Patch Alignment for Dimensionality Reduction

Tianhao Zhang, Dacheng Tao, *Member*, *IEEE*, Xuelong Li, *Senior Member*, *IEEE*, and Jie Yang

Abstract—Spectral analysis based dimensionality reduction algorithms are important and have been popularly applied in data mining and computer vision applications. To date many algorithms have been developed, e.g., principal component analysis, locally linear embedding, Laplacian eigenmaps, and local tangent space alignment. All of these algorithms have been designed intuitively and pragmatically, i.e., on the base of the experience and knowledge of experts for their own purposes. Therefore, it will be more informative to provide a systematic framework for understanding the common properties and intrinsic difference in different algorithms. In this paper, we propose such a framework, named "*patch alignment*", which consists of two stages: *part optimization* and *whole alignment*. The framework reveals that: i) algorithms are intrinsically different in the patch optimization stage; and ii) all algorithms share an almost-identical whole alignment stage. As an application of this framework, we develop a new dimensionality reduction algorithm, termed *Discriminative Locality Alignment* (DLA), by imposing discriminative information in the part optimization stage. DLA can: i) attack the distribution nonlinearity of measurements; ii) preserve the discriminative ability; and iii) avoid the small sample size problem. Thorough empirical studies demonstrate the effectiveness of DLA compared with representative dimensionality reduction algorithms.

Index Terms—Dimensionality reduction, spectral analysis, patch alignment, discriminative locality alignment.

1 Introduction

DIMENSIONALITY reduction based on spectral analysis is the process of transform measurements from a high-dimensional space to a low-dimensional subspace through the spectral analysis on specially constructed matrices [28]. It aims to reveal the intrinsic structure of the distribution of measurements in the original high-dimensional space and plays an important role in data mining, computer vision, and machine learning to deal with "curse of dimensionality" [4] for various applications, e.g., biometrics [29], [31], [35], [36], multimedia information retrieval [1], [16], [17], [30], document clustering [8], [14], and data visualization [23]. Representative spectral analysis based dimensionality reduction algorithms can be classified into two groups: i) conventional linear dimensionality reduction algorithms and ii) manifold learning based algorithms.

Representative conventional linear dimensionality reduction algorithms include principal component analysis (PCA) [20] and linear discriminant analysis (LDA) [13]. PCA maximizes the mutual information between original high-dimensional Gaussian distributed measurements and projected low-dimensional measurements. PCA, which is unsupervised, does not utilize the class label information. While, LDA finds a projection matrix that maximizes the trace of the between-class scatter matrix and minimizes the trace of the within-class scatter matrix in the projected subspace simultaneously. LDA is supervised since it utilizes class label information. The global linearity of PCA and LDA prohibit their effectiveness for non-linear distributed measurements.

Representative manifold learning based dimensionality reduction algorithms include locally linear embedding (LLE) [26], ISOMAP [32], Laplacian eigenmaps (LE) [3], Hessian eigenmaps (HLLE) [11], and local tangent space alignment (LTSA) [39]. LLE uses linear coefficients, which reconstruct a given measurement by its neighbours, to represent the local geometry, and then seeks a low-dimensional embedding, in which these coefficients are still suitable for reconstruction. ISOMAP, a variant of MDS [12], preserves global geodesic distances of all pairs of measurements. LE preserves proximity relationships by manipulations on an undirected weighted graph, which indicates neighbour relations of pairwise measurements. LTSA exploits the local tangent information as a representation of the local geometry and this local tangent information is then aligned to provide a global coordinate. HLLE obtains the final low-dimensional representations by applying eigen-analysis to a matrix which is built by estimating the Hessian over neighbourhood. All of these algorithms suffer from the out of sample problem [5]. One common response to this problem is to apply a linearization procedure to construct explicit maps over new measurements. Examples of this approach include locality preserving projections (LPP) [19], a linearization of LE; neighbourhood preserving embedding (NPE) [18], a linearization of LLE; orthogonal neighbourhood preserving projections (ONPP) [22], a linearization of LLE with the orthogonal constraint over the projection matrix; and linear local tangent space alignment (LLTSA) [38], a linearization of LTSA.

- *T. Zhang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, P. R. China. E-Mail: z.tianhao@gmail.com.*
- *D. Tao is with School of Computer Engineering, The Nanyang Technological University, 50 Nanyang Avenue, Blk N4, Singapore, 639798. E-Mail: dacheng.tao@gmail.com.*
- *X. Li is with the School of Computer Science and Information Systems, Birkbeck, University of London, U.K. E-Mail: xuelong@dcs.bbk.ac.uk.*
- *J. Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, P. R. China. E-Mail: jieyang@sjtu.edu.cn.*

Manuscript received in November 2007.

TABLE 1
IMPORTANT NOTATIONS USED IN THE PAPER

Notation	Description	Notation	Description
X	given dataset in a high-dimensional space	L_i	representation of part optimization
Y	dimension-reduced dataset	S_i	selection matrix
N	size of the dataset X	F_i	indices for the i^{th} patch
m	dimension of original measurements	I_d	$d \times d$ identity matrix
d	reduced dimension	$\vec{e}_l$	$[1,\cdots,1]^T \in \mathbb{R}^l$
$\mathbb{R}^m$	m-dimensional Euclidean space	$\vec{c}_i$	reconstruction coefficients in LLE
C	number of classes	$\vec{w}_i$	weighting vector in LE
X_i	i^{th} patch	R_N	centralization matrix
$\vec{x}_i$	given measurement in X	H_i	Hessian matrix
N_i	number of measurements in i^{th} class	k	number of neighbours
S_T	total scatter matrix	β	scaling factor
S_W	within-class scatter matrix	$\vec{\omega}_i$	coefficients vector in DLA
S_B	between-class scatter matrix	U	projection matrix

The above analysis shows that all the aforementioned algorithms are designed according to specific intuitions and solutions are given by optimizing intuitive and pragmatic objectives. That is, these algorithms have been developed based on the experience and knowledge of field experts for their own purposes. As a result, common properties and intrinsic differences of these algorithms are not completely clear. Therefore, it is essential and more informative to provide some a systematic framework for better understanding the common properties and intrinsic differences in algorithms.

In this paper, we propose such a framework termed "patch alignment" to unify spectral analysis based dimensionality reduction algorithms. This framework consists of two stages: part optimization and whole alignment. For part optimization, different algorithms have different optimization criteria over patches, each of which is built by one measurement associated with its related ones. For whole alignment, all part optimizations are integrated to form the final global coordinate for all independent patches based on the alignment trick, originally used by Zhang and Zha [39]. This framework discovers that: i) algorithms are intrinsically different in the patch optimization stage; and ii) all algorithms share an almost identical whole alignment stage. As an application of this framework, we also develop a new dimensionality reduction algorithm, termed Discriminative Locality Alignment (DLA), by imposing discriminative information in the part optimization stage. Benefits of DLA are threefold: i) because it takes into account the locality of measurements, it can deal with the nonlinearity of the measurement distribution; ii) because the neighbour measurements of different classes are considered, it well preserves discriminability of classes; and iii) because it obviates the need to compute the inverse of a matrix, it avoids the small sample size problem.

The rest of the paper is organized as follows: Section 2 introduces the proposed framework. In Section 3 we use this framework to explain existing spectral analysis based dimensionality reduction algorithms. In Section 4, DLA, the new algorithm, is described. In Section 5 we evaluate DLA in comparison with popular dimensionality reduction algorithms and Section 6 concludes.

For convenience, Table 1 lists important notations used in the rest of the paper.

2 PATCH ALIGNMENT FRAMEWORK

Consider a dataset X, which consists of N measurements $\vec{x}_i$ $(1 \le i \le N)$ in a high-dimensional space $\mathbb{R}^m$. That is $X = [\vec{x}_1,\cdots,\vec{x}_N] \in \mathbb{R}^{m\times N}$. The objective of a dimensionality reduction algorithm is to compute the corresponding low-dimensional representations $Y = [\vec{y}_1,\cdots,\vec{y}_N] \in \mathbb{R}^{d\times N}$, where $d < m$, of X. For the linear dimensionality reduction, it is necessary to find a projection matrix $U \in \mathbb{R}^{m\times d}$, such that $Y = U^T X$. For the non-linear dimensionality reduction, it is usually difficult to provide a explicit mapping to transform measurements from a high-dimensional space to a low-dimensional subspace.

In this framework, we first build N patches for each measurement in the dataset. Each patch consists of a measurement and its related ones, which depend on both characteristics of the dataset and the objective of an algorithm. Two cases are given in Fig. 1. As shown in Fig. 1a, global patches should be built based on each measurement and all the others, because measurements in this case are Gaussian distributed. In Fig. 1b, measurements are sampled at random from the S-curve manifold embedded in a 3-dimensional space. In this case, local patches should be built based on a given measurement and its nearest neighbours to capture the local geometry (locality). Global patches are usually built for conventional linear algorithms, e.g., PCA and LDA, while local patches are usually formed in manifold learning based

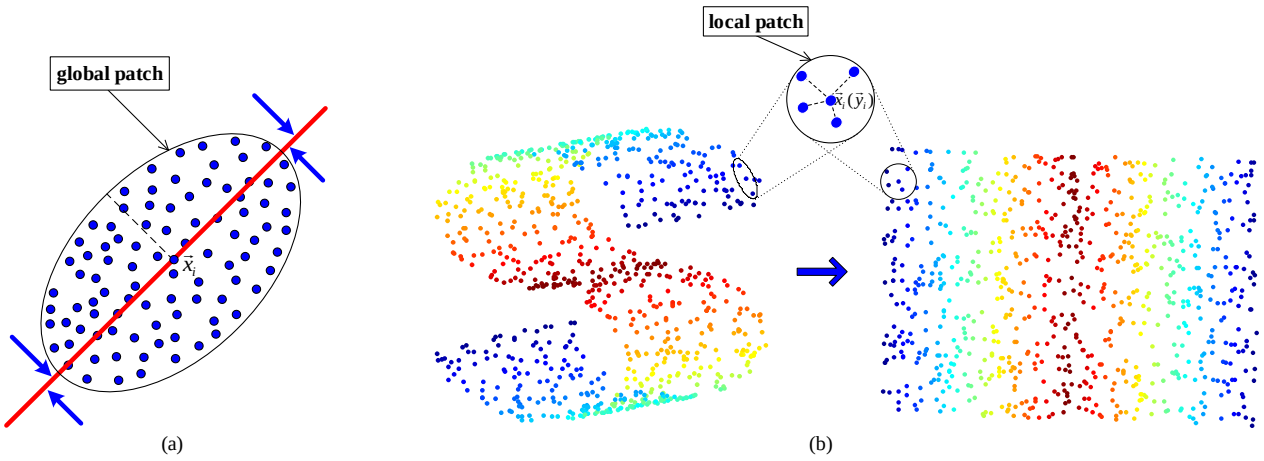


Fig. 1. Framework Illustration

ones, e.g., LLE and LE. Section 3 gives details of building patches for various algorithms. It is worth reminding that each measurement is associated with a patch. For better representation, we show only one patch for both Figs. 1a and 1b. With these built patches, optimization can be imposed on them based on an objective function, and then alignment trick [39] can be utilized to form a global coordinate.

