

Received January 21, 2020, accepted February 10, 2020, date of publication March 13, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980757

# Supervised Structural Sparse Subspace Learning Based on Hierarchical Locality Preservation for Multimodal and Mixmodal Data

QI ZHANG<sup>1,2</sup>, CHAOYI SHI<sup>3</sup>, AND TIANGUANG CHU<sup>3</sup>

<sup>1</sup>School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China

<sup>2</sup>Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China

<sup>3</sup>College of Engineering, Peking University, Beijing 100871, China

Corresponding author: Tianguang Chu (chutg@pku.edu.cn)


This work was supported in part by the Beijing Natural Science Foundation under Grant 19L2037, in part by the NSFC under Grant 61673027, in part by the Humanity and Social Science Youth Foundation of Ministry of Education under Grant 20YJCZH228, in part by the Excellent Young Scholars Funding Project in UIBE under Grant 19YQ10, in part by the Fundamental Research Funds for the Central Universities in UIBE under Grant CXTD10-05 and Grant 18QD18, and in part by the Foundation for Disciplinary Development of School of Information Technology and Management, University of International Business and Economics.

**ABSTRACT** We study the multimodal and mixmodal data-driven supervised structural sparse subspace learning problem in this paper, and present the  $\alpha$ -structural regularization based hierarchical locality analysis ( $\alpha$ -SRHLA) model. Unlike most existing sparse subspace learning models that merely constrain the cardinalities of the subspace basis vectors, the present  $\alpha$ -SRHLA model takes into account the structural correlations of the original data variables and generates “variable groups” for sparse subspace learning. As a result, the sparsity is induced in the scale of the variable group instead of the single variable, i.e., “structural sparsity”. In addition, the  $\alpha$ -SRHLA considers the “hierarchical locality” of multimodal and mixmodal data, and derives the weighted local affinity correlations in both data-level and class-level. This helps to reveal the intrinsic distribution characteristics of the considered multimodal and mixmodal manifold structures. A series of experiments on normal and multimodal data classification, multimodal and mixmodal digit as well as face recognition verify the effectiveness of the present  $\alpha$ -SRHLA model in dealing with both multimodal and mixmodal data.

**INDEX TERMS** Dimensionality reduction, sparse subspace learning, structural sparsity, multimodal and mixmodal data, face recognition.

## I. INTRODUCTION

Supervised learning, as an efficient way that incorporates the information of label supervision with data distribution, can decrease the temporal and computational burden of the learning processes. In the past few years, numbers of supervised learning models have been presented such as [1]–[7], but most existing models might not always obtain satisfying results when dealing with multimodal (i.e., *samples within the same class have separated clusters*) and mixmodal (i.e., *some samples from different classes have relatively closer distances than those from the same class*) data, as the hierarchical distribution properties (i.e., *different distribution properties are shown in within-class and between-class scales*) exhibited in

The associate editor coordinating the review of this manuscript and approving it for publication was Meng-Lin Ku .

multimodal or mixmodal data are sometimes neglected in the context of feature extraction.

In real applications, problems involving data with both multimodality and mixmodality are frequently encountered. For example, fulfilling multi-class data classification by utilizing a series of “one-versus-rest” binary data classification tasks induces within-class multimodality. Besides, separating odd and even numbers from 0-9 upon the handwritten digit images follows between-class multimodal properties as images of different odd/even numbers have the same labels. For the between-class mixmodality, when the face images have noises (e.g., *different illuminations or decorations*), data samples from different classes might be located more adjacently in the Euclidean space, whereas those of the same class might be separated with each other [8]. Moreover, the handwriting recognition task also follows mixmodal rules as the

same handwriting contents written by different people are with different class labels, but in the original space, these data samples share comparatively close Euclidean distances.

To tackle the aforementioned multimodal and mixmodal data-driven problems, a locality preserving discriminant analysis (LPDA) model was proposed in [9], which is concerned with the hierarchical locality (*i.e.*, the different local geometric information in within-class and between-class scales) of data with complex distributions, and efficiently derives the inherited characteristics of both multimodal and mixmodal data. The LPDA model is designed for nonsparse scenarios, where all data variables are treated without discrimination, and the low-dimensional projections are correlated with all the original data variables. However in practice, it is also desirable to find a few variables in a more interpretable way during the learning process, especially for the cases where the variables have physical meanings, e.g., gene analysis, face recognition, etc [10].

