Hybrid Manifold Embedding

Yang Liu, Yan Liu, Keith C. C. Chan, *Member, IEEE*, and Kien A. Hua, *Fellow, IEEE*

***Abstract*— In this brief, we present a novel supervised manifold learning framework dubbed hybrid manifold embedding (HyME). Unlike most of the existing supervised manifold learning algorithms that give linear explicit mapping functions, the HyME aims to provide a more general nonlinear explicit mapping function by performing a two-layer learning procedure. In the first layer, a new clustering strategy called geodesic clustering is proposed to divide the original data set into several subsets with minimum nonlinearity. In the second layer, a supervised dimensionality reduction scheme called locally conjugate discriminant projection is performed on each subset for maximizing the discriminant information and minimizing the dimension redundancy simultaneously in the reduced low-dimensional space. By integrating these two layers in a unified mapping function, a supervised manifold embedding framework is established to describe both global and local manifold structure as well as to preserve the discriminative ability in the learned subspace. Experiments on various data sets validate the effectiveness of the proposed method.**

***Index Terms*— Dimensionality reduction, geodesic clustering (GC), hybrid manifold embedding (HyME), locally conjugate discriminant projection (LCDP), supervised manifold learning.**

I. Introduction

Manifold learning, which aims to discover the compact representation of high-dimensional data based on the assumption that the original observations lie on or close to a low-dimensional nonlinear manifold, has received much attention recently [1], [5], [18], [28], [30], [31], [37]. Many real-world data sets match this manifold assumption very well in their original observed spaces [26], [27] and thus make manifold learning become an appropriate tool in various data analysis applications. Motivated by the success of manifold learning in unsupervised learning tasks, such as data embedding and visualization, the supervised versions of manifold learning have also been explored [12], [32].

The above manifold learning models, however, do not provide explicit mapping functions and thus cannot project new data to the low-dimensional space without reembedding the entire data set. Based on this observation, many linear manifold learning algorithms have been proposed [4], [11], [16], [25], [33]–[35], [38], [39], [41], [42]. These algorithms introduce linear explicit mapping functions into the objective formulas, and thus make the projection of new data straightforward.

Although the linear mapping functions provide an easy way to project test data points to the low-dimensional space, some of the manifold information contained in the original data sets will be lost inevitably, even if the objective functions are formulated for preserving the nonlinear structure. Moreover, the data sets in many learning tasks are very complicated so that a single linear learning model might be too limited to capture the subtleties of the data sets [7], [21], [23], [36]. In order to analyze complex nonlinear data, models with more powerful ability in data representation are needed. For example, Bregler and Omohundro [2] proposed a nonlinear manifold learning technique based on hidden Markov models for visual speech recognition. Two multilevel frameworks, i.e., the algebraic framework and the geometric framework, for unsupervised manifold learning were introduced in [8] and [9].

To capture the nonlinear information while inheriting the easy mapping property from the linear manifold learning algorithms, we propose a supervised manifold learning framework dubbed hybrid manifold embedding (HyME), which aims to provide a nonlinear explicit mapping function by performing a two-layer learning procedure. In the first layer, a new clustering strategy called geodesic clustering (GC) is proposed to divide the original data set into several subsets with minimum nonlinearity. This step could be viewed as a coarse learning of the nonlinear structure on the global level. Then, in the second layer, a supervised learning scheme called locally conjugate discriminant projection (LCDP) is carried out on each subset for maximizing the discriminant information, whereas minimizing the dimension redundancy in the reduced subsapce. This step could be considered as a fine learning of the nonlinear details from the local perspective. After performing LCDP, the subtleties of the data set ignored by the GC should be well captured. Since GC is a manifold-based clustering and LCDP is a manifold-based projection, the proposed framework that integrates GC and LCDP is called HyME.

Fig. 1 illustrates the idea behind HyME. Given the 2-D data set composed of three classes, the target of supervised dimensionality reduction (DR) is to find the optimal projection direction, on which the data points from different classes are well separated. Fig. 1(a) shows the projection direction found by maximizing the ratio of between-class scatter to the within-class scatter. Since the data set is nonlinear, it is impossible to find a globally linear projection direction that can separate the three classes, no matter what kind of local information or manifold regularization has been incorporated into the objective function. Fig. 1(b) gives the desired projection direction generated by HyME. Using the proposed GC, the entire data set is divided into two approximately linear subsets. Then, the proposed LCDP is performed on each subset and the optimal projection directions $\mathbf{w}_1$ for subset 1 and $\mathbf{w}_2$ for subset 2 are

Manuscript received July 11, 2013; accepted February 5, 2014. Date of publication February 25, 2014; date of current version November 17, 2014. This work was supported by 4.61.xx.681C Cognitive Computing Laboratory.

Y. Liu, Y. Liu, and K. C. C. Chan are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: csygliu@comp.polyu.edu.hk; csyliu@comp.polyu.edu.hk; cskcchan@comp.polyu.edu.hk).

K. A. Hua is with the Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: kienhua@cs.ucf.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2305760

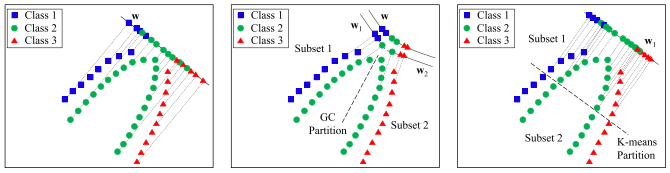

Fig. 1. Schematic illustration of the main idea behind HyME. (a) Globally linear projection direction $\mathbf{w}$. (b) Partition result of GC and globally nonlinear projection direction $\mathbf{w}$ of HyME. (c) Partition result of k-means clustering and projection direction $\mathbf{w}_1$ of subset 1.

learned, respectively. By combining $\mathbf{w}_1$ and $\mathbf{w}_2$, the data points from different classes are well separated on the globally nonlinear direction $\mathbf{w}$.

