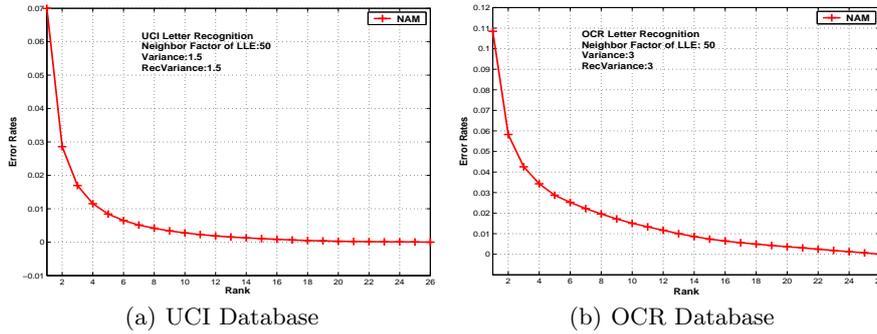
Furthermore, several parameters need to be predefined. Without loss of generality, neighbor factor  $K$  of LLE is set to 50. Consider the dimensionality difference of the two mentioned databases, parameters  $\sigma^2$  and  $(\sigma')^2$  are equal to 1.5 in UCI character database while being equal to 3 in OCR database. And then 26 independent NAMs are established for 26 different classes. The experimental results are the average of 100 runs.

The experimental results of two databases are illustrated as in Figure 7. From Figure 7 it can be seen that when intrinsic principal dimensions are equal to 12, the lowest error rates of two character databases are obtained. We therefore assume that the possible number of principal features of character manifolds should be extracted is 12 or so.

For comparing the recognition performance between the proposed NAMs and other known state-of-the-art algorithms, we cite the recently experimental results from literature [18] as in Table 3. As in Table 3, our proposed NAM than other classifiers has better recognition rates. For instances, in UCI character database, the error rates of NAMs is about 67.13% of the K-NN, 32.8% of the MLP; while in OCR character database, the error rates of NAMs is about 98.6% of the K-NN, 43.52% of the MLP. Furthermore, our proposed NAMs for two character databases using fewer features (12) to model intrinsic feature spaces. It is also noticeable that our experimental results are the average of 100 runs, whereas other results are the average of 10 runs. We also investigate the error rates of character recognition in the top  $n$  matches with

**Table 3.** The Average Error Rates of several algorithms on the two character databases for NAM, K-nearest Neighbor (K-NN, K=3), Maximum Likelihood classifier (MLC), Bayesian pairwise classifier with single Gaussian with voting combination method (BPC(1,V)), and MAP estimate combination (BPC(1,M)), and Bayesian pairwise classifier with mixture of Gaussian for voting (BPC(n,v)) and MAP estimate (BPC(n,M)) combination

CLASSIFIER	UCI %	OCR %
NAM	6.78 (12DIM)	10.358 (12DIM)
K-NN	10.10	10.5
MLP	20.7	23.8
MLC	17.3	20.5
BPC(1,V)	14.6	17.9
BPC(1,M)	14.7	16.7
BPC(N,V)	13.8	16.9
BPC (N,M)	12.4	13.7