2.1 Part optimization

Considering an arbitrary measurement $\vec{x}_i$ and its K related ones (e.g., nearest neighbours) $\vec{x}_{i_1},\cdots,\vec{x}_{i_K}$, the matrix $X_i=[\vec{x}_i,\vec{x}_{i_1},\cdots,\vec{x}_{i_K}]\in\mathbb{R}^{m\times(K+1)}$ is formed to denote the patch. For X_i, we have a part mapping $f_i:X_i\mapsto Y_i$ and $Y_i=[\vec{y}_i,\vec{y}_{i_1},\cdots,\vec{y}_{i_K}]\in\mathbb{R}^{d\times(K+1)}$. The part optimization is defined as

$$\arg\min_{Y_i}\operatorname{tr}\left(Y_iL_iY_i^T\right), \tag{1}$$

where $\operatorname{tr}(\cdot)$ is the trace operator; $L_i\in\mathbb{R}^{(K+1)\times(K+1)}$ encodes the objective function for the i^{th} patch; and L_i varies with the different algorithms.

2.2 Whole alignment

For each patch X_i, there is a low-dimensional representation Y_i. All Y_is can be unified together as a whole one by assuming that the coordinate for the i^{th} patch $Y_i=[\vec{y}_i,\vec{y}_{i_1},\cdots,\vec{y}_{i_K}]$ is selected from the global coordinate $Y=[\vec{y}_1,\cdots,\vec{y}_N]$, such that

$$Y_i=YS_i, \tag{2}$$

where $S_i\in\mathbb{R}^{N\times(K+1)}$ is the selection matrix and an entry is defined as:

$$(S_i)_{pq}=\begin{cases}1 & \text{if } p=F_i\{q\}\\ 0 & \text{else},\end{cases} \tag{3}$$

where $F_i=\{i,i_1,\cdots,i_K\}$ denotes the set of indices for i^{th} patch which is built by the measurement $\vec{x}_i$ (or $\vec{y}_i$) and its K related ones. Then, (1) can be rewritten as:

$$\arg\min_{Y}\operatorname{tr}\left(YS_iL_iS_i^TY^T\right). \tag{4}$$

By summing over all the part optimizations described as (4), we can obtain the whole alignment as:

$$\begin{aligned}&\arg\min_{Y}\sum_{i=1}^{N}\operatorname{tr}\left(YS_iL_iS_i^TY^T\right)\\ &=\arg\min_{Y}\operatorname{tr}\left(Y\left(\sum_{i=1}^{N}S_iL_iS_i^T\right)Y^T\right)\\ &=\arg\min_{Y}\operatorname{tr}\left(YLY^T\right),\end{aligned} \tag{5}$$

where $L=\sum_{i=1}^{N}S_iL_iS_i^T\in\mathbb{R}^{N\times N}$ is the alignment matrix [39]. It is obtained based on an iterative procedure:

$$L(F_i,F_i)\leftarrow L(F_i,F_i)+L_i, \tag{6}$$

for $i=1,\cdots,N$ with the initialization $L=0$. Note that $L(F_i,F_i)$ is a submatrix constructed by selecting certain rows and columns from L according to the index set F_i.

To uniquely determine Y, the constraint $YY^T=I_d$ is imposed on (5), where I_d is a $d\times d$ identity matrix. The objective function is then defined as:

$$\arg\min_{Y}\operatorname{tr}\left(YLY^T\right)\qquad \text{s.t. } YY^T=I_d. \tag{7}$$

For linearization [19], we can consider $Y=U^TX$ and (7) is deformed as:

$$\arg\min_{U}\operatorname{tr}\left(U^TXLX^TU\right)\qquad \text{s.t. } U^TXX^TU=I_d. \tag{8}$$

In addition, we can impose $U^TU=I_d$ as another way to uniquely determine the projection matrix U such that $Y=U^TX$. So, the objective function can be written as:

$$\arg\min_{U}\operatorname{tr}\left(U^TXLX^TU\right)\qquad \text{s.t. } U^TU=I_d. \tag{9}$$

Equations (7), (8), and (9) are basic optimization problems which can be solved by using Lagrangian multiplier method [21] and their solutions can be obtained by conducting the generalized or standard eigenvalue decomposition on XLX^T, i.e., $L\vec{\alpha}=\lambda\vec{\alpha}$, $XLX^T\vec{\alpha}=\lambda XX^T\vec{\alpha}$, and $XLX^T\vec{\alpha}=\lambda\vec{\alpha}$, respectively. The optimal solution for (7), (8), or (9) is the d eigenvectors associated with d smallest eigenvalues.

3 Unifying Various Algorithms of Dimensionality Reduction

Based on the proposed framework, in this section, we unify various spectral analysis based dimensionality re-

duction algorithms, e.g., LLE/NPE/ONPP, ISOMAP, LE/LPP, LTSA/LLTSA, HLLE, PCA and LDA.

The proposed framework identifies that these algorithms intrinsically differ in how to build patches and the corresponding optimizations on patches. All algorithms use an almost identical whole alignment procedure. Therefore, for each algorithm, we mainly provide how to build the patch X_i and the part optimization L_i.

3.1 LLE/NPE/ONPP

LLE represents the local geometry by using the linear coefficients which reconstruct a given measurement $\vec{x}_i$ by its k nearest neighbours $\vec{x}_{i_1},\cdots,\vec{x}_{i_k}$. Therefore, the patch is $X_i=\left[\vec{x}_i,\vec{x}_{i_1},\cdots,\vec{x}_{i_k}\right]$ and $\vec{x}_i$ can be linearly reconstructed from $\vec{x}_{i_1},\cdots,\vec{x}_{i_k}$ as:

$$\vec{x}_i=\left(\vec{c}_i\right)_1 x_{i_1}+\left(\vec{c}_i\right)_2 x_{i_2}+\cdots+\left(\vec{c}_i\right)_k x_{i_k}+\vec{\varepsilon}_i, \tag{10}$$

where $\vec{c}_i$ is a k-dimensional vector to encode reconstruction coefficients and $\vec{\varepsilon}_i$ is the reconstruction error. Minimizing the error yields:

$$\arg\min_{\vec{c}_i}\left\|\vec{\varepsilon}_i\right\|^2=\arg\min_{\vec{c}_i}\left\|\vec{x}_i-\sum_{j=1}^{k}\left(\vec{c}_i\right)_j\vec{x}_{i_j}\right\|^2. \tag{11}$$

With the sum-to-one constraint: $\sum_{j=1}^{k}\left(\vec{c}_i\right)_j=1$, $\vec{c}_i$ can be computed in a closed form as:

$$\left(\vec{c}_i\right)_j=\frac{\sum_{t=1}^{k}G_{jt}^{-1}}{\sum_{p=1}^{k}\sum_{q=1}^{k}G_{pq}^{-1}}, \tag{12}$$

where $G_{jt}=(\vec{x}_i-\vec{x}_{i_j})^T(\vec{x}_i-\vec{x}_{i_t})$ is called the local Gram matrix [27].

LLE assumes that $\vec{c}_i$ reconstructs both $\vec{x}_i$ from $\vec{x}_{i_1},\cdots,\vec{x}_{i_k}$ in the high-dimensional space and $\vec{y}_i$ from $\vec{y}_{i_1},\cdots,\vec{y}_{i_k}$ in the low-dimensional subspace. Based on this point, the cost function can be reformulated as:

$$\begin{aligned}\arg\min_{Y_i}\left\|\vec{\sigma}_i\right\|^2&=\arg\min_{Y_i}\left\|\vec{y}_i-\sum_{j=1}^{k}\left(\vec{c}_i\right)_j\vec{y}_{i_j}\right\|^2\\&=\arg\min_{Y_i}\operatorname{tr}\left(Y_i\begin{bmatrix}-1\\ \vec{c}_i\end{bmatrix}\begin{bmatrix}-1 & \vec{c}_i^{\,T}\end{bmatrix}Y_i^T\right)\\&=\arg\min_{Y_i}\operatorname{tr}\left(Y_iL_iY_i^T\right),\end{aligned} \tag{13}$$

where, $L_i=\begin{bmatrix}-1\\ \vec{c}_i\end{bmatrix}\begin{bmatrix}-1 & \vec{c}_i^{\,T}\end{bmatrix}=\begin{bmatrix}1 & -\vec{c}_i^{\,T}\\ -\vec{c}_i & \vec{c}_i\vec{c}_i^{\,T}\end{bmatrix}$. With L_i, (6), and (7), we can obtain the low-dimensional representations under the proposed framework.

In our framework, NPE [18] is the case that LLE changes its objective function to (8) and it is the linearization of LLE. ONPP [22] can be seen as the orthogonal linearization of LLE and its objective function is given by (9).

3.2 ISOMAP

ISOMAP preserves the pairwise geodesic distances [32] and its objective function is defined as

$$\arg\min\sum_{m,n}\left(d_G(m,n)-d'(m,n)\right)^2, \tag{14}$$

where $d_G(m,n)$ is an approximated geodesic distance between the m^{th} measurement and the n^{th} measurement in the high-dimensional space and $d'(m,n)$ is the corresponding Euclidean distance in the low-dimensional subspace. According to [32], these distances can be converted to inner products [12]. Denoting $D_G=\left[d_G(m,n)\right]$ as the matrix whose entries are approximated geodesic distances between the m^{th} measurement and the n^{th} measurement in the high-dimensional space, the inner product matrix $\tau(D_G)$ is obtained by $\tau(D_G)=-R_NS_GR_N/2$, where $(S_G)_{ij}=(D_G)_{ij}^2$; $R_N=I_N-\vec{e}_N\vec{e}_N^{\,T}/N$ is the centralization matrix; $\vec{e}_N=[1,\cdots,1]^T\in\mathbb{R}^N$; and I_N is an $N\times N$ identity matrix. Therefore, the objective function of ISOMAP described in (14) can be converted to:

$$\arg\min_{Y}\left\|\tau(D_G)-\tau(D_Y)\right\|^2. \tag{15}$$

Equation (15) can be further transformed to

$$\begin{aligned}&\arg\min_{Y}\left\|\tau(D_G)-Y^TY\right\|^2\\&=\arg\min_{Y}\operatorname{tr}\left(\tau(D_G)\tau(D_G)^T-2Y\tau(D_G)Y^T+Y^TYY^TY\right).\end{aligned} \tag{16}$$

Assuming that Y^TY is a constant matrix, (16) can be reformulated to

$$\begin{aligned}&\arg\max_{Y}\operatorname{tr}\left(Y\tau(D_G)Y^T\right)\\&=\arg\max_{Y}N\operatorname{tr}\left(Y\frac{1}{N}\tau(D_G)Y^T\right)\\&=\arg\max_{Y_i}\sum_{i=1}^{N}\operatorname{tr}\left(Y_i\frac{1}{N}\tau\left(D_G{}^i\right)Y_i^T\right),\end{aligned} \tag{17}$$

where $D_G{}^i=\left[d_G\left(F_i\{m\},F_i\{n\}\right)\right]$, $d_G\left(F_i\{m\},F_i\{n\}\right)$ is the approximated geodesic distance between the $F_i\{m\}^{\text{th}}$ measurement and the $F_i\{n\}^{\text{th}}$ measurement both of which are on the i^{th} patch, and $Y_i=[\vec{y}_i,\vec{y}_{i_1},\cdots,\vec{y}_{i_{N-1}}]$ denotes the i^{th} patch which is built by the given measurement $\vec{y}_i$ and all the rest ones $\vec{y}_{i_1},\cdots,\vec{y}_{i_{N-1}}$. The patch is global because it contains all measurements.