The rest of this brief is organized as follows. Section II formulates the proposed HyME. In Section III, we evaluate HyME on standard data sets and compare it with other representative supervised manifold learning algorithms. Conclusions are given in Section IV.

II. Hybrid Manifold Embedding

Given the training data set $\mathcal{X} = \{(\mathbf{x}_i, l_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the ith data point, and $l_i \in \{1, \ldots, L\}$ is the corresponding class label. In order to provide a nonlinear explicit mapping function, the proposed HyME performs a two-layer supervised learning procedure. In the first layer, a new algorithm GC is presented to divide $\mathcal{X}$ into C subsets $\mathcal{X}_1, \ldots, \mathcal{X}_C$, each with nonlinearity minimized. In the second layer, for each subset $\mathcal{X}_c$ ($c = 1, \ldots, C$), an individual transformation matrix $\mathbf{W}_c \in \mathbb{R}^{D\times d}$ is learned by a novel supervised DR algorithm LCDP, where D denotes the dimension of original data and d denotes the dimension of projected data. Each $\mathbf{W}_c$ maximizes the discriminant information contained in the corresponding subset $\mathcal{X}_c$ and minimizes the dimension redundancy via the conjugate constraint. Finally, a unified mapping function of HyME is formulated to integrate these two layers.

A. Geodesic Clustering

Some recent work showed that the data sets can be very complex in many real-world learning tasks, and using a single global model does not have the delicacy of discrimination to learn all of the subtleties involved [7], [21], [23], [36]. In order to enhance learnability of the proposed model, we investigate GC to minimize the nonlinearity of each subset of the data set. Here, the nonlinearity of a subset is measured by the maximum geodesic distance of that subset. The objective function of GC is given as follows:

$$P = \arg\min_{P} \max_{c=1,\ldots,C} \max_{i,j:\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}_c} \text{dist}_G(\mathbf{x}_i, \mathbf{x}_j) \tag{1}$$

where P denotes a partition of the data set, C the number of subsets, $\mathcal{X}_c$ the cth subset, and $\text{dist}_G(\mathbf{x}_i, \mathbf{x}_j)$ the geodesic distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ (approximated by the shortest path [31] between $\mathbf{x}_i$ and $\mathbf{x}_j$).

First, we build a faithful neighborhood graph by employing a two-way connection criterion [22]: we connect $\mathbf{x}_i$ and $\mathbf{x}_j$ only if $\mathbf{x}_i$ is one of the K-nearest neighbors of $\mathbf{x}_j$, and, $\mathbf{x}_j$ is also one of the K-nearest neighbors of $\mathbf{x}_i$. After the neighborhood graph is constructed, the shortest paths of all pairs of data points are then computed using Floyd's algorithm or Dijkstra's algorithm [6] in order to approximate the geodesic distances. Since finding the optimal solution of (1) is NP-hard [24], in our algorithm, we fix the value of C in advance. Then, a greedy procedure [13] is utilized to select C data samples $\mathbf{s}_1, \ldots, \mathbf{s}_C$ from $\mathcal{X}$ as the representative points of subsets $\mathcal{X}_1, \ldots, \mathcal{X}_C$, respectively. Finally, a data point $\mathbf{x}_i$ is assigned to the subset $\mathcal{X}_c$ if $\text{dist}_G(\mathbf{x}_i, \mathbf{s}_c) = \min_{c'=1,\ldots,C} \text{dist}_G(\mathbf{x}_i, \mathbf{s}_{c'})$. The detailed procedure of GC is described in Algorithm 1, where $\mathcal{N}_i$ denotes the K-nearest neighborhood set of data point $\mathbf{x}_i$ and $\text{subset}(\mathbf{x}_i) = c$ indicates that $\mathbf{x}_i \in \mathcal{X}_c$.

Algorithm 1 Geodesic Clustering (GC)

Input: Training dataset: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$; number of subsets: C
Output: Partitioned subsets: $\mathcal{X}_1, \ldots, \mathcal{X}_C$; representative points: $\mathbf{s}_1, \ldots, \mathbf{s}_C$

```
for all pairs (x_i, x_j) do
    if x_i ∈ N_j and x_j ∈ N_i then
        dist_0(x_i, x_j) = ||x_i − x_j||^2;
    else
        dist_0(x_i, x_j) = +∞;
for all pairs (x_i, x_j) do
    Compute the shortest paths dist_G(x_i, x_j);
Randomly select x_r from X, let s_1 = x_r, S = {s_1};
for i = 1, ..., n do
    dist(x_i) = dist_G(x_i, s_1); subset(x_i) = 1;
for c = 2, ..., C do
    x_r = arg max_{x_r : x_r ∈ X\S} dist(x_r);
    s_c = x_r, S = S ∪ {s_c};
    for i = 1, ..., n do
        if dist_G(x_i, s_c) ≤ dist(x_i) then
            dist(x_i) = dist_G(x_i, s_c); subset(x_i) = c;
for c = 1, ..., C do
    X_c = {x_i : subset(x_i) = c};
```

B. Locally Conjugate Discriminant Projection

For each subset $\mathcal{X}_c$, we define the data matrix $\mathbf{X}_c = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_c}]$, where n_c is the number of data points in $\mathcal{X}_c$. The aim of LCDP is to learn a $D \times d$ transformation matrix $\mathbf{W}_c = [\mathbf{w}_{c1}, \ldots, \mathbf{w}_{cd}]$, which maximizes the discriminant information, whereas preserving the local geometry in the reduced subspace of $\mathcal{X}_c$.