An efficient way to achieve this is to impose sparsity regularization on the learning models to extract sparse subspaces [11]–[14]. Through learning sparse subspaces, merely a subset of the original data variables matter during the dimensionality reduction processes, and thus the most discriminative features can be yielded. Inspired by this, the problem of sparse subspace learning has raised increasing interests, and several efficient models that explore sparse subspaces have been proposed [15]–[18]. For example, the unified sparse subspace learning (USSL) model determines the sparse subspaces in a framework of  $l_1$ -regularization based regression [19]. In [20], the discriminant locality preserving projection in terms of  $l_1$ -norm maximization (DLPP-L1) was used along with sparsity for locality preserving analysis. In addition, the sparse locality preserving projection (spLPP) model [21], the supervised discriminative sparse PCA (SDSPCA) model, the sparse local discriminant projections (SLDP) model [22], the online sparse supervised learning of extreme learning machine (ELM) model [23], and the sparse locality preserving discriminant analysis (SLPDA) model [9] were also devised for different problems. Most of the existing models utilize  $l_1$ - or  $l_\alpha$ -regularization in sparse subspace learning. Particularly, the  $l_1$ -regularization-based models (e.g., [19]–[21], etc) utilize  $l_1$ -norm to convexly surrogate the  $l_0$ -norm, and thus can be solved efficiently. The  $l_\alpha$ -regularization-based models (e.g., [24]–[26], etc) invoke non-convex  $l_\alpha$  ( $0 < \alpha < 1$ ) quasi-norms, which have been verified to perform more efficiently and robustly in inducing sparsity because the  $l_\alpha$ -norm is closer to the  $l_0$ -norm [25], [27], [28].

Notice that the sparse models as mentioned above all aim to decrease the cardinality (*i.e.*, the nonzero elements' amount) of the subspace basis vectors. However, in many applications, it might not be enough to merely constrain the cardinality of the subspace without considering the structural interrelation of the data variables. For instance, the sense organs on face images usually act as important roles in recognition, where the pixels contained in these organs on a face image are

actually correlated with each other. If one merely decreases the cardinality of the basis vectors in the derived subspace without emphasizing on the pixel composition, it might be difficult to physically interpret the face recognition process. In genetics, specific biological characters are usually dominated by a set of genes, instead of a single one. It is desirable to find groups of genes to interpret the biological expressions. In these cases, making use of  $l_1$ - or  $l_\alpha$ -regularization might not always be very efficient to encode such structural interrelations [29]. Recently, several sparse subspace learning models that take into account the structural sparsity (*i.e.*, sparsity encoding of the structural correlations of the data variables) were presented, such as the structured sparse PCA (SSPCA) method [30], the supervised principal coefficients embedding (SPCE) model [31], the simple linear iterative clustering superpixel-based  $l_{2,1}$ -norm robust principal component analysis (SURPCA<sub>21</sub>) model [32], the structural sparse locality preserving projection (SSLPP) model [24], and the dictionary learning algorithm based on the structural sparse preserving (SSP-DL) model [33], etc. However, most of these models emphasize more on the data-level locality, and thus might not always efficiently deal with the data involving properties of multimodality or mixmodality.

Due to the above considerations, here we intend to develop an  $\alpha$ -structural regularization based hierarchical locality analysis ( $\alpha$ -SRHLA) model that can incorporate both the “structural sparsity” and the “hierarchical locality” into consideration. Our method exploits the “hierarchical locality” of data and can capture the local affinity information of the samples within a same class (*i.e.*, data-level locality) as well as the local geometric correlations of different classes (*i.e.*, class-level locality). This allows for the extraction of the inherent nonlinear distribution characteristics in the considered multimodal or mixmodal data samples. Furthermore, we take the benefit of the non-convex  $l_\alpha$ -based structural norm to introduce the  $\alpha$ -structural-regularization into sparse subspace learning scenarios, encoding the structural interrelations of the data variables. An efficient algorithm are proposed to solve the non-convex optimization problem. The derived sparse subspace will be with not only small cardinality, but also the desired “sparse pattern”, and the sparsity will arise in the scale of variable group instead of the single variable, *i.e.*, the “structural sparsity”. To evaluate the efficiency of the present model, both multimodal and mixmodal experiments are conducted for data classification, digit and face recognition. Followings are the key features of our study.