Equation (17) stands for the whole alignment for all measurements. For each measurement, we have the part optimization: $\arg\max_{Y_i}\operatorname{tr}\left(Y_iL_iY_i^T\right)$, where $L_i=(1/N)\cdot\tau\left(D_G{}^i\right)$. With L_i and (6), we can form the alignment matrix L which is equivalent to $\tau(D_G)$.

Note that ISOMAP is different from other non-linear algorithms which have the constraint $YY^T=I_d$ in (7). In ISOMAP [32], the imposed constraint is $YY^T=diag\left(\left[\lambda_1,\cdots,\lambda_d\right]\right)$, where λ_i is the corresponding eigenvalue of L and $\operatorname{diag}(\cdot)$ is the diagonalisation operation.

3.3 LE/LPP

LE preserves the local geometry based on manipulations on an undirected weighted graph, which indicates neighbour relations of pairwise measurements. The objective function of LE is:

$$\arg\min_{Y}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\vec{y}_i-\vec{y}_j\right\|^2W(i,j), \tag{18}$$

where $W\in\mathbb{R}^{N\times N}$ is the relation matrix weighted by the heat kernels [25]: $W(i,j)=\exp\left(-\left\|\vec{x}_i-\vec{x}_j\right\|^2/t\right)$ if $\vec{x}_i$ is one of the k nearest neighbours of $\vec{x}_j$ or $\vec{x}_j$ is one of the k nearest neighbours of $\vec{x}_i$, otherwise 0, and t is a tuning parameter.

To unify LE into the proposed framework, we rewrite

(18) as:

$$\arg\min_{\vec{y}_i}\sum_{i=1}^{N}\sum_{j=1}^{l}\left\|\vec{y}_i-\vec{y}_{i_j}\right\|^2\left(\vec{w}_i\right)_j, \tag{19}$$

where, $\vec{y}_{i_j}$, $j=1,\cdots,l$, are l connected measurements of the given measurement $\vec{y}_i$ in the graph and $\vec{w}_i$ is the l-dimensional column vector weighted by $\left(\vec{w}_i\right)_j=\exp\left(-\left\|\vec{x}_i-\vec{x}_{i_j}\right\|^2/t\right)$. Therefore, (19) can be reformulated to

$$\begin{aligned}&\arg\min_{\vec{y}_i}\sum_{i=1}^{N}\mathrm{tr}\left(\begin{bmatrix}\left(\vec{y}_i-\vec{y}_{i_1}\right)^T\\ \vdots\\ \left(\vec{y}_i-\vec{y}_{i_l}\right)^T\end{bmatrix}\left[\vec{y}_i-\vec{y}_{i_1},\ \cdots,\ \vec{y}_i-\vec{y}_{i_l}\right]\mathrm{diag}\left(\vec{w}_i\right)\right)\\ &=\arg\min_{Y_i}\sum_{i=1}^{N}\mathrm{tr}\left(Y_i\begin{bmatrix}-\vec{e}_l^{\,T}\\ I_l\end{bmatrix}\mathrm{diag}\left(\vec{w}_i\right)\left[-\vec{e}_l\quad I_l\right]Y_i^T\right)\\ &=\arg\min_{Y_i}\sum_{i=1}^{N}\mathrm{tr}\left(Y_iL_iY_i^T\right),\end{aligned} \tag{20}$$

where

$$L_i=\begin{bmatrix}-\vec{e}_l^{\,T}\\ I_l\end{bmatrix}\mathrm{diag}\left(\vec{w}_i\right)\left[-\vec{e}_l\quad I_l\right]=\begin{bmatrix}\sum_{j=1}^{l}\left(\vec{w}_i\right)_j & -\vec{w}_i^{\,T}\\ -\vec{w}_i & \mathrm{diag}\left(\vec{w}_i\right)\end{bmatrix};$$

$Y_i=\left[\vec{y}_i,\vec{y}_{i_1},\cdots,\vec{y}_{i_l}\right]$; $\vec{e}_l=\left[1,\cdots,1\right]^T\in\mathbb{R}^l$; and I_l is an $l\times l$ identity matrix.

Equation (20) serves as the whole alignment for all measurements. For each measurement, we have the part optimization: $\arg\min_{Y_i}\mathrm{tr}\left(Y_iL_iY_i^T\right)$. Therefore, the patch X_i is built by the given measurement $\vec{x}_i$ and its connected measurements $\vec{x}_{i_1},\cdots,\vec{x}_{i_l}$, which consist i) the k nearest neighbours of the given measurement and ii) measurements which see the given measurement as one of the k nearest neighbours. With L_i and (6), we can construct the alignment matrix L which is equal to the Laplacian matrix [3] used in LE. Finally, we can obtain the low-dimensional representations by using (7). Note that in LE [3], the imposed constraint is $YDY^T=I_d$, where D is the diagonal weighting matrix and $D_{ii}=\sum_{j=1}^{N}W(i,j)$. In our framework, $D_{ii}=\sum_{j=1}^{l}\left(\vec{w}_i\right)_j+1$.

In the proposed framework, LPP [19] is the linearization of LE by using (8) as the objective function.

3.4 LTSA/LLTSA

LTSA uses tangent coordinates to represent the local geometry. The patch X_i is defined the same as that in LLE. To obtain the optimal tangent coordinates, LTSA has the objective function on each patch:

$$\arg\min_{\Theta_i,Q_i}\left\|X_iR_{k+1}-Q_i\Theta_i\right\|^2, \tag{21}$$

where, $R_{k+1}=I_{k+1}-\vec{e}_{k+1}\vec{e}_{k+1}^{\,T}/\left(k+1\right)$ denotes the centralization matrix; $\vec{e}_{k+1}=\left[1,\cdots,1\right]^T\in\mathbb{R}^{k+1}$; I_{k+1} is a $\left(k+1\right)\times\left(k+1\right)$ identity matrix; $Q_i\in\mathbb{R}^{m\times d}$ is an orthonormal basis matrix of the tangent space; and $\Theta_i\in\mathbb{R}^{d\times(k+1)}$ is the tangent coordinates corresponding to Q_i. The optimal Q_i is the matrix of d left singular vectors of X_iR_{k+1} corresponding to its d largest singular values and the optimal tangent coordinates Θ_i are defined as

$$\Theta_i=Q_i^TX_iR_{k+1}. \tag{22}$$

Assume that there is an affine projection matrix, which projects tangent coordinates Θ_i to the low-dimensional coordinates Y_i. Then, we have:

$$Y_iR_{k+1}=T_i\Theta_i+E_i, \tag{23}$$

where T_i is the projection matrix and E_i is the error term. To preserve the local geometry in the low-dimensional space, LTSA finds Y_i and T_i by minimizing the error E_i:

$$\arg\min_{Y_i,T_i}\left\|E_i\right\|^2=\arg\min_{Y_i,T_i}\left\|Y_iR_{k+1}-T_i\Theta_i\right\|^2. \tag{24}$$

Therefore, the optimal affine projection matrix $T_i=Y_iR_{k+1}\Theta_i^+$, where Θ_i^+ is the Moore-Penrose generalized inverse of Θ_i. Equation (24) can be written as:

$$\arg\min_{Y_i}\left\|Y_iR_{k+1}\left(I_{k+1}-\Theta_i^+\Theta_i\right)\right\|^2. \tag{25}$$

Let V_i denote the matrix of d right singular vectors of X_iR_{k+1} corresponding to its d largest singular values, (25) can be converted to:

$$\begin{aligned}&\arg\min_{Y_i}\left\|Y_iR_{k+1}\left(I_{k+1}-V_iV_i^T\right)\right\|^2\\ &=\arg\min_{Y_i}\mathrm{tr}\left(Y_iR_{k+1}\left(I_{k+1}-V_iV_i^T\right)\left(I_{k+1}-V_iV_i^T\right)^TR_{k+1}Y_i^T\right)\\ &=\arg\min_{Y_i}\mathrm{tr}\left(Y_iL_iY_i^T\right),\end{aligned} \tag{26}$$

where, $L_i=R_{k+1}\left(I_{k+1}-V_iV_i^T\right)\left(I_{k+1}-V_iV_i^T\right)^TR_{k+1}=R_{k+1}-V_iV_i^T$. With L_i, (6), and (7), we can obtain the low-dimensional representations under the proposed framework.

LLTSA [38] is the linearization of LTSA and its objective function is defined by (8).

3.5 HLLE

HLLE assumes the low-dimensional representations can be obtained from a $\left(d+1\right)$-dimensional null-space of $\tilde{H}$ which indicates the curviness of the manifold M^d, if the manifold is locally isometric to an open connected subset of $\mathbb{R}^d$.

$\tilde{H}$ can be measured by averaging the Frobenius Norm of the Hessians on the manifold as

$$\tilde{H}\left(f\right)=\int_M\left\|H_f\left(m\right)\right\|_F^2dm, \tag{27}$$

where $f:X\mapsto Y$ describes the smooth functions and H_f denotes the Hessian of f.

To define the Hessian, HLLE uses orthogonal coordinates on the tangent planes of M^d. Suppose that tangent coordinates of X_i are given by Θ_i, where X_i is the patch built by the given measurement and its nearest neighbours. Let $g\left(\left(\Theta_i\right)_j\right)=f\left(\vec{x}_{i_j}\right)$ define a function on $g:U_i\mapsto\mathbb{R}$, where U_i is the neighbourhood formed by the components of tangent coordinates of X_i. The Hessian of f at $\vec{x}_i$ in tangent coordinates can be defined as a matrix, whose each entry is defined by:

$$\left(H_f^{\tan}\left(\vec{x}_i\right)\right)_{p,q}=\frac{\partial}{\partial\left(\Theta_i\right)_{j_p}}\frac{\partial}{\partial\left(\Theta_i\right)_{j_q}}g\left(\left(\Theta_i\right)_j\right)\Big|_{\left(\Theta_i\right)_j=0}, \tag{28}$$

$p,q=1,\cdots,d$.