In order to describe the discriminant information of the subset, we define the following discriminant scatter matrix:

$$\begin{aligned}\mathbf{S}_c^d &= \sum_{i,j} (\mathbf{A}_c^d)_{ij} (\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)^T \\ &= \sum_{i,j} (\mathbf{A}_c^d)_{ij} \mathbf{W}_c^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}_c\end{aligned} \tag{2}$$

where $\mathbf{A}_c^d = [(\mathbf{A}_c^d)_{ij}]$ is the $n_c \times n_c$ discriminant coefficient matrix and $\mathbf{y}_i \in \mathbb{R}^d$ is the low-dimensional representation of $\mathbf{x}_i$ projected by the transformation matrix $\mathbf{W}_c$. Similarly, we define the locality scatter matrix to emphasize the details of local geometry

$$\mathbf{S}_c^l = \sum_{i,j} (\mathbf{A}_c^l)_{ij} \mathbf{W}_c^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}_c \tag{3}$$

where $\mathbf{A}_c^l = [(\mathbf{A}_c^l)_{ij}]$ is the $n_c \times n_c$ locality coefficient matrix. Actually, $\mathbf{S}_c^d$ and $\mathbf{S}_c^l$ can be rewritten as follows:

$$\mathbf{S}_c^d = \mathbf{W}_c^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{W}_c, \quad \mathbf{S}_c^l = \mathbf{W}_c^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{W}_c \tag{4}$$

where $\mathbf{L}_c^d = \mathbf{D}_c^d - \mathbf{A}_c^d$ and $\mathbf{L}_c^l = \mathbf{D}_c^l - \mathbf{A}_c^l$ are two Laplacian matrices [1], and $\mathbf{D}_c^d$ and $\mathbf{D}_c^l$ are diagonal matrices defined as $(\mathbf{D}_c^d)_{ii} = \sum_{j=1}^{n_c}(\mathbf{A}_c^d)_{ij}$ and $(\mathbf{D}_c^l)_{ii} = \sum_{j=1}^{n_c}(\mathbf{A}_c^l)_{ij}$ $(i = 1, \ldots, n_c)$, respectively.

Accordingly, the following objective function of the LCDP for $\mathcal{X}_c$ is formulated to maximize the discriminant information and preserve the manifold structure simultaneously:

$$\mathbf{W}_c = \arg\max_{\mathbf{W}_c} \left\{ \mathrm{tr}\left(\left(\mathbf{S}_c^l\right)^{\dagger} \mathbf{S}_c^d \right) \right\} \tag{5}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operation and $(\mathbf{S}_c^l)^{\dagger}$ denotes the Moore–Penrose pseudoinverse of $\mathbf{S}_c^l$.

In order to eliminate the correlation between different projection directions, we further introduce the conjugate constraint

$$\mathbf{W}_c^T \mathbf{B}_c \mathbf{W}_c = \mathbf{I}_d \tag{6}$$

which indicates that different projection directions are conjugate with respect to $\mathbf{B}_c$, i.e., $\mathbf{w}_{ci}^T \mathbf{B}_c \mathbf{w}_{cj} = 0$ when $i \neq j$.

The above constraint actually could be viewed as the generalization of the orthogonal constraint and uncorrelated constraint. In particular, if we set $\mathbf{B}_c = \mathbf{I}_d$, the proposed constraint is reduced to $\mathbf{W}_c^T \mathbf{W}_c = \mathbf{I}_d$, which is called the orthogonal constraint in many DR algorithms such as orthogonal isometric projection (IsoP) [40] and supervised optimal locality preserving projection (SOLPP) [35]. When we set $\mathbf{B}_c = \sum_{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$, the proposed constraint is reduced to the uncorrelated constraint in many DR algorithms such as uncorrelated linear discriminant analysis (LDA) [20] and uncorrelated discriminant IsoP [11].

To achieve good discriminant performance, it is crucial to separate close data points of different classes and keep them far away from each other after embedding. So, we specify the discriminant coefficient matrix $\mathbf{A}_c^d$ as follows:

$$\left(\mathbf{A}_c^d\right)_{ij} = \begin{cases} e^{\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma}}, & \text{if } j \in \mathcal{N}_i,\ i \in \mathcal{N}_j,\ \text{and } l_i \neq l_j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

in which l_i denotes the class label of $\mathbf{x}_i$ and σ is empirically set by $\sigma = \sum_{i=1}^{n_c} ||\mathbf{x}_i - \mathbf{x}_{i_K}||^2 / n_c$, where $\mathbf{x}_{i_K}$ is the Kth nearest neighbor of $\mathbf{x}_i$. Similarly, we specify $\mathbf{A}_c^l$

$$\left(\mathbf{A}_c^l\right)_{ij} = \begin{cases} e^{\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma}}, & \text{if } j \in \mathcal{N}_i,\ i \in \mathcal{N}_j,\ \text{and } l_i = l_j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

unlike the existing orthogonal constraint or uncorrelated constraint that treats all of the data points equally, the conjugate constraint matrix $\mathbf{B}_c$ in LCDP emphasizes the relationship between nearby data points

$$\begin{aligned} \mathbf{B}_c &= \sum_{i,j} \left(\left(\mathbf{A}_c^d\right)_{ij} + \left(\mathbf{A}_c^l\right)_{ij} \right) (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \\ &= \mathbf{X}_c \left(\mathbf{L}_c^d + \mathbf{L}_c^l \right) \mathbf{X}_c^T. \end{aligned} \tag{9}$$

Note that under the proposed LCDP strategy, three matrices $\mathbf{A}_c^d$, $\mathbf{A}_c^l$, and $\mathbf{B}_c$ could be modified flexibly according to various learning objectives.