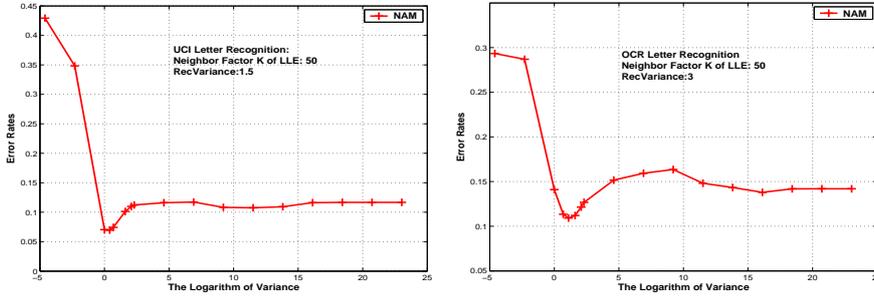


**Fig. 8.** Rank Performance

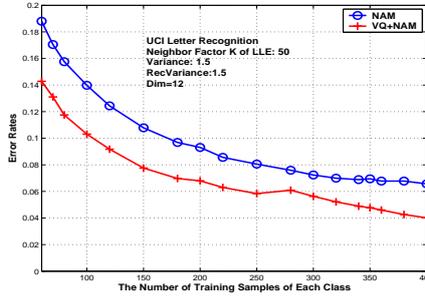
our proposed NAM. This lets one know how many characters have to be examined to get a desired level of performance. The performance statistics are reported as cumulative error rates, which are plotted on Figure 8. The horizontal axis of the graph is rank and the vertical axis of the percentage of error rates. For example, when in the top 3 matches, the error rate of NAM for UCI is 1.7%, and the error rate of NAM for OCR database is about 4.25%. Meanwhile, when in the top 6 matches, the error rate of NAM for UCI database is 0.65%, and the error rate of NAM for OCR database is about 2.5%.

It is worthy noting that several parameters, such as neighbor factor  $K$  of LLE algorithm, variance  $\sigma^2$ , and reconstructed variance  $(\sigma')^2$ , and the number of training sample, influence recognition results. We observe that the curve of error rates about parameter  $\sigma^2$  will appear a 'valley' which is corresponding to the lowest error rates when adjusting parameter  $\sigma^2$ . For example, the

influences of variance  $\sigma^2$  on two character recognition are illustrated as in Figure 9(a) and Figure 9(b). And the lowest error rates are obtained when  $\sigma^2$  is equal to 1.5 for UCI and 3 for OCR, respectively. Meanwhile, we observe that parameter  $(\sigma')^2$ , independent of the selection of parameter  $\sigma^2$ , has similar phenomenon on the curve of error rates. Therefore, the parameter  $\sigma^2$  and  $(\sigma')^2$  may be automatically selected in future work. And the influences of the



(a) Variance Influence of NAMs for UCI Database (b) Variance Influence of NAMs for OCR Database



(c) Influence of Training samples for UCI Character Recognition

**Fig. 9.** Parameter Influences

number of training samples about recognition rates are investigated and the results for the UCI database are illustrated in Figure 9(c). From the Figure, it can be noted that there are remarkable improvement in the error rates of the recognition tasks as the number of training samples increases. Experiments on the influence on selecting Gaussian RBF centers of our proposed NAM through vector quantization techniques (VQ) are also carried out on UCI character databases as in Figure 9(c), it is no difficult to see that with the combination of NAM and VQ the error rate has remarkable decreased, and therefore NAM+VQ can obtain the same error rates as NAMs with fewer RBF centers and training samples.

## 6 Conclusions

In this paper, we propose two recognition approaches (MLA+LDA and NAM) based on manifold learning. If data to be classified belongs to the same or similar cognitive category such face, MLA+LDA is employed. Otherwise, NAM approach is implemented.

Assuming that objects manifolds of different classes lie on the same feature subspace, object manifolds of different classes are first reduced into the intrinsic principal feature subspace with proposed MLA. And then the within-classes distances is further enlarged and the between-classes is decreased with LDA. The final classification task is completed based on the reduced dimensions of MLA+LDA.

If data to be classified are from remarkable different categories (for example, character 'a' and 'b'), recognition is achieved under the common feature subspace seems to be unreasonable. And the proposed nonlinear dimension reduction method MLA+LDA is not effective in this case. We therefore propose a new constructive nonlinear auto-associative modeling based on manifold learning. Based on our proposed NAM, the intrinsic principal features are extracted for preserving the principal structure of each manifold, and then reconstruction is achieved. With auto-associative mechanism, the reconstructed data through NAM having the same cognitive concept than through NAMS having the different cognitive concept will have less deviation. Therefore, the probability metric for recognition is naturally established. Our proposed NAM has several obvious merits: Firstly, it avoids problems with local minimum and convergence. Secondly, it is constructive, and geometrically intuitive. Meanwhile, our nonlinear auto-associative modeling can be used for the construction of both mapping and inverse mapping relationship between the observation space and corresponding feature spaces without dimensionality limitation.