HLLE uses Gram-Schmidt orthonormalization to estimate each tangent Hessian, and one can refer to [11] for the details. Using these tangent Hessians, $\tilde{H}$ can be con-

structed based under our framework as follows. For each patch X_i, we have $L_i = H_i H_i^T$, where H_i is the Hessian matrix. With L_i and (6), we can obtain the alignment matrix L which is equivalent to $\tilde{H}$ and on which the final solutions can be found through the eigenvalue decomposition as described in (7).

3.6 PCA

PCA maximizes the trace of total scatter matrix in the projected subspace, that is:

$$\arg\max_{\vec{y}_i} \mathrm{tr}(S_T) = \arg\max_{\vec{y}_i} \mathrm{tr}\left(\sum_{i=1}^{N} (\vec{y}_i - \vec{y}^m)(\vec{y}_i - \vec{y}^m)^T \right), \quad (29)$$

where $\vec{y}^m$ is the centroid of all measurements. To unify PCA into the proposed framework, we rewrite (29) as:

$$\arg\max_{\vec{y}_i} \sum_{i=1}^{N} \mathrm{tr}\left(\frac{1}{N^2} \left(\sum_{j=1}^{N-1} (\vec{y}_i - \vec{y}_{i_j}) \right) \left(\sum_{j=1}^{N-1} (\vec{y}_i - \vec{y}_{i_j}) \right)^T \right), \quad (30)$$

where, $\vec{y}_{i_j}$, $j = 1, \cdots, N-1$, are the rest measurements for $\vec{y}_i$. Equation (30) reduces to

$$\begin{aligned} &\arg\max_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(\frac{1}{N^2} \left(Y_i \begin{bmatrix} N-1 \\ -\vec{e}_{N-1} \end{bmatrix} \right) \left(Y_i \begin{bmatrix} N-1 \\ -\vec{e}_{N-1} \end{bmatrix} \right)^T \right) \\ &= \arg\max_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(Y_i L_i Y_i^T \right), \end{aligned} \quad (31)$$

where,

$$L_i = \frac{1}{N^2} \begin{bmatrix} N-1 \\ -\vec{e}_{N-1} \end{bmatrix} \begin{bmatrix} N-1 & -\vec{e}_{N-1}^T \end{bmatrix} = \frac{1}{N^2} \begin{bmatrix} (N-1)^2 & -(N-1)\vec{e}_{N-1}^T \\ -(N-1)\vec{e}_{N-1} & \vec{e}_{N-1}\vec{e}_{N-1}^T \end{bmatrix};$$

$Y_i = \left[\vec{y}_i, \vec{y}_{i_1}, \cdots, \vec{y}_{i_{N-1}} \right]$; and $\vec{e}_{N-1} = [1, \cdots, 1]^T \in \mathbb{R}^{N-1}$.

Equation (31) can be seen as the whole alignment for all measurements, each of which has the part optimization: $\arg\max_{Y_i} \mathrm{tr}(Y_i L_i Y_i^T)$. It is clear that the patch X_i is built by the given measurement $\vec{x}_i$ and all the rest ones $\vec{x}_{i_1}, \cdots, \vec{x}_{i_{N-1}}$. It is global since it contains all measurements.

Under our framework, we can build the alignment matrix L, with L_i and (6). Because PCA is an orthogonal linear algorithm, we can form its objective function based on (9).

3.7 LDA

LDA tries to find the subspace that discriminates different classes by minimizing the trace of the within-class scatter matrix S_W, while maximizing the trace of the between-class scatter matrix S_B.

For S_W, it has:

$$\begin{aligned} &\arg\min \mathrm{tr}(S_W) \\ &= \arg\min_{\vec{y}_j^{(i)}} \mathrm{tr}\left(\sum_{i=1}^{C} \sum_{j=1}^{N_i} (\vec{y}_i^{(j)} - \vec{y}_i^m)(\vec{y}_i^{(j)} - \vec{y}_i^m)^T \right), \end{aligned} \quad (32)$$

where, C is the number of classes; N_i is the number of measurements in the i^{th} class; $\vec{y}_i^{(j)}$ is the j^{th} measurement in the i^{th} class; and $\vec{y}_i^m$ is the centroid of the i^{th} class. To unify LDA into the proposed framework, we rewrite (32) as:

$$\arg\min_{\vec{y}_i} \mathrm{tr}\left(\sum_{i=1}^{N} \frac{1}{N_i^2} \left(\sum_{j=1}^{N_i-1} (\vec{y}_i - \vec{y}_{i_j}) \right) \left(\sum_{j=1}^{N_i-1} (\vec{y}_i - \vec{y}_{i_j}) \right)^T \right), \quad (33)$$

where, $\vec{y}_{i_j}$, $j = 1, \cdots, N_i - 1$, are $N_i - 1$ rest measurements in the same class of $\vec{y}_i$. Equation (33) reduces to

$$\begin{aligned} &\arg\min_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(\frac{1}{N_i^2} \left(Y_i \begin{bmatrix} N_i - 1 \\ -\vec{e}_{N_i-1} \end{bmatrix} \right) \left(Y_i \begin{bmatrix} N_i - 1 \\ -\vec{e}_{N_i-1} \end{bmatrix} \right)^T \right) \\ &= \arg\min_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(Y_i L_i^W Y_i^T \right), \end{aligned} \quad (34)$$

where,

$$L_i^W = \frac{1}{N_i^2} \begin{bmatrix} N_i - 1 \\ -\vec{e}_{N_i-1} \end{bmatrix} \begin{bmatrix} N_i - 1 & -\vec{e}_{N_i-1}^T \end{bmatrix} = \frac{1}{N_i^2} \begin{bmatrix} (N_i-1)^2 & -(N_i-1)\vec{e}_{N_i-1}^T \\ -(N_i-1)\vec{e}_{N_i-1} & \vec{e}_{N_i-1}\vec{e}_{N_i-1}^T \end{bmatrix};$$

$Y_i = \left[\vec{y}_i, \vec{y}_{i_1}, \cdots, \vec{y}_{i_{N_i-1}} \right]$; and $\vec{e}_{N_i-1} = [1, \cdots, 1]^T \in \mathbb{R}^{N_i-1}$.

Equation (34) can be seen as the whole alignment for all measurements, each of which has the part optimization: $\arg\min_{Y_i} \mathrm{tr}(Y_i L_i^W Y_i^T)$. Clearly, the patch X_i is built by a given measurement $\vec{x}_i$ and the rest ones in the same class $\vec{x}_{i_1}, \cdots, \vec{x}_{i_{N_i-1}}$. It is global since it contains all measurements in one class.

With L_i^W and (6), we obtain the whole alignment under the proposed framework:

$$\arg\min_{Y} \mathrm{tr}(Y L^W Y^T), \quad (35)$$

which is equivalent to (32).

For S_B, it has:

$$\begin{aligned} &\arg\max \mathrm{tr}(S_B) \\ &= \arg\max_{\vec{y}_i^m} \mathrm{tr}\left(\sum_{i=1}^{C} N_i (\vec{y}_i^m - \vec{y}^m)(\vec{y}_i^m - \vec{y}^m)^T \right). \end{aligned} \quad (36)$$

It is shown that we can only exploit the centroids of different classes to represent the between-class scatter. Equation (36) can be rewritten as:

$$\arg\max_{\vec{y}_i^m} \mathrm{tr}\left(\sum_{i=1}^{C} N_i \frac{1}{C^2} \left(\sum_{j=1}^{C-1} (\vec{y}_i^m - \vec{y}_{i_j}^m) \right) \left(\sum_{j=1}^{C-1} (\vec{y}_i^m - \vec{y}_{i_j}^m) \right)^T \right), \quad (37)$$

where, $\vec{y}_{i_j}^m$, $j = 1, \cdots, C-1$, are centroids of the different classes from $\vec{y}_i^m$. Equation (37) reduces to

$$\begin{aligned} &\arg\max_{Y_i^m} \sum_{i=1}^{C} \mathrm{tr}\left(\frac{N_i}{C^2} \left(Y_i^m \begin{bmatrix} C-1 \\ -\vec{e}_{C-1} \end{bmatrix} \right) \left(Y_i^m \begin{bmatrix} C-1 \\ -\vec{e}_{C-1} \end{bmatrix} \right)^T \right) \\ &= \arg\max_{Y_i^m} \sum_{i=1}^{C} \mathrm{tr}\left(Y_i^m L_i^B (Y_i^m)^T \right), \end{aligned} \quad (38)$$

where,

$$L_i^B = \frac{N_i}{C^2} \begin{bmatrix} C-1 \\ -\vec{e}_{C-1} \end{bmatrix} \begin{bmatrix} C-1 & -\vec{e}_{C-1}^T \end{bmatrix} = \frac{N_i}{C^2} \begin{bmatrix} (C-1)^2 & -(C-1)\vec{e}_{C-1}^T \\ -(C-1)\vec{e}_{C-1} & \vec{e}_{C-1}\vec{e}_{C-1}^T \end{bmatrix};$$

$Y_i^m = \left[\vec{y}_i^m, \vec{y}_{i_1}^m, \cdots, \vec{y}_{i_{C-1}}^m \right]$; and $\vec{e}_{C-1} = [1, \cdots, 1]^T \in \mathbb{R}^{C-1}$.

Equation (38) can be seen as the whole alignment for all centroids, each of which has the part optimization: $\arg\max_{Y_i^m} \mathrm{tr}\left(Y_i^m L_i^B (Y_i^m)^T \right)$. The patch X_i^m is built by a given centroid $\vec{x}_i^m$ and the rest centroids of different classes $\vec{x}_{i_1}^m, \cdots, \vec{x}_{i_{C-1}}^m$. It is also global since it does not take the local geometry into account.

With L_i^B and (6), we can obtain the whole alignment:

$$\arg\max_{Y} \mathrm{tr}\left(Y^m L^B (Y^m)^T \right), \quad (39)$$

where $Y^m = \left[\vec{y}_1^m, \vec{y}_2^m, \cdots, \vec{y}_C^m \right]$. The above equation is equivalent to (36).

TABLE 2
SUMMARY OF THE MANIFOLD LEARNING ALGORITHMS

Algorithms	Patch: X_i	Representation of part optimization : L_i	Objective function
LLE/ NPE/ ONPP	Given measurement and its neighbours	$\begin{bmatrix} 1 & -\vec{c}_i^{\,T} \\ -\vec{c}_i & \vec{c}_i\vec{c}_i^{\,T} \end{bmatrix}$	Non-linear/ linear/ orthogonal linear
ISOMAP	Given measurement and the rest ones	$(1/N)\cdot\tau\left(D_G{}^i\right)$	Non-linear
LE/ LPP	Given measurement and its connected ones in the undirected graph	$\begin{bmatrix} \sum_{j=1}^{l}(\vec{w}_i)_j & -\vec{w}_i^{\,T} \\ -\vec{w}_i & \mathrm{diag}(\vec{w}_i) \end{bmatrix}$	Non-linear/ linear
LTSA/ LLTSA	Given measurement and its neighbours	$R_{k+1}-V_iV_i^T$, where V_i denotes d largest right singular vectors of X_iR_{k+1}	Non-linear/ linear
HLLE	Given measurement and its neighbours	$H_iH_i^T$	Non-linear

Note that one can refer to Table 1 for the explanations of the notations in this table.