Let $\mathbf{B}_c = \mathbf{U}\Sigma\mathbf{U}^T$ be the eigendecomposition of $\mathbf{B}_c$, where $\mathbf{U}$ is orthogonal, $\Sigma = \begin{bmatrix} \Sigma_\gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, $\Sigma_\gamma \in \mathbb{R}^{\gamma \times \gamma}$ is diagonal, and $\gamma = \mathrm{rank}(\mathbf{B}_c)$. Then, let $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$, where $\mathbf{U}_1 \in \mathbb{R}^{D \times \gamma}$ and $\mathbf{U}_2 \in \mathbb{R}^{D \times (D-\gamma)}$, we have

$$\begin{aligned} \begin{bmatrix} \Sigma_\gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} &= \mathbf{U}^T \mathbf{B}_c \mathbf{U} \\ &= \begin{bmatrix} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 & \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_2 \\ \mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 & \mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_2 \end{bmatrix} \\ &\quad + \begin{bmatrix} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_1 & \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_2 \\ \mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_1 & \mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_2 \end{bmatrix}. \end{aligned} \tag{10}$$

Obviously, $\mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_2 + \mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_2 = \mathbf{0}$. So, we know that $\mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_2 = \mathbf{0}$ and $\mathbf{U}_2^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_2 = \mathbf{0}$ since both of them are positive semi-definite. Then, we know that $\mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_2 = \mathbf{0}$ and $\mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_2 = \mathbf{0}$ as both matrices at the end of (10) are positive semidefinite. We also have $\mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 + \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_1 = \Sigma_\gamma$, then

$$\Sigma_\gamma^{-\frac{1}{2}} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 \Sigma_\gamma^{-\frac{1}{2}} + \Sigma_\gamma^{-\frac{1}{2}} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{U}_1 \Sigma_\gamma^{-\frac{1}{2}} = \mathbf{I}_\gamma. \tag{11}$$

Let $\mathbf{G}_c = \mathbf{P}\Lambda\mathbf{P}^T$ be the eigendecomposition of $\mathbf{G}_c$, where $\mathbf{G}_c = \Sigma_\gamma^{-1/2} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 \Sigma_\gamma^{-1/2}$ and $\mathbf{P}$ is orthogonal. Define the transformation matrix $\mathbf{W}_c = \mathbf{U}_1 \Sigma_\gamma^{-1/2} \mathbf{P}_d$, where $\mathbf{P}_d = \mathbf{P} \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix}$ denotes the first d columns of $\mathbf{P}$ corresponding to the d largest eigenvalues of $\mathbf{G}_c$, then we have

$$\mathbf{S}_c^d = \mathbf{W}_c^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{W}_c = \Lambda_d \tag{12}$$

$$\mathbf{S}_c^l = \mathbf{W}_c^T \mathbf{X}_c \mathbf{L}_c^l \mathbf{X}_c^T \mathbf{W}_c = \mathbf{I}_d - \Lambda_d \tag{13}$$

$$\mathbf{W}_c^T \mathbf{B}_c \mathbf{W}_c = \mathbf{I}_d. \tag{14}$$

Above analysis indicates that $\mathbf{W}_c$ is the solution of the optimization problem in (5) with the constraint in (6). The detailed procedure of the LCDP is described in Algorithm 2.

Algorithm 2 Locally Conjugate Discriminant Projection (LCDP)

Input: Partitioned subsets: $\mathcal{X}_1, \ldots, \mathcal{X}_C$; the low dimension: d
Output: Transformation matrices: $\mathbf{W}_1, \ldots, \mathbf{W}_C \in \mathbb{R}^{D \times d}$

for $c = 1, \ldots, C$ **do**
- Construct data matrix $\mathbf{X}_c = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_c}]$ of $\mathcal{X}_c$;
- Construct $\mathbf{A}_c^d$, $\mathbf{A}_c^l$, and $\mathbf{B}_c$ using (7), (8), and (9);
- Obtain $\mathbf{U}$ by eigendecomposition: $\mathbf{B}_c = \mathbf{U}\Sigma\mathbf{U}^T$;
- $\gamma = rank(\mathbf{B}_c)$, $\mathbf{U}_1 = \mathbf{U} \begin{bmatrix} \mathbf{I}_\gamma \\ \mathbf{0} \end{bmatrix}$, $\Sigma_\gamma = [\mathbf{I}_\gamma, \mathbf{0}] \Sigma \begin{bmatrix} \mathbf{I}_\gamma \\ \mathbf{0} \end{bmatrix}$;
- $\mathbf{G}_c = \Sigma_\gamma^{-\frac{1}{2}} \mathbf{U}_1^T \mathbf{X}_c \mathbf{L}_c^d \mathbf{X}_c^T \mathbf{U}_1 \Sigma_\gamma^{-\frac{1}{2}}$;
- Obtain $\mathbf{P}$ by eigendecomposition: $\mathbf{G}_c = \mathbf{P}\Lambda\mathbf{P}^T$;
- $\mathbf{P}_d = \mathbf{P} \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix}$;
- $\mathbf{W}_c = \mathbf{U}_1 \Sigma_\gamma^{-\frac{1}{2}} \mathbf{P}_d$;

C. Unified Mapping Function

For a new data point $\mathbf{x}$, we first identify the subset that it belongs to, and then project it to the low-dimensional subspace using the corresponding transformation matrix. These two steps could be unified by the following mapping function:

$$\mathbf{y} = \sum_{c=1}^{C} \mathbb{1}\{\mathrm{dist}_G(\mathbf{x}, \mathbf{s}_c) = \min_{c'=1,\ldots,C} \mathrm{dist}_G(\mathbf{x}, \mathbf{s}_{c'})\} \cdot \mathbf{W}_c^T \mathbf{x} \tag{15}$$

where $\mathbf{y}$ is the low-dimensional representation of $\mathbf{x}$, $\mathbb{1}\{\cdot\}$ is the indicator function that equals one if the argument holds and zero otherwise, $\mathbf{s}_c$ is the representative point of the subset $\mathcal{X}_c$ selected by

Algorithm 3 Hybrid Manifold Embedding (HyME)

Input: Training dataset: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$; number of subsets: C; the low dimension: d; new data point: $\mathbf{x}$

Output: Low-dimensional embedding of $\mathbf{x}$: $\mathbf{y}$

1 $\{\mathcal{X}_1, \ldots, \mathcal{X}_C; \mathbf{s}_1, \ldots, \mathbf{s}_C\} = \text{GC}(\mathcal{X}, C)$;