- 1) Our  $\alpha$ -SRHLA model utilizes the  $\alpha$ -structural regularization, which differs from many existing sparse models in adopting  $l_1$ - or  $l_\alpha$ -regularizations. The present SHLPA model induces structurally sparse subspaces with desired “sparse patterns”, which benefits for the preservation of the structural correlations of the original variables and facilitates a better interpretation of the feature extraction procedure.
- 2) The proposed  $\alpha$ -SRHLA model can better deal with multimodal and mixmodal data. The  $\alpha$ -SRHLA model

preserves the “hierarchical locality” of data, which differs from many existing supervised learning models that merely consider the data-level locality. Through exploiting the local affinity information in both data-level (i.e., samples within a same class) and class-level (i.e., different classes) scales, the multimodal and mixmodal data manifold can be effectively learned.

- 3) For the structural regularization, different grouping approaches are provided in our model to generate structural sparsity in the scale of variable group instead of the single variable. Simulations for various multimodal and mixmodal classification/recognition tasks verify that discriminative structurally sparse subspaces can be extracted when the variables have structural interrelations with each other.

In the rest of the paper, Section 2 states the problem considered in this work. Section 3 introduces the proposed  $\alpha$ -SRHLA model. Then, we test it by a series of classification and recognition experiments in Section 4, with the conclusions given in Section 5.

## II. PROBLEM STATEMENT

Suppose that the training data matrix  $\mathbf{X} \in \mathbb{R}^{p \times a}$  is constituted by the  $p$ -dimensional samples  $\mathbf{x}_i$ 's for  $i = 1, \dots, a$  from  $C$  different classes. We aim to exploit the intrinsic characteristics inhered in the training data to determine a low-dimensional subspace. For later use, let  $\mathbf{X} = [\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(C)}]$  be the training data that follows multimodal and mixmodal distributions, where  $\mathbf{X}^{(c)} = [\mathbf{x}_1^{(c)}, \dots, \mathbf{x}_{a_c}^{(c)}]$  indicates the data matrix for the  $c$ th class,  $a_c$  denotes the size of this class, and  $c = 1, \dots, C$ .

The sparse supervised learning task considered in this paper is to find a mapping matrix  $\mathbf{\Omega} = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_q] \in \mathbb{R}^{p \times q}$  ( $q \ll p$ ), whose column vectors determine the basis vectors of the low-dimensional sparse subspace. The learned subspace is expected to be able to seize the intrinsic distribution characteristics of the multimodal and mixmodal data, with sparse patterns reflecting the structural correlations of the original data variables. By linearly mapping the data into the learned subspace, i.e., setting

$$\mathbf{Y} = \mathbf{\Omega}^T \mathbf{X}, \quad (1)$$

we can have a low-dimensional representation  $\mathbf{Y} = [\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(C)}]$  of the original high-dimensional training data matrix  $\mathbf{X}$ . Similarly, testing data samples can be linearly projected into the subspace in an inductive way for subsequent classification or recognition tasks. For better understanding, the flowchart of the proposed  $\alpha$ -SHLPA model is given in Fig. 1.

### A. USSL(UNIFIED SPARSE SUBSPACE LEARNING) [19]

The USSL framework can be applied towards the graph-based subspace learning methods like locality preserving projection [34], LDA [35], etc, with the sparse solutions computed by

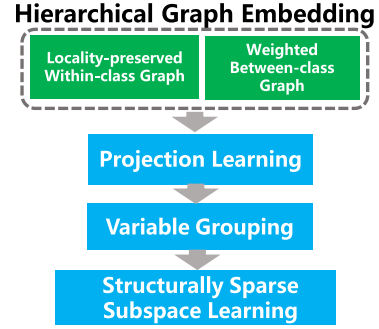


FIGURE 1. The flowchart of the  $\alpha$ -SRHLA model.

a  $l_1$ -norm regularizer as follows:

$$\min_{\boldsymbol{\omega}} \|y_i - \boldsymbol{\omega}^T \mathbf{x}_i\| + \lambda \|\boldsymbol{\omega}\|_1,$$

where  $\lambda$  is the sparsity parameter.

### B. SLPDA(SPARSE LOCALITY PRESERVING DISCRIMINANT ANALYSIS) [9]

The SLPDA model aims to yield the nonlinear characteristics of the data structure in an  $l_\alpha$ -regularization scheme, with its objective function defined as:

$$\min_{\mathbf{\Omega}=[\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_q]} \sum_{j=1}^q \sum_{i=1}^a (\boldsymbol{\omega}_j^T \mathbf{x}_i - y_{ji})^2 + \gamma \sum_{j=1}^q \|\boldsymbol{\omega}_j\|_\alpha,$$

where  $0 < \alpha < 1$ ,  $\gamma \geq 0$  is the sparsity parameter,  $\mathbf{\Omega} = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_q]$  is the mapping matrix, and the projection matrix  $\mathbf{Y} = (y_{ji})$  is determined by the locality preserving discriminant analysis [9].