TABLE 3
SUMMARY OF THE CONVENTIONAL LINEAR ALGORITHMS

Algorithms	Patch: X_i		Representation of part optimization : L_i	Objective function
PCA	Given measurement and the rest ones		$\frac{1}{N^2}\begin{bmatrix} (N-1)^2 & -(N-1)\vec{e}_{N-1}{}^T \\ -(N-1)\vec{e}_{N-1} & \vec{e}_{N-1}\vec{e}_{N-1}{}^T \end{bmatrix}$	Orthogonal linear
LDA	Within-class	Given measurement and the rest ones of a same class	$L_i{}^W=\frac{1}{N_i{}^2}\begin{bmatrix} (N_i-1)^2 & -(N_i-1)\vec{e}_{N_i-1}{}^T \\ -(N_i-1)\vec{e}_{N_i-1} & \vec{e}_{N_i-1}\vec{e}_{N_i-1}{}^T \end{bmatrix}$	Dual-objective optimization
	Between-class	Given centroid of one class and the rest ones of different classes	$L_i{}^B=\frac{N_i}{C^2}\begin{bmatrix} (C-1)^2 & -(C-1)\vec{e}_{C-1}{}^T \\ -(C-1)\vec{e}_{C-1} & \vec{e}_{C-1}\vec{e}_{C-1}{}^T \end{bmatrix}$	

Note that one can refer to Table 1 for the explanations of the notations in this table.

Considering (35) and (39) together, we get the special dual-objective optimization model which is different from the other algorithms in our framework:

$$\begin{cases} \arg\min\limits_{U} \mathrm{tr}\left(U^T X L^W X^T U\right) \\ \arg\max\limits_{U} \mathrm{tr}\left(U^T X^m L^B \left(X^m\right)^T U\right) \end{cases} \quad \text{s.t.}\, UU^T = I_d\,, \tag{40}$$

where, $X^m=\left[\vec{x}_1^m,\vec{x}_2^m,\cdots,\vec{x}_C^m\right]$ corresponds to Y^m. According to the fashion of the original LDA, we have the following criterion:

$$\arg\max_{U} \frac{\mathrm{tr}\left(U^T X^m L^B \left(X^m\right)^T U\right)}{\mathrm{tr}\left(U^T X L^W X^T U\right)} \quad \text{s.t.}\, UU^T = I_d\,. \tag{41}$$

The above optimization can be converted to solving the generalized eigenvalue problem as follows:

$$X^m L^B \left(X^m\right)^T \alpha = \lambda X L^W X^T \alpha\,, \tag{42}$$

and the optimal solutions are the d eigenvectors associated with d largest eigenvalues.

3.7 Discussions

Discovered by the proposed unifying framework, all algorithms have an almost identical whole alignment stage and intrinsic differences of them are how to build patches and the associated optimization, as shown both in the above sub-sections 3.1–3.6 in detail and in Tables 2 and 3 briefly. Based on this point of view, we have the following observations, which are helpful: i) to understand existing dimensionality reduction algorithms; and ii) to guide us to design new algorithms with specific properties for dimensionality reduction.

Observation 1: patches in manifold learning algorithms consider local geometry of measurements, while conventional linear algorithms do not. In detail, LLE, LTSA, and HLLE build each patch by a measurement and its nearest neighbours. Each patch of LE consists of two parts: i) a measurement x_i and its nearest neighbours and ii) measurements which deem x_i as their nearest neighbours. Each patch in PCA is built by all measurements in a dataset. For LDA, there are two types of patches: i) each first type patch is built by all measurements in a class and ii) each second type patch is built by all centroids of different classes. PCA and LDA build patches globally without considering the local geometry so they cannot discover the non-linear structure hidden in high-dimensional data. ISOMAP builds global patches each of which contains all the measurements like PCA. However, local geometry is

still considered by ISOMAP since geodesic distances used in the algorithm contain the neighbourhood information.

Observation 2: different types of geometry are preserved in patches. LLE preserves reconstruction coefficients, which are obtained in the original high-dimensional space for patch representation, in the low-dimensional subspace. LE preserves the nearby relations of a patch. Both LTSA and HLLE preserve local geometry represented by tangent coordinates of a patch. LTSA uses the linear transformation on tangent coordinates, while HLLE employs the quadratic form. ISOMAP preserves pairwise geodesic distances of measurements on patches. Among the manifold learning algorithms, LE has less computational cost than others because it only minimizes the sum of distances on local patches. PCA and LDA preserve the global geometry by minimizing or maximizing scatters of each patch.

4 DLA: Discriminative Locality Alignment

In this section, a new linear discriminative dimensionality reduction algorithm termed Discriminative Locality Alignment (DLA) is developed as an application of the proposed framework. In DLA, the discriminative information, i.e., labels of measurements, is imposed on the part optimization stage and then the whole alignment stage constructs the global coordinate in the projected low-dimensional subspace.

4.1 Part optimization

For a given measurement $\vec{x}_i$, according to the label information, we can divide the other measurements into two groups: measurements in the same class with $\vec{x}_i$ and measurements from different classes with $\vec{x}_i$. We select k_1 nearest neighbours with respect to $\vec{x}_i$ from measurements in the same class with $\vec{x}_i$ and term them Neighbour Measurements of a Same Class: $\vec{x}_{i^1},\cdots,\vec{x}_{i^{k_1}}$. We select k_2 nearest neighbours with respect to $\vec{x}_i$ from measurements in different classes with $\vec{x}_i$ and term them Neighbour Measurements of Different Classes denoted by $\vec{x}_{i_1},\cdots,\vec{x}_{i_{k_2}}$. By putting $\vec{x}_i$, $\vec{x}_{i^1},\cdots,\vec{x}_{i^{k_1}}$, and $\vec{x}_{i_1},\cdots,\vec{x}_{i_{k_2}}$ together, we can build the local patch for the measurement $\vec{x}_i$ as $X_i=\left[\vec{x}_i,\vec{x}_{i^1},\cdots,\vec{x}_{i^{k_1}},\vec{x}_{i_1},\cdots,\vec{x}_{i_{k_2}}\right]$.

For each patch, the corresponding output in the low-dimensional space is denoted by $Y_i=\left[\vec{y}_i,\vec{y}_{i^1},\cdots,\vec{y}_{i^{k_1}},\vec{y}_{i_1},\cdots,\vec{y}_{i_{k_2}}\right]$. In the low-dimensional space, we expect that distances between the given measurement and the Neighbour Measurements of a Same Class are as small as possible, while distances between the given measurement and the Neighbour Measurements of Different Classes are as large as possible. Fig. 2 illustrates this idea. The left part of the figure shows the i^{th} patch in the original high-dimensional space and the patch consists of $\vec{x}_i$, Neighbour Measurements of a Same Class (i.e., $\vec{x}_{i^1}$, $\vec{x}_{i^2}$, and $\vec{x}_{i^3}$), and Neighbour Measurements of Different Classes (i.e., $\vec{x}_{i_1}$ and $\vec{x}_{i_2}$). The expected results on the patch in the low-dimensional space are shown as the right part of the figure. Low-dimensional measurements $\vec{y}_{i^1}$, $\vec{y}_{i^2}$, and $\vec{y}_{i^3}$ are as close as possible to $\vec{y}_i$, while low-dimensional measurements $\vec{y}_{i_1}$ and $\vec{y}_{i_2}$ are as far as possible away from $\vec{y}_i$.

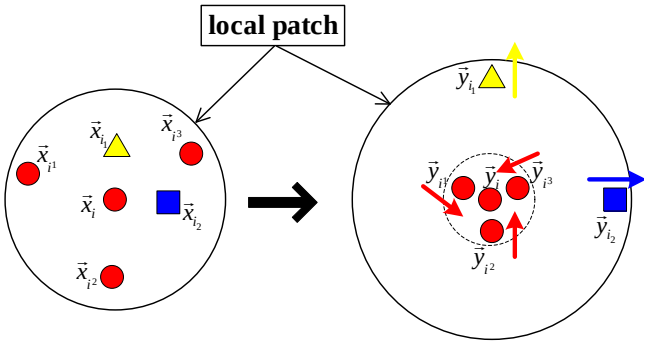


Fig. 2. Part optimization of DLA. The measurements with the same shape and color come from the same class.

For each patch in the low-dimensional subspace, we expect that distances between $\vec{y}_i$ and the Neighbour Measurements of a Same Class are as small as possible, so we have:

$$\arg\min_{\vec{y}_i}\sum_{j=1}^{k_1}\left\|\vec{y}_i-\vec{y}_{i^j}\right\|^2. \tag{43}$$

Meanwhile, we expect that distances between $\vec{y}_i$ and the Neighbour Measurements of Different Classes are as large as possible, so we have:

$$\arg\max_{\vec{y}_i}\sum_{p=1}^{k_2}\left\|\vec{y}_i-\vec{y}_{i_p}\right\|^2. \tag{44}$$

Since the patch formed by the local neighbourhood can be regarded approximately linear [27], we formulate the part discriminator by using the linear manipulation as follows:

$$\arg\min_{\vec{y}_i}\left(\sum_{j=1}^{k_1}\left\|\vec{y}_i-\vec{y}_{i^j}\right\|^2-\beta\sum_{p=1}^{k_2}\left\|\vec{y}_i-\vec{y}_{i_p}\right\|^2\right), \tag{45}$$

where β is a scaling factor in $[0,1]$ to unify different measures of the within-class distance and the between-class distance. Define the coefficients vector

$$\vec{\omega}_i=\left[\overbrace{1,\ \cdots,\ 1}^{k_1},\ \overbrace{-\beta,\ \cdots,\ -\beta}^{k_2}\right]^T, \tag{46}$$

then, (45) reduces to:

$$\begin{aligned}
&\arg\min_{\vec{y}_i}\left(\sum_{j=1}^{k_1}\left\|\vec{y}_i-\vec{y}_{i^j}\right\|^2(\vec{\omega}_i)_j+\sum_{p=1}^{k_2}\left\|\vec{y}_i-\vec{y}_{i_p}\right\|^2(\vec{\omega}_i)_{p+k_1}\right)\\
&=\arg\min_{\vec{y}_i}\left(\sum_{j=1}^{k_1+k_2}\left\|\vec{y}_{F_i(1)}-\vec{y}_{F_i(j+1)}\right\|^2(\vec{\omega}_i)_j\right)\\
&=\arg\min_{Y_i}\operatorname{tr}\left(Y_i\begin{bmatrix}-\vec{e}_{k_1+k_2}^{\,T}\\ I_{k_1+k_2}\end{bmatrix}\operatorname{diag}(\vec{\omega}_i)\begin{bmatrix}-\vec{e}_{k_1+k_2} & I_{k_1+k_2}\end{bmatrix}Y_i^T\right)\\
&=\arg\min_{Y_i}\operatorname{tr}\left(Y_iL_iY_i^T\right),
\end{aligned} \tag{47}$$

where

$$L_i=\begin{bmatrix}\sum_{j=1}^{k_1+k_2}(\vec{\omega}_i)_j & -\vec{\omega}_i^T\\ -\vec{\omega}_i & \operatorname{diag}(\vec{\omega}_i)\end{bmatrix}; \tag{48}$$

$F_i=\left\{i,i^1,\cdots,i^{k_1},i_1,\cdots,i_{k_2}\right\}$ is the set of indices for measurements on the patch; $\vec{e}_{k_1+k_2}=[1,\cdots,1]^T\in\mathbb{R}^{k_1+k_2}$; and $I_{k_1+k_2}$ is a $(k_1+k_2)\times(k_1+k_2)$ identity matrix.