2 $\{\mathbf{W}_1, \ldots, \mathbf{W}_C\} = \text{LCDP}(\mathcal{X}_1, \ldots, \mathcal{X}_C; d)$;

3 $\mathbf{y} = \sum_{c=1}^{C} \mathbb{1}\{\text{dist}_G(\mathbf{x}, \mathbf{s}_c) = \min_{c'=1,\ldots,C} \text{dist}_G(\mathbf{x}, \mathbf{s}_{c'})\} \cdot \mathbf{W}_c^T \mathbf{x}$;

TABLE I
TRAINING AND TEST TIME COSTS OF HYME

	Training time cost	Test time cost
HyME	$O(KDn + n^2 + Dn^2/C)$	$O(D(C + d))$

TABLE II
PERFORMANCE OF LINEAR PROJECTION [FIG. 1(a)], GC + LCDP [FIG. 1(b)], AND k-MEANS + LCDP [FIG. 1(c)] ON SYNTHETIC DATA SET

	LP	k-means+LCDP	GC+LCDP
Separable/Total	22/39	34/39	39/39

the GC strategy, and $\mathbf{W}_c$ is the transformation matrix of $\mathcal{X}_c$ learned by the LCDP scheme.

The entire procedure of HyME, including GC, LCDP, and the mapping process of new data point, is described in Algorithm 3. The training and test time costs of HyME are listed in Table I. For ease of analysis, we simply assume that all C subsets are of the same size, i.e., $n_1 = \cdots = n_C = n/C$.

III. EXPERIMENTS

In this section, we demonstrate the performance of HyME on a simple synthetic example as well as three standard data sets: the United State Postal Service (USPS) digit data set [19], the CMU pose, illumination, and expression (PIE) face data set [29], and the UMIST face data set [14]. In our experiments, the K-nearest-neighbor classifier is utilized in the learned subspace for final classification and we set $K = 5$.

A. Synthetic Example

First, we use the synthetic example given in Fig. 1 to show the effect of the proposed GC technique. The synthetic data set is 2-D and composed of three classes. There are 8, 21, and 10 data points in class 1, 2, and 3, respectively.

Table II reports the ratio of the number of separable data points to the total number of data points of linear projection [Fig. 1(a)], GC + LCDP [Fig. 1(b)], and k-means + LCDP [Fig. 1(c)], respectively, where separable data points refer to the data points that do not fall into the range of other classes. Since the original data set is linearly nonseparable, the linear projection performs the worst: more than one third of the projected data points fall into the range of other classes. By dividing the data set into two subsets, k-means + LCDP obtains better performance: the subset 2 is linearly separable and the number of nonseparable data points reduces to 5. However, subset 1 is still linearly nonseparable after partition, which limits the performance of the following discriminant projection. Finally, GC + LCDP (i.e., HyME) effectively simplifies the structure of the data set by minimizing the nonlinearity of each subset, and thus makes all the projected data points separable.

B. USPS Digit Data Set

The USPS data set of hand written digital characters contains 11 000 normalized grayscale images of size 16 × 16, with 1100 images in each class [19].

In order to illustrate the discriminant ability of HyME, we conduct an visualization experiment. We map 1000 images (100 for each class) onto the 2-D subspaces learned by the following DR algorithms: 1) principal component analysis (PCA) [17]; 2) LDA [10]; 3) LPP [16]; 4) locality-preserved maximum information projection (LPMIP) [34]; 5) SOLPP [35]; and 6) HyME. For HyME, we demo the results when $C = 1$ and $C = 2$, respectively.

From Fig. 2(a)–(f), we can see that the projection results of LDA, LPMIP, SOLPP, and HyME have better separability than the results of PCA and LPP. Actually, HyME is reduced to a linear manifold learning algorithm when $C = 1$. So, it is reasonable that its performance is similar to the LDA and LPMIP. When $C = 2$, which means that HyME first divides the data set into two subsets and then learns the subspace of each subset individually, we can observe from Fig. 2(g) that the overlap between different classes is largely reduced.

In the second experiment, we evaluate the performance of HyME by comparing it with nine competitive DR algorithms: 1) LDA; 2) local LDA (LLDA) [7]; 3) LPP; 4) LPMIP; 5) SOLPP; 6) neighborhood preserving embedding [15]; 7) local discriminant embedding [4]; 8) IsoP [3]; and 9) uncorrelated discriminant IsoP [11] on the entire USPS data set. In each class, 10 images are selected for training and the rest are used for test. We repeat the experiment 10 times on different randomly selected training sets and report the average results. As shown in Table III, HyME achieves higher recognition accuracy than the other algorithms.

In order to show the performance of the proposed method on Subset 1 and Subset 2 ($C = 2$) separately, we provide more details of all the 10 trials in Table IV. Here, we fix dim = 8. Although HyME shows good overall performance in Table III, its accuracy might not be so satisfactory when the data are biased to some subsets or classes. For example, in trial 2, the majority of the data set is in Subset 1. Moreover, within each subset, the data points are biased to some classes as well. In this situation, we can observe from the table that the accuracy on Subset 2 of trial 2 is low. One possible solution to this problem is constructing more balanced subsets/classes in the clustering process.

To further demonstrate the effect of GC and LCDP, we compare HyME with k-means + LCDP and GC + LDA. By comparing HyME (i.e., GC + LCDP) with k-means + LCDP, we can see the effect of GC directly. By comparing HyME with GC + LDA, the effectiveness of the LCDP is highlighted independently. The experiment settings are the same as those in the the previous experiment. In this experiment, we set $C = 2$. Table V lists the recognition results of these three algorithms. By combining GC and LCDP, HyME performs better than the algorithm that employs only GC (GC + LDA) or LCDP (k-means + LCDP).