### III. THE $\alpha$ -STRUCTURAL REGULARIZATION BASED HIERARCHICAL LOCALITY ANALYSIS ( $\alpha$ -SRHLA)

For multimodal and mixmodal data, merely considering the data-level locality like most existing methods might be insufficient to capture the complex distribution characteristic. On the other side, simply constraining the cardinality of the learned subspaces might not always effectively describe the structural correlations of the original data variables. In view of this, here we propose an  $\alpha$ -SRHLA model that incorporates the “hierarchical locality” with “structural sparsity” and is hence capable of dealing with the multimodal and mixmodal cases.

The  $\alpha$ -SRHLA model consists of two steps, i.e., projection learning and sparse subspace learning. To be specific, in the projection learning step, the  $\alpha$ -SRHLA model aims to learn the low-dimensional projections that preserve the “hierarchical locality” of the multimodal and mixmodal data. Then in the step of sparse subspace learning, the structural correlations of the data variables are taken into account to induce structurally sparse subspaces with desired “sparse pattern” (i.e., the derived sparsity reflects the structural correlations between variables). Considering the efficiency of the non-convex penalties (e.g.,  $l_\alpha$  quasi-norms with  $0 < \alpha < 1$ )

in inducing sparsity, we will make use of non-convex regularization [24], [25]. Through recursively conducting the steps of projection learning as well as sparse subspace learning, subspaces capturing the intrinsic distribution characteristics of the considered data with multimodal and mixmodal properties can be yielded.

**A. PROJECTION LEARNING**

The projection learning procedure is to get low-dimensional projections that capture the hierarchical local characteristics of multimodal and mixmodal data. To this end, we introduce the within-class and between-class affinity matrices as follows. For the  $c$ th class, let  $\mathbf{x}_i^{(c)}$  denote the data sample indexed by  $i$ , then the within-class affinity matrix  $\mathbf{W}^{(c)}$  encoding the data-level locality of this class is defined by

$$W_{ij}^{(c)} = \begin{cases} \exp\left[-\frac{\|\mathbf{x}_i^{(c)} - \mathbf{x}_j^{(c)}\|^2}{\theta^{(c)^2}}\right], & \text{if } \|\mathbf{x}_i^{(c)} - \mathbf{x}_j^{(c)}\|^2 < \varepsilon^{(c)}, \\ 0, & \text{otherwise,} \end{cases}$$

where  $i, j = 1, \dots, a_c, c = 1, 2, \dots, C, \theta^{(c)}$  is the heat kernel parameter,  $\varepsilon^{(c)}$  measures the size of the data-level local geometric affinity of the  $c$ th class. Similarly, the between-class weight matrix  $\mathbf{W}$  capturing the class-level locality is defined by

$$W_{c_1 c_2} = \begin{cases} \exp\left[-\frac{\|\mathbf{e}_{c_1} - \mathbf{e}_{c_2}\|^2}{\theta^2}\right], & \text{if } \|\mathbf{e}_{c_1} - \mathbf{e}_{c_2}\|^2 < \varepsilon, \\ 0, & \text{otherwise,} \end{cases}$$

where  $c_1, c_2 = 1, \dots, C, \theta$  and  $\varepsilon$  are the kernel and class-level locality parameters respectively,

$$\mathbf{e}_{c_1} = \frac{1}{a_{c_1}} \sum_{t=1}^{a_{c_1}} \mathbf{x}_t^{(c_1)}, \quad \mathbf{e}_{c_2} = \frac{1}{a_{c_2}} \sum_{t=1}^{a_{c_2}} \mathbf{x}_t^{(c_2)},$$

computed as the mean vectors, are respectively the representatives of the  $a_{c_1}$  and  $a_{c_2}$  samples for the  $c_1$ th and  $c_2$ th classes.