4.2 Whole alignment

With the constructed part optimization L_i, the matrix L can be built to achieve the whole alignment by (6). To obtain the linear and orthogonal projection matrix U with d columns, the objective function is designed as (9) and then the problem is converted to a standard eigenvalue problem. Different from algorithms, e.g., LDA, LPP, and Marginal Fisher Analysis (MFA) [37], which lead to a generalized eigenvalue problem, DLA successfully avoids the matrix singularity problem since it has no inverse operation over a matrix. However, the PCA step is still recommended to reduce noise. The procedure of the proposed DLA is listed as follows:

1. Use PCA to project the dataset X to the subspace by eliminating useless information. To keep it simple, we still use X to denote the dataset in the PCA subspace in the following steps. We denote by U_{PCA} the PCA projection matrix. Note that this step is optional.
2. For each measurement $\vec{x}_i$ in dataset X, $i=1,\cdots,N$, search k_1 Neighbour Measurements of a Same Class and k_2 Neighbour Measurements of Different Classes, and then build the patch $X_i=[\vec{x}_i,\vec{x}_{i_1},\cdots,\vec{x}_{i_{k_1}},\vec{x}_{i_1},\cdots,\vec{x}_{i_{k_2}}]$.
3. Compute the matrix L_i, by (48), construct the alignment matrix L by the iterative procedure described as (6).
4. Solve the standard eigenvalue problem: $XLX^T\vec{u}=\lambda\vec{u}$ to obtain the DLA projection matrix $U_{DLA}=[\vec{u}_1,\vec{u}_2\cdots,\vec{u}_d]$, whose vectors are the eigenvectors corresponding to the d smallest eigenvalues. The final projection matrix is as follows:

$$U=U_{PCA}U_{DLA}. \tag{49}$$

5 Experiments

This section evaluates the performance of the proposed DLA in comparison with six representative algorithms, i.e., PCA [34], Generative Topographic Mapping (GTM) [6], [7] Probabilistic Kernel Principal Components Analysis (PKPCA) [33], [40], LDA [2], SLPP (LPP1 in [10]) and MFA [37], on three face image databases, i.e., YALE [2], UMIST [15], and FERET [24]. Among these algorithms, PCA, PKPCA and GTM are unsupervised algorithms which do not consider the class label information. A semi-linear GTM described in [7] is used, since the standard GTM [6] grows exponentially with number of latent dimensions and lead to a high computational complexity in real applications. SLPP and MFA are recently proposed manifold learning based algorithms. All face images from three databases were cropped with reference to the eyes and cropped images were normalized to the 40×40 pixel arrays with 256 gray levels per pixel. Each image was reshaped to one long vector by arranging its pixel values in a fixed order.

All datasets constructed from each database were randomly divided into three separate sets: *training set*, *validation set* and *testing set*. Training set was used to learn the low-dimensional subspace along with the projection matrix. Validation set was used to determine the optimal parameters in algorithms. For the proposed DLA algorithm, the important parameters include k_1 (the number of Neighbour Measurements of a Same Class), k_2 (the number of Neighbour Measurements of Different Classes), and d (the subspace dimension). Testing set was used to report the final recognition accuracy. During both validation and testing phases, the Nearest Neighbour (NN) rule was used in classification.

Fig. 3. Sample images from YALE.

For all algorithms but PCA itself, the first step is the PCA projection. Because the number of measurements is often much smaller than the dimension of measurements, i.e., $N \ll m$, we retain $N-C$ dimensions in the PCA step to ensure S_W (referring to [2]) in LDA and $X\left(D^p-W^p\right)X^T$ (referring to [37]) in MFA nonsingular for both LDA and MFA algorithms. Since DLA has no singularity problem, in the algorithm PCA subspace is set $N-1$ dimensions and all the energies can be preserved in this step. The same is true of KPCA and GTM. In SLPP, we only need to ensure XX^T is full rank, so we can retain $N-1$ (referring to [10]) dimensions in PCA subspace.

5.1 YALE

The YALE database [2] contains face images collected from fifteen individuals, eleven images for each individual and showing varying facial expressions and configurations. Fig. 3 shows the image set for one individual. For training, we randomly selected different numbers (3, 5, 7, 9) of images per individual, used 1/2 of the rest images for validation, and 1/2 of the rest images for testing. Such trial was independently performed ten times, and then the average recognition results were calculated. Fig. 4 shows the average recognition rates versus subspace dimensions on the validation sets, which help to select the best subspace dimension. Table 4 reports the final recognition rates (%) on the testing sets. It can be seen that DLA outperforms the other algorithms. Table 4 also provides the optimal values of the parameters k_1 and k_2 for DLA, which are crucial because they have the special sense for building the local patches. In Section 5.5, we will describe how to determine k_1 and k_2 for DLA.

5.2 UMIST

The UMIST database [15] consists of a total of 564 face images of twenty people. The individuals are a mix of race, sex and appearance and are photographed in a range of poses from profile to frontal views. Fig. 5 shows some images of an individual. For each individual, different number (3, 5, 7, 9) of images were randomly selected for training, 1/2 of the rest were used for validation and 1/2 of the rest were used for testing. We repeated these trials ten times and computed the average results. Fig. 6

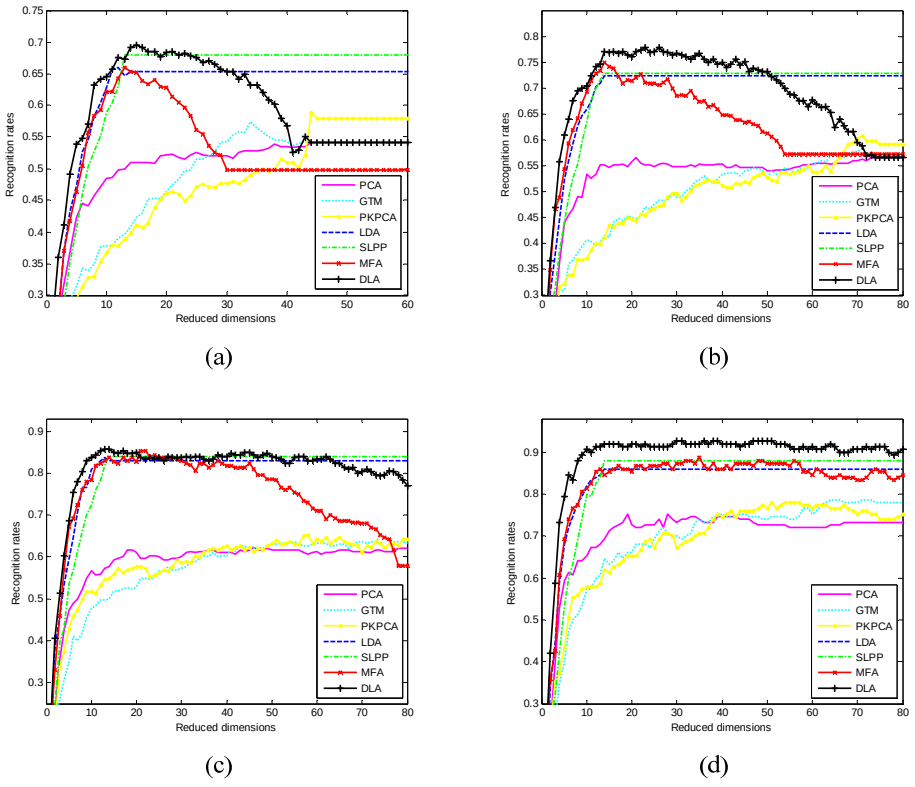


Fig. 4. Recognition rate vs. subspace dimension on the validation sets of YALE. (a) three measurements for training. (b) five measurements for training. (c) seven measurements for training. (d) nine measurements for training.

TABLE 4
BEST RECOGNITION RATES (%) OF SEVEN ALGORITHMS ON THE TESTING SETS OF YALE

Number of Training	3	5	7	9
PCA	50.50 (44)	56.44(21)	64.00 (46)	62.67 (19)
GTM	52.00 (34)	60.22 (71)	65.33 (74)	66.00 (64)
PKPCA	53.50 (44)	60.44 (71)	66.67 (58)	67.33 (51)
LDA	61.33 (12)	73.11 (14)	76.67 (13)	80.00 (14)
SLPP	64.17 (13)	74.00 (14)	80.67 (14)	80.67(14)
MFA	61.00 (13)	74.67 (14)	79.67 (21)	82.00 (35)
DLA	66.83 (15, 2, 2)	78.89 (23, 2, 1)	81.33 (13, 4, 2)	85.33 (30, 4, 4)

For PCA, PKPCA, GTM, LDA SLPP, and MFA, the numbers in the parentheses are the selected subspace dimensions. For DLA, the first numbers in the parentheses are the selected subspace dimensions, the second and the third numbers are the parameters k_1 *and* k_2 *, respectively.*

Fig. 5. Sample images from UMIST.

shows the recognition rates versus subspace dimensions on the validation sets and Table 5 lists the final recognition rates (%) on the testing sets. Again, DLA performs better than the other algorithms.

5.3 FERET

The FERET database [24] contains a total of 13,539 face images of 1,565 subjects. The images vary in size, pose, illumination, facial expression and age. We randomly selected 100 individuals, having seven images. Fig. 7 shows images of one individual. We randomly chose different number (3, 5) of images per individual for training, 1/2 of the rest were used for validation and 1/2 of the rest were used for testing. All the trials were repeated ten times, and we then calculated the average recognition results. The recognition rates versus subspace dimensions on the validation sets are given in Fig. 8 and the final recognition rates (%) on the testing sets are list in Table 6. DLA outperforms the other algorithms.