As mentioned in Section II-B, the parameter σ in HyME is an empirical setting. In order to examine the reasonableness of this setting, we test the performance of HyME under different values of σ ($\sigma = 0.1\sigma_0, 0.2\sigma_0, 0.5\sigma_0, \sigma_0, 2\sigma_0, 5\sigma_0, 10\sigma_0$, where σ_0 denotes the empirical setting). We fix $C = 2$ and Dim = 8. The recognition accuracies and standard derivations are reported in Table VI.

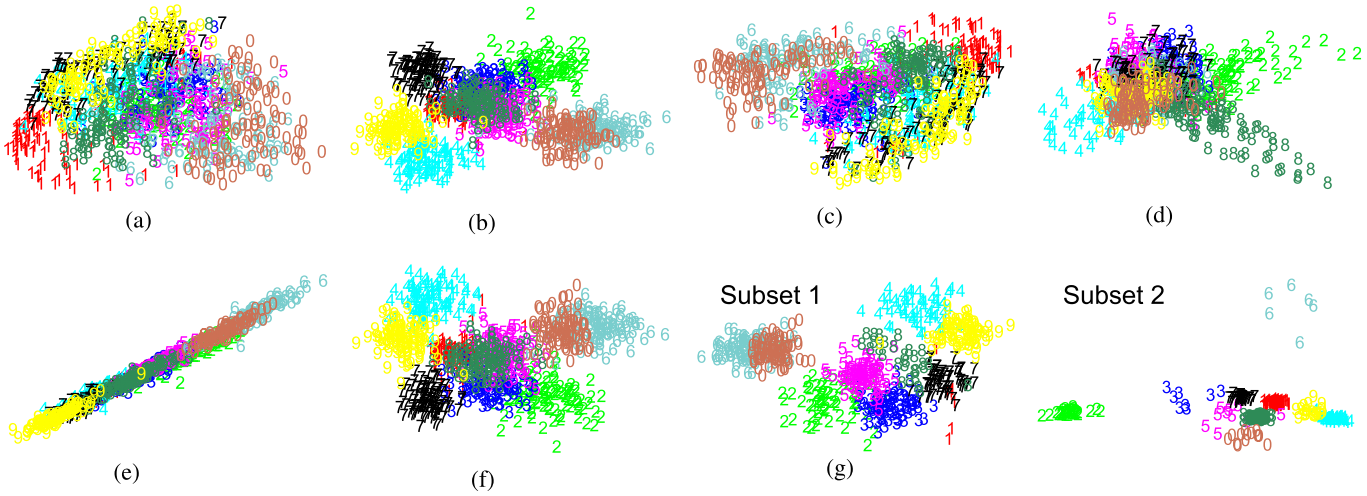

Fig. 2. Low-dimensional 2-D representations of USPS handwritten digits (0–9) generated by various DR algorithms. (a) PCA. (b) LDA. (c) LPP. (d) LPMIP. (e) SOLPP. (f) HyME with $C = 1$. (g) HyME with $C = 2$.

TABLE III
PERFORMANCE OF VARIOUS DR ALGORITHMS ON USPS AND CMU PIE DATA SETS

		LDA	LLDA	LPP	SOLPP	NPE	LDE	LPMIP	IsoP	UDIsoP	HyME ($C = 1$)	HyME ($C = 2$)	HyME ($C = 3$)
USPS	Accu.(%)	56.9	54.1	50.0	57.9	50.9	54.8	56.3	48.1	56.9	56.0	**59.8**	59.1
	Std.(%)	4.0	1.0	3.8	3.8	3.2	3.4	3.1	3.1	3.0	4.1	3.2	3.3
	Dim.	9	6	20	18	27	10	9	16	15	10	8	9
PIE	Accu.(%)	63.8	60.4	66.5	66.7	66.1	64.1	62.9	66.4	66.2	64.5	**68.0**	59.3
	Std.(%)	2.3	1.1	2.7	1.2	2.7	3.1	3.3	2.9	2.6	2.7	3.2	4.3
	Dim.	19	10	19	30	19	20	17	20	26	19	19	19

TABLE IV
PERFORMANCE OF HyME ON SUBSET 1 AND SUBSET 2 OF 10 TRIALS ON USPS AND CMU PIE DATA SETS

			1	2	3	4	5	6	7	8	9	10
USPS	Subset 1	Accu.(%)	61.5	60.4	67.2	60.4	65.3	55.3	59.6	53.2	63.3	55.6
		Number of Classes	10	10	9	10	9	10	8	10	10	10
		Number of Training Data	70	80	64	68	52	59	38	72	69	63
	Subset 2	Accu.(%)	56.8	40.6	61.4	58.8	60.8	67.0	52.8	59.5	59.7	68.9
		Number of Classes	8	7	8	9	7	9	9	7	9	8
		Number of Training Data	30	20	36	32	48	41	62	28	31	37
PIE	Subset 1	Accu.(%)	72.9	78.2	73.2	71.1	68.6	68.2	66.2	81.8	80.2	69.2
		Number of Classes	20	18	20	20	20	20	20	20	19	20
		Number of Training Data	107	81	115	161	110	133	135	128	103	131
	Subset 2	Accu.(%)	65.2	66.8	62.9	33.3	62.1	59.6	56.5	48.8	68.0	68.7
		Number of Classes	20	20	19	14	20	18	17	19	20	18
		Number of Training Data	93	119	85	39	90	67	65	72	97	69

From the results, we can see that the performance of HyME is relatively robust to the variation of σ around the empirical setting. Therefore, we keep using the empirical setting in the remaining experiments.

C. CMU PIE Face Data Set

The CMU PIE data set contains 41 368 images of 68 people with different poses, illumination conditions, and facial expressions [29]. Currently, 11 554 images in PIE are downloadable [16], which are grayscale and normalized to a resolution of 32×32 pixels. In our experiments, we use a subset, which contains 3400 images of the first 20 people (170 for each).