Some potential problems remains. First, several parameters influence the experimental results. How to select parameter automatically is worthy to further research. Second, the number of intrinsic principal features of manifold is related to the error rates of recognition. In the future work, it is necessary to find a more effective approach which can estimate the number of principal features. The proposed NAM has semi-supervised learning characteristics which can find unlabelled sample through predefined threshold and can facilitates adding new NAMS without redesigning original NAMS, moreover, how to utilize these properties for modelling unknown concepts and designing new algorithms are also our future work.

## References

1. T. Hastie, and W. Stuetzle, (1988), "Principia Curves," *Journal of the American Statistical Association*, 84(406), pp. 502-516

2. B. Kégl, A. Krzyzak, T. Linder, and K. Zeger, (2000), "Learning and design of principal curves", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 281-297.
3. C.M. Bishop, M. Sevensen, and C.K.I. Williams, (1998), "GTM: The generative topographic mapping," *Neural Computation*, 10, pp. 215-234
4. K. Chang and J. Ghosh, (2001), "A unified model for probabilistic principal surfaces," *IEEE transactions on Pattern Analysis and Machine Intelligence*, 23(1), pp. 22-41
5. A.J. Smola, S.Mika, et al, (1999), "Regularized Principal Manifolds," *In Computational Learning Theory: 4th European Conference*, Vol 1572 of Lecture Notes in Artificial Intelligence, New York: Springer, pp. 251-256
6. J. B. Tenenbaum, de Silva, V. & Langford, J.C, (2000), "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290, pp. 2319-2323
7. V. S.Silva, J. B. Tenenbaum, (2002), "Unsupervised Learning of curved manifolds", *Nonlinear Estimation and Classification*, Springer-Verlag, New York.
8. S. T. Roweis, and K. S. Lawrence, (2000), "Nonlinear Dimensionality reduction by locally linear embedding," *Science*, 290, pp. 2323-2326
9. Mikhail Belkin, and Partha Niyogi. (2003) *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation Neural Computation*, 15(6), 1373 - 1396.
10. G. Hinton and S. Roweis, (2002), "Stochastic Neighbor Embedding," *Neural Information Proceeding Systems: Natural and Synthetic*, Vancouver, Canada, December 9-14.
11. M. Brand, MERL, (2002), "Charting a manifold," *Neural Information Proceeding Systems: Natural and Synthetic*, Vancouver, Canada, December 9-14.
12. Dick de Ridder, Robert P.W. Duin, "Locally Linear Embedding for Classification", technical reports, Delft University of Technology, The Netherlands
13. Daniel L. Swets and John (Juyang) Weng, (1996), "Using Discriminant Eigenfeatures for Image Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, pp.831-836.
14. F. S. Samaria, (1994), "Face Recognition Using Hidden Markov Models", PhD thesis, University of Cambridge.
15. H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie and T. S. Huang (eds), (1998), "em Characterizing virtual Eigensignatures for General Purpose Face Recognition", Daniel B Graham and Nigel M Allinson. In *Face Recognition: From Theory to Applications*; NATO ASI Series F, Computer and Systems Sciences, Vol. 163, pp. 446-456.
16. Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu, (1999), "Automatic classification of Single Facial Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, no. 12, pp. 1357-1362.
17. Frey, P. W., Slate, D. J. (1991). letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6, 161-182.
18. Kumar, S., Ghosh, J., Crawford, M. (2000) A Bayesian Pairwise Classifier for Character Recognition. *Cognitive and Neural Models for Word Recognition and Document Processing*. Nabeel Mursheed (Ed), World Scientific Press.
19. Steve Lawrence, C. Lee Giles, A. C. Tsoi, and A. D. Back, (1997) "Face recognition: A convolutional neural network approach", *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98-113.

20. Y. Moses, Y. Adini, and S. Ullman, (1994), "Face Recognition: The Problem of compensating for changes in illumination direction", in *Proceedings of the European Conference on Computer Vision*, vol. A, pp: 286-296.
21. M. Turk, and A. Pentland, (1991), "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol 3, no.1, pp.71-86
22. Junping Zhang, Stan Z. Li, Jue Wang, (2004), "Nearest Manifold Approach for Face Recognition", *The 6th IEEE International Conference on Automatic Face and Gesture Recognition*, May 17-19, Seoul, Korea, 2004.