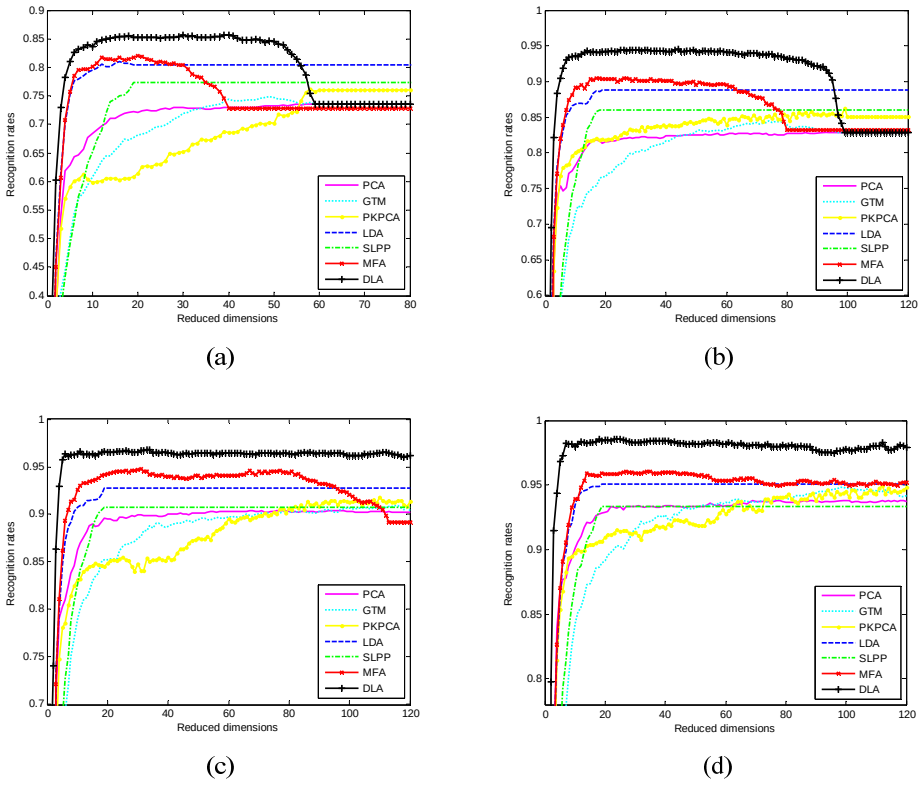


Fig. 6. Recognition rate vs. subspace dimension on the validation sets of UMIST. (a) three measurements for training. (b) five measurements for training. (c) seven measurements for training. (d) nine measurements for training.

TABLE 5
BEST RECOGNITION RATES (%) OF SEVEN ALGORITHMS ON THE TESTING SETS OF UMIST

Number of Training	3	5	7	9
PCA	71.13 (58)	82.86 (99)	90.62 (71)	93.79 (69)
GTM	73.17 (49)	84.08 (74)	90.93 (112)	94.98 (98)
PKPCA	74.04(58)	84.82 (99)	93.16 (110)	95.02 (112)
LDA	78.23 (16)	88.24 (19)	93.82 (19)	95.10 (19)
SLPP	75.58(19)	85.92 (19)	91.91 (19)	93.36 (19)
MFA	79.55 (20)	90.00 (17)	94.76 (31)	96.05 (27)
DLA	84.04 (40, 2, 1)	93.35 (44, 2, 2)	96.76 (33, 4, 5)	98.58 (24, 6, 5)

For PCA, PKPCA, GTM, LDA SLPP, and MFA, the numbers in the parentheses are the selected subspace dimensions. For DLA, the first numbers in the parentheses are the selected subspace dimensions, the second and the third numbers are the parameters k_1 *and* k_2 *, respectively.*

Fig. 7. Sample images from FERET.

5.4 Building patches

In this subsection, we study effects of k_1 (the number of Neighbour Measurements of a Same Class) and k_2 (the number of Neighbour Measurements of Different Classes) on the recognition rates in the validation phase based on the YALE database with nine measurements in each class for training. The selected subspace dimension was fixed to 30.

By fixing k_2 to an arbitrary value and varying k_1 from 1 to $N_i - 1$ (= 8), we can obtain the recognition rate curve with respect to k_1 as shown in Fig. 9a. There is a peak on the curve when $k_1 = 4$. By fixing $k_1 = 4$ and varying k_2 from 0 to $N - N_i$ (= 126), another recognition rate curve with respect to k_2 can be obtained as shown in Fig. 9b. There is a peak on the curve when $k_2 = 4$. By varying k_1 from 1 to 8 and k_2 from 0 to 126 simultaneously, the recognition rate surface can be obtained as shown in Fig. 10. In this figure, there is a peak which corresponds to $k_1 = 4$ and $k_2 = 4$, so, the local neighbourhood in DLA characterizes the discriminability better than global structure.

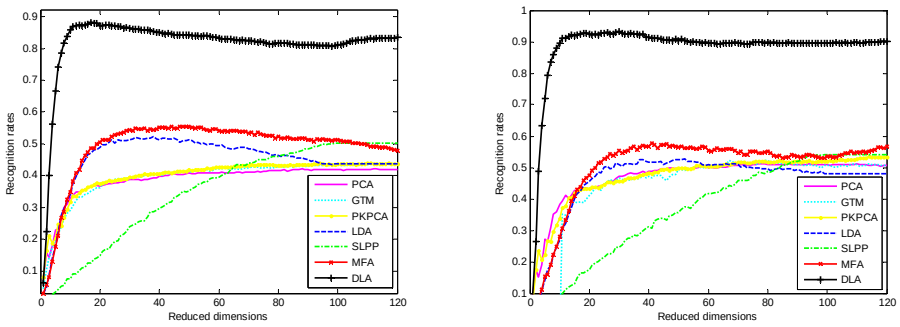


Fig. 8. Recognition rate vs. subspace dimension on the validation sets of FERET. (a) three measurements for training. (b) five measurements for training.

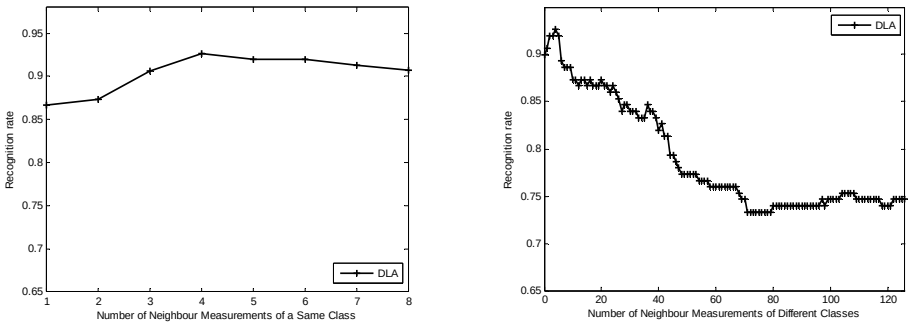


Fig. 9. Building patches. (a) recognition rate vs. number of Neighbour Measurements of a Same Class. (b) recognition rate vs. number of Neighbour Measurements of Different Classes.

TABLE 6
BEST RECOGNITION RATES (%) OF SEVEN ALGORITHMS ON THE TESTING SETS OF FERET

Number of Training	3	5
PCA	41.45 (88)	51.20 (78)
GTM	43.15 (114)	52.90 (67)
PKPCA	43.65 (94)	53.10 (114)
LDA	50.05 (38)	54.50 (52)
SLPP	48.95 (99)	55.10 (99)
MFA	55. 05 (48)	57. 20 (41)
DLA	87. 90 (17, 1, 6)	93.00 (30, 2, 6)

For PCA, PKPCA, GTM, LDA SLPP, and MFA, the numbers in the parentheses are the selected subspace dimensions. For DLA, the first numbers in the parentheses are the selected subspace dimensions, the second and the third numbers are the parameters k_1 and k_2, respectively.

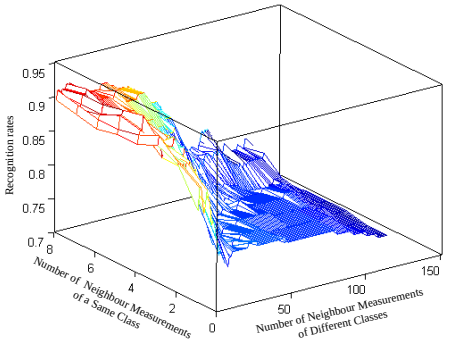


Fig. 10. Recognition rates vs. the number of Neighbour Measurements of a Same Class and the number of Neighbour Measurements of Different Classes.

5.5 Discussion

Based on the experimental results in the subsections 5.1–5.4, we have the following observations:

1. DLA considers both the discriminative information and the locality of measurements. Therefore it works better that LDA, PCA, and SLPP. Although MFA takes these two aspects into account, it might not be as good as DLA in terms of classification accuracy. This is because some discriminative information is discarded by the PCA step in MFA;
2. in experiments on building patches, by setting $k_1 = 8$ and $k_2 = 126$, DLA is similar to LDA because the global structure is considered. With this setting, DLA ignores the local geometry and performs poorly for classification. Therefore, by setting k_1 and k_2 suitably, DLA captures both the local geometry and the dis-

criminative information of measurements. It is not necessary to traverse all possible values of k_1 and k_2 for parameter selection because k_1 and k_2 are usually small to represent the locally Euclidean property; and

3. it is demonstrated by Figs. 4a, 4b, 4c, 6a, and 6b: when an optimal value of DLA subspace dimension is achieved (i.e., classification error is minimized), the classification accuracy decreases fast while the dimension of DLA subspace is increased. This is because the effective DLA subspace is determined by the rank of XLX^T (as described in (9)), which depends on the cardinality of the training set (N) if $m >> N$. In Figs. 4a, 4b, 4c, 6a, and 6b, cardinalities of training sets are 45, 75, 105, 60, and 100, respectively. These low cardinalities limit the dimension of effective DLA selected subspaces and thus the classification accuracies drop quickly after the dimension of DLA subspace achieves an optimal value.

6 Conclusions

In this paper, a unifying framework, "patch alignment", was proposed as a powerful analysis and development tool for dimensionality reduction. It implements the idea, "part optimization and whole alignment". The proposed framework was first applied to reformulate various existing spectral analysis based dimensionality reduction algorithms into a unified form, allowing the different algorithms to be analyzed and compared. Different algorithms were shown to construct whole alignment matrices (for global coordinate construction in the subspace) in an almost identical way, but vary with patch optimizations (associated with different measurements and have different objectives on these built patches).

Based on this framework, we developed a new discriminative dimensionality reduction algorithm, Discriminative Locality Alignment (DLA). DLA can be seen as a special case of the framework and it: 1) overcomes the nonlinear distribution of measurements; 2) preserves the discriminative ability; and 3) avoids the matrix singularity problem. Experimental results show the effectiveness of DLA on YALE, UMIST, and FERET face image databases in comparison with popular dimensionality reduction algorithms.