In the first experiment, we compare the aforementioned 10 DR algorithms. For each class, 10 images are randomly selected for training and the rest are used for test. The results are listed in Table III. By taking both the global and local manifold structures as well as the discriminant information into account in a unified

TABLE V
PERFORMANCE OF k-MEANS + LCDP, GC + LDA, AND HYME ON USPS DATA SET

	k-means+LCDP	GC+LDA	HyME
Accu.(%)	58.7	59.1	**59.8**
Std.(%)	3.4	3.0	3.2
Dim.	7	8	8

TABLE VI
PERFORMANCE OF HYME ON USPS DATA SET UNDER DIFFERENT SETTINGS OF σ

	$0.1\sigma_0$	$0.2\sigma_0$	$0.5\sigma_0$	σ_0	$2\sigma_0$	$5\sigma_0$	$10\sigma_0$
Accu.(%)	60.5	60.1	60.4	59.8	59.2	60.6	60.9
Std.(%)	4.3	3.5	3.8	3.2	3.9	3.9	4.3

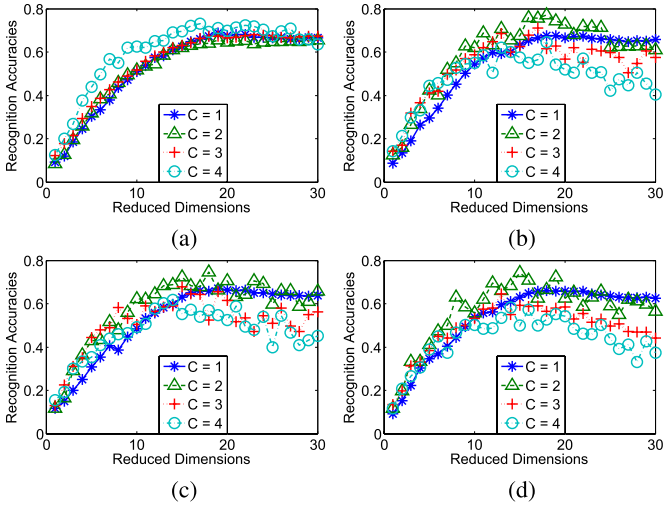

Fig. 3. Recognition accuracies of HyME on CMU PIE data set under different settings of neighborhood size and subset number. (a) $K = 2$. (b) $K = 3$. (c) $K = 4$. (d) $K = 5$.

nonlinear framework, HyME outperforms the other algorithms in terms of accuracy.

Similar to the experiment on USPS data set, in Table IV, we provide details of the performance of proposed method in all the 10 trials on CMU PIE data set. In the result, we can observe a trend: less training data points in a subset generally causes the lower recognition accuracy. The worst performance appears in Subset 2 of trial 4, which has only 39 data points for 14 classes. This observation indicates that the performance of the proposed algorithm might be affected by the proportion of data points in each subset. Therefore, giving a good partition of the data set could improve the accuracy directly.

Another observation from Table IV is that about half of the subsets do not contain full number of classes, which indicates that the GC can differentiate classes in some situations. However, this preclassification does not affect the final accuracy of the proposed algorithm too much. The reason might be that the classes differentiated by clustering are already well separated in the original feature space, and some closely overlapped classes are still staying

TABLE VII
PERFORMANCE OF HYME ON UMIST DATA SET UNDER DIFFERENT SETTINGS OF SUBSET NUMBER AND TRAINING DATA NUMBER

	C	1	2	3	4	5
$p = 5$	Accu.(%)	51.3	50.3	**56.2**	55.0	55.0
	Std.(%)	4.9	7.3	5.3	5.2	4.8
$p = 10$	Accu.(%)	62.9	73.9	78.2	**79.1**	78.6
	Std.(%)	4.0	8.3	3.7	4.8	5.2
$p = 15$	Accu.(%)	65.4	79.3	87.2	89.9	**90.3**
	Std.(%)	3.6	6.6	5.6	4.5	3.1

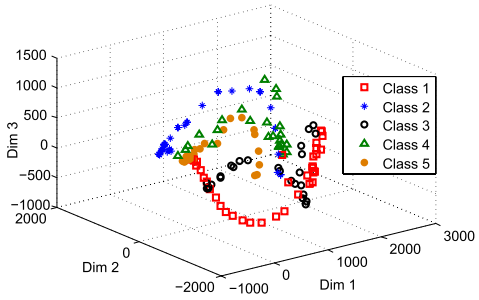

Fig. 4. 3-D PCA representations of data points of the first five classes from UMIST data set. Clear manifold structure could be observed.

in the same subset. The proposed GC could benefit the performance of latter classification since it simplifies the structure of the data set by minimizing the nonlinearity, no matter whether it differentiates classes or not.

Furthermore, we examine the performance of HyME under different settings of neighborhood sizes ($K = 2, 3, 4, 5$) and subset numbers ($C = 1, 2, 3, 4$). For each class, 10 images are randomly selected for training. Fig. 3 shows the accuracies of HyME versus reduced dimensions under different settings of K and C. As can be seen, HyME performs similarly when $K = 3, 4, 5$, which indicates that the proposed method is not so sensitive to K in this range. By observing Fig. 3(b)–(d), we find that the accuracies of HyME start to decrease when C is larger than 2. One possible reason of this tendency is that although the GC partition can simplify the globally complicated structure of the data set, over-partition of the small training set may cause the insufficiency of training data in each partitioned subset, and thus affect the performance.

D. UMIST Face Data Set

In order to further evaluate the effect of C on the proposed algorithm, we conduct an experiment on the UMIST face data set [14]. This data set consists of 575 images of 20 individuals. Each individual is shown in a range of poses from profile to frontal views. Each image is grayscale and downsampled to the resolution of 56×46. For each class, p ($p = 5, 10, 15$) images are randomly selected for training and the rest are used for test. We fix $K = 5$ and Dim $= 5$.