Acknowledgment

The authors thank the handling Associate Editor Prof. Sameer Singh and four anonymous reviewers for their constructive comments on this paper. The first author thanks Mr. Deli Zhao (CUHK) for beneficial discussions and selfless help. The work was supported by National Science Foundation of China (No. 60675023) and Chinese National 863 High Technology Plan (No. 2007AA01Z164).

References

[1] Y. Aslandogan and C. Yu, "Techniques and Systems for Image and Video Retrieval," *IEEE Trans. Knowledge and Data Engineering*, vol. 11, no. 1, pp. 56- 63, 1999.

[2] P. Belhumeour, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[3] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing System*, vol. 14, pp. 585-591, 2002.

[4] R. Bellman, *Adaptive Control Processes*: *A Guided Tour*. Princeton University Press, 1961.

[5] Y. Bengio, J. Paiement, P. Vincent, O. Dellallaeu, L. Roux , and M. Quimet, "Out-of sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering," *Advances in Neural Information Processing System*, vol. 16, 2004.

[6] C.M. Bishop, M. Svensén, and C.K. I. Williams, "GTM: The Generative Topographic Mapping," *Neural Computation*, vol. 10, no. 1, pp. 215-234, 1998.

[7] C.M. Bishop, M. Svensén, and C.K. I. Williams, "Developments of the Generative Topographic Mapping," *Neurocomputing*, vol. 21, 203-224, 1998.

[8] D. Cai, X. He and J, Han, "Document Clustering Using Locality Preserving Indexing," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 12, pp. 1624-1637, 2005.

[9] D. Cai, X. He and J, Han, "SRDA: an Efficient Algorithm for Large Scale Discriminant Analysis," *IEEE Trans. Knowledge and Data Engineering*, vol. 20, no. 1, pp. 1-12, 2008.

[10] D. Cai, X. He and J, Han, "Using Graph Model for Face Analysis," Department of Computer Science Technical Report No. 2636, University of Illinois at Urbana-Champaign, Sept. 2005.

[11] D.L. Donoho and C. Grimes, "Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-dimensional Data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591-5596, 2003.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed., Wiley, 2000.

[13] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[14] B. Gao, T. Liu, G. Feng, T. Qin, Q. Cheng, and W. Ma, "Hierarchical Taxonomy Preparation for Text Categorization Using Consistent Bipartite Spectral Graph Copartitioning," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1263-1273, 2005.

[15] D.B. Graham and N.M. Allinson, "Characterizing Virtual Eigensignatures for General Purpose Face Recognition, " in *Face Recognition*: *From Theory to Applications*, NATO ASI Series F, Computer and Systems Science, vol.163, H. Wechsler, P.J. Pillips, V. Bruce, F. Fogelman-Soulie and T.S. Huang, eds. Springer, 1998, pp.446-456.

[16] X. He, D. Cai, and J. Han, "Learning a Maximum Margin Subspace for Image Retrieval," *IEEE Trans. Knowledge and Data Engineering*, vol. 20, no. 2, pp. 189-201, 2008.

[17] X. He, D. Cai, J. Wen, W. Ma, and H. Zhang, "Clustering and Searching WWW Images Using Link and Page Layout Analysis," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 3, no. 2, 2007.

[18] X. He, D. Cai, S. Yan, and H. Zhang, "Neighborhood Preserving Embedding," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1208-1213, 2005.

[19] X. He and P. Niyogi, "Locality Preserving Projections," *Advances in Neural Information Processing System*, vol. 16, 2004.

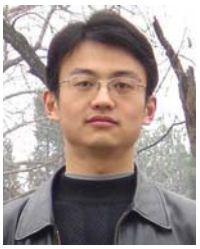
[20] H. Hotelling, "Analysis of A Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, vol. 24, pp. 417-441, 1933.

[21] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer-Verlag, 2002.

[22] E. Kokiopoulou and Y. Saad, "Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143-2156, 2007.

[23] Y. Koren and L. Carmel, "Robust Linear Dimensionality Reduction," *IEEE Trans. Visualization and Computer Graphics*, vol. 10, no. 4, pp. 459-470, 2004.

[24] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.
[25] S. Rosenberg, *The Laplacian on a Riemannian Manifold*. Cambridge University Press, 1997.
[26] S.T. Roweis and L.K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
[27] L.K. Saul and S.T. Roweis, "Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifold," *Journal of Machine Learning Research*, vol. 4, pp.119-155, 2003.
[28] L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee, "Spectral Methods for Dimensionality Reduction," In *Semisupervised Learning*, O. Chapelle, B. Schoelkopf, and A. Zien, eds. MIT Press, 2006.
[29] G. Shakhnarovich and B. Moghaddam, "Face Recognition in Subspaces," *Handbook of Face Recognition*, Stan Z. Li and Anil K. Jain, Eds. Springer-Verlag, 2004.
[30] D. Tao, X. Li, and S.J. Maybank, "Negative Samples Analysis in Relevance Feedback," *IEEE Trans. Knowledge and Data Engineering*, vol. 19, no. 4, pp. 568-580, 2007.
[31] D. Tao, X. Li, X. Wu, and S.J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700-1715, 2007.
[32] J. Tenenbaum, V. Silva, and J. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, pp. 2319-2323, 2000.
[33] M.E. Tipping and C.M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society B*, vol. 21, no. 3, pp. 611-622, 1999.
[34] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 586-591, 1991.
[35] D. Xu, S. Lin, S. Yan, and X. Tang, "Rank-one Projections with Adaptive Margin for Face Recognition, " *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 37, no. 5, pp. 1226-1236, 2007.
[36] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Multilinear Discriminant Analysis for Face Recognition," *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 212-220, 2007.
[37] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.
[38] T. Zhang, J. Yang, D. Zhao, and X. Ge, "Linear Local Tangent Space Alignment and Application to Face Recognition," *Neurocomputing*, vol. 70, pp. 1547-1553, 2007.
[39] Z. Zhang and H. Zha, "Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment," *SIAM J. Scientific Computing*, vol. 26, no. 1, pp. 313-338, 2005.
[40] S. Zhou, "Probabilistic Analysis of Kernel Principal Components: Mixture Modeling and Classification," CfAR Technical Report, CAR-TR-993, 2003.



Tianhao Zhang received the PhD degree in Pattern Recognition and Intelligence System from Shanghai Jiao Tong University in 2008. He is currently a Postdoctoral Researcher at the Section of Biomedical Image Analysis (SBIA), Department of Radiology, University of Pennsylvania. He has spent one year at Department of Computing, the Hong Kong Polytechnic University as a Research Assistant (Visiting PhD) from 2007 to 2008. His research interests include machine learning, computer vision and medical image analysis. He has published extensively in the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Systems, Man and Cybernetics, Part B (TSMC-B), Pattern Recognition (PR), the European Conference on Computer Vision (ECCV), the International Joint Conference on Neural Networks (IJCNN), etc. He is a reviewer for TSMC-B, Neurocomputing, and IJCM. He has been a program committee of PAKDD 2009, ICIAR 2008, PSIVT 2007 and 2008, DMAMH 2007, and a workshop in ICDM 2008. He also reviews for IJCNN 2008 and ICASSP 2009.

Dacheng Tao (M'07) received the B.Eng. degree from the University of Science and Technology of China (USTC), the MPhil degree from the Chinese University of Hong Kong (CUHK), and the PhD degree from the University of London (Lon). Currently, he is a Nanyang Assistant Professor with the School of Computer Engineering in the Nanyang Technological University, a Visiting Professor in Xi'Dian University, a Guest Professor in Wuhan University, and a Visiting Research Fellow at Birkbeck in Lon. His research is mainly on applying statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and visual surveillance. He has published more 90 scientific papers including IEEE TPAMI, TKDE, TIP, TMM, TCSVT, TSMC, CVPR, ECCV, ICDM; ACM TKDD, Multimedia, KDD etc., with one best paper runner up award. Previously he gained several Meritorious Awards from the International Interdisciplinary Contest in Modeling, which is the highest level mathematical modeling contest in the world, organized by COMAP. He is an associate editor of Neurocomputing (Elsevier) and the Official Journal of the International Association for Statistical Computing -- Computational Statistics & Data Analysis (Elsevier). He has authored/edited six books and eight special issues, including CVIU, PR, PRL, SP, and Neurocomputing. He has (co-)chaired for special sessions, invited sessions, workshops, and conferences. He has served with more than 50 major international conferences including CVPR, ICCV, ECCV, ICDM, KDD, and Multimedia, and more than 15 top international journals including TPAMI, TKDE, TOIS, TIP, TCSVT, TMM, TIFS, TSMC-B, Computer Vision and Image Understanding (CVIU), and Information Science. He is a member of IEEE, IEEE SMC Society, IEEE Signal Processing Society, and IEEE SMC Technical Committee on Cognitive Computing.

Xuelong Li (M'02-SM'07) holds a permanent post at Birkbeck College, University of London and a visiting/guest professorship at Tianjin University and University of Science and Technology of China. His research focuses on cognitive computing, image/video processing, pattern recognition, and multimedia. His research activities are partly sponsored by EPSRC, the British Council, Royal Society, and the Chinese Academy of Sciences. He has over a hundred scientific publications with several Best Paper Awards and finalists. He is an author/editor of four books, an associate editor of IEEE Trans. on Image Processing, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Systems, Man and Cybernetics Part B, and IEEE Trans. on Systems, Man and Cybernetics Part C. He is also an associate editor (editorial board member) of ten other international journals and a guest co-editor of eight special issues. He has served as a chair of around twenty conferences and a program committee member for more than eighty conferences. He has been a reviewer for over a hundred journals and conferences, including eleven IEEE transactions. He is a academic committee member of the China Society of Image and Graphics, a senior member of the IEEE, the chair of IEEE Systems, Man and Cybernetics Society Technical Committee on Cognitive Computing, and a member of several other technical committees of IEEE Systems, Man and Cybernetics Society and IEEE Signal Processing Society Technical Committee on Machine Learning for Signal Processing (MLSP). He is a Chapters Coordinator of the IEEE Systems, Man and Cybernetics Society.

Jie Yang is the Professor and Director of Institute of Image Processing and Pattern recognition in Shanghai Jiao Tong University, Shanghai, China. He was born in Shanghai, in August 1964. In 1985, he received his bachelor degree in Automatic Control in Shanghai Jiao Tong University, where a master degree in Pattern Recognition and Intelligent System was achieved three years later. In 1989, as one of the nation's first-class graduate students, he was enrolled by the Department of Computer Science, University of Hamburg, Germany. In 1994, he came back with a PhD degree. He has been the principal investigator of more than 30 nation and ministry scientific research projects in image processing, pattern recognition, data mining, and artificial intelligence, including two 973 projects, three national 863 projects, four NSFC projects, three international cooperative projects with institutions from France, Korea, and Japan. He has published more than three hundreds of articles in national or international academic journals and conferences. Up to now, he has guided 3 postdoctoral, 28 doctors and 38 masters, awarded four research achievement prize from ministry of Education, China and Shanghai municipality.