Table VII gives the recognition accuracies of HyME on UMIST with different values of p and C. We can see that the optimal value of C on this data set is larger than that on USPS and CMU PIE. Moreover, the optimal C increases as p increases. One reasonable explanation of above observations is that the UMIST data set contains obvious manifold structure (as shown in Fig. 4) so that a small C might not be able to make the subsets simple

enough. When p increases, the distribution becomes more complex, which calls HyME to give a larger C to simplify the structure of the data set.

IV. Conclusion

In this brief, we present a supervised manifold learning framework dubbed HyME, which provides a nonlinear explicit mapping function by performing a two-layer learning procedure. By introducing a nonlinear clustering algorithm and a LCDP strategy, HyME discovers both the global and local manifold structure of the data set, as well as captures the discriminant information for classification. Experiments demonstrate good performance of the proposed HyME.

Future work will be explored from the following several aspects. First, the proposed framework is a combination of two algorithms so that it is not as compact as a single learning model. How to design a more compact framework is the first topic that we are going to investigate. Moreover, how to automatically determine the optimal number of subset as well as a good partition of the data set is also a meaningful future work, which could directly benefit the final recognition performance. Last but not least, we aim to extend the proposed framework for large-scale problems, which are very common in real-world applications.

Acknowledgment

The authors would like to thank the editors and reviewers for their constructive comments and suggestions.

References

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. NIPS*, 2001, pp. 585–591.

[2] C. Bregler and S. Omohundro, "Nonlinear manifold learning for visual speech recognition," in *Proc. 5th ICCV*, 1995, pp. 494–499.

[3] D. Cai, X. He, and J. Han, "Isometric projection," in *Proc. AAAI*, 2007, pp. 528–533.

[4] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 846–853.

[5] J. Chen, Z. Ma, and Y. Liu, "Local coordinates alignment with global preservation for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 106–117, Jan. 2013.

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.

[7] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.

[8] H. R. Fang and Y. Saad, "Enhanced multilevel manifold learning," Dept. Comput. Sci. Eng., Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. ys-2012-2, 2012.

[9] H. R. Fang, S. Sakellaridi, and Y. Saad, "Multilevel manifold learning with application to spectral clustering," in *Proc. 19th ACM CIKM*, 2010, pp. 419–428.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[11] B. Ge, Y. Shao, and Y. Shu, "Uncorrelated discriminant isometric projection for face recognition," in *Proc. Inf. Comput. Appl.*, vol. 307. 2012, pp. 138–145.

[12] X. Geng, D.-C. Zhan, and Z.-H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 6, pp. 1098–1107, Dec. 2005.

[13] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Comput. Sci.*, vol. 38, pp. 293–306, 1985.

[14] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Proc. Face Recognit., From Theory Appl., NATO ASI Ser. F, Comput. Syst. Sci.*, vol. 163. 1998, pp. 446–456.

[15] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE ICCV*, Oct. 2005, pp. 1208–1213.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[17] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[18] Y. Hou, P. Zhang, X. Xu, X. Zhang, and W. Li, "Nonlinear dimensionality reduction by locally linear inlaying," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 300–315, Feb. 2009.

[19] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[20] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.

[21] T.-K. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 318–327, Mar. 2005.

[22] Y. Liu, Y. Liu, and K. C. C. Chan, "Ordinal regression via manifold learning," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 398–403.

[23] D. Luo, F. Nie, C. Ding, and H. Huang, "Multi-subspace representation and discovery," in *Proc. ECML/PKDD*, 2011, pp. 405–420.

[24] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard," *Theoretical Comput. Sci.*, vol. 442, pp. 13–21, Jul. 2012.

[25] F. Nie, S. Xiang, and C. Zhang, "Neighborhood minmax projections," in *Proc. 20th IJCAI*, 2007, pp. 993–998.

[26] R. Pless, "Image spaces and video trajectories: Using isomap to explore video sequences," in *Proc. 9th IEEE ICCV*, Oct. 2003, pp. 1433–1440.

[27] A. Rahimi, B. Recht, and T. Darrell, "Learning appearance manifolds from video," in *Proc. CVPR*, 2005, pp. 868–875.

[28] S. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[29] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 1615–1618, Dec. 2003.

[30] M. Sun, J. Yang, C. Liu, and J. Yang, "Similarity preserving principal curve: An optimal 1-D feature extractor for data representation," *IEEE Trans. Neural Netw.*, vol. 21, no. 9, pp. 1445–1456, Sep. 2010.

[31] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[32] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-linear dimensionality reduction techniques for classification and visualization," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 645–651.

[33] H. Wang, H. Huang, and C. H. Q. Ding, "Discriminant Laplacian embedding," in *Proc. 24th AAAI*, 2010, pp. 618–623.

[34] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 571–585, Apr. 2008.

[35] W. Wong and H. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognit.*, vol. 45, no. 1, pp. 186–197, 2012.

[36] S. Yan, Y. Hu, D. Xu, H.-J. Zhang, B. Zhang, and Q. Cheng, "Nonlinear discriminant analysis on embedded manifold," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 468–477, Apr. 2007.

[37] J. Zhang, X. Wang, U. Krüger, and F.-Y. Wang, "Principal curve algorithms for partitioning high-dimensional data spaces," *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 367–380, Mar. 2011.

[38] L. Zhang, L. Qiao, and S. Chen, "Graph-optimized locality preserving projections," *Pattern Recognit.*, vol. 43, no. 6, pp. 1993–2002, 2010.

[39] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[40] Y. Zheng, B. Fang, Y. Y. Tang, T. Zhang, and R. Liu, "Learning orthogonal projections for isomap," *Neurocomputing*, vol. 103, pp. 149–154, Mar. 2013.

[41] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 244–257, Jan. 2013.

[42] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: A unified framework for sparse dimension reduction," *Data Mining Knowl. Discovery*, vol. 22, no. 3, pp. 340–371, 2011.