By using the within-class affinity matrices  $\mathbf{W}^{(c)}$ , we define  $\mathbf{S}_W$  and  $s_w$  as below to measure the weighted covariances of the original and projected samples from the same class respectively:

$$\begin{aligned} \mathbf{S}_W &= \sum_{c=1}^C \mathbf{X}^{(c)} \mathbf{L}^{(c)} \left[ \mathbf{X}^{(c)} \right]^T \\ &= \left[ \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(C)} \right] \begin{bmatrix} \mathbf{L}^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{L}^{(C)} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(C)} \end{bmatrix} \end{aligned} \quad (2)$$

and

$$\begin{aligned} s_w &= \sum_{c=1}^C \mathbf{y}^T \mathbf{L}^{(c)} \mathbf{y} \\ &= \mathbf{y}^T \begin{bmatrix} \mathbf{L}^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{L}^{(C)} \end{bmatrix} \mathbf{y} \end{aligned} \quad (3)$$

where  $\mathbf{y} = \mathbf{X}^T \boldsymbol{\omega}$  denotes the 1-dimensional projection of the data matrix  $\mathbf{X}, \mathbf{L}^{(c)}$  is the Laplacian matrix determined by

$$\mathbf{L}^{(c)} = \mathbf{D}^{(c)} - \mathbf{W}^{(c)},$$

and  $\mathbf{D}^{(c)} \in \mathbb{R}^{a_c \times a_c}$  is a diagonal matrix with each of its diagonal entries computed as  $D_{ij}^{(c)} = \sum_j W_{ij}^{(c)}$ . Besides, we introduce an auxiliary matrix  $\mathbf{T}_1$  as

$$\mathbf{T}_1 = \text{diag}[\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(C)}] \in \mathbb{R}^{a \times a},$$

and thus the within-class scatter matrices  $\mathbf{S}_W$  and  $s_w$  can be reformulated by

$$\mathbf{S}_W = \mathbf{X} \mathbf{T}_1 \mathbf{X}^T, \quad s_w = \mathbf{y}^T \mathbf{T}_1 \mathbf{y}.$$

Evidently,  $\mathbf{T}_1$  encodes the local geometric information in the data-level scale, and thus  $\mathbf{S}_W$  and  $s_w$  capture the locality-preserved within-class distances in the original space and the low-dimensional subspace respectively.

Moreover, we compute the Laplacian matrix  $\mathbf{L} = \mathbf{D} - \mathbf{W}$  with the diagonal matrix determined by  $D_{c_1 c_1} = \sum_{c_2} W_{c_1 c_2}$ , and introduce another auxiliary matrix  $\mathbf{T}_2 = \text{diag}\left[\frac{1(a_1)}{a_1}, \dots, \frac{1(a_C)}{a_C}\right] \in \mathbb{R}^{a \times C}$  to formulate  $\mathbf{S}_B$  and  $s_b$  that describe the weighted covariances of the original and projected samples from different classes as follows:

$$\mathbf{S}_B = \mathbf{E} \mathbf{L} \mathbf{E}^T = \mathbf{X} \mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T \mathbf{X}^T, \quad (4)$$

where

$$\begin{aligned} \mathbf{E} &= [\mathbf{e}_1, \dots, \mathbf{e}_C] \\ &= \left[ \mathbf{X}^{(1)}, \dots, \mathbf{X}^{(C)} \right] \begin{bmatrix} \frac{1(a_1)}{a_1} & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & \dots & 0 & \frac{1(a_C)}{a_C} \end{bmatrix}, \end{aligned}$$

$\mathbf{1}(a_c) = [1, \dots, 1] \in \mathbb{R}^{a_c}$ , and

$$s_b = \mathbf{y}^T \mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T \mathbf{y}.$$

Clearly,  $s_b$  indicates the weighted separability between the projected samples from different classes, i.e., the weighted between-class separability in the derived subspace.

Now, we are ready to formulate the projection learning as the following optimization problem:

$$\max_{\mathbf{y}} \frac{s_b}{s_w} \quad \text{i.e.,} \quad \max_{\mathbf{y}} \frac{\mathbf{y}^T \mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T \mathbf{y}}{\mathbf{y}^T \mathbf{T}_1 \mathbf{y}}, \quad (5)$$

which maximizes the weighted between-class separability and minimizes the locality-preserved within-class distances. If data follows within-class multimodal distributions, the projections of the samples from the same/different modalities within a same class will be assigned with larger/smaller weights by  $\mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T$  so as to maintain such local geometric correlations in the low-dimensional subspace. Similarly, for mixmodal data, the locality in class-level scale is preserved by  $\mathbf{T}_1$ , with larger/smaller values given to the farther/closer pairs of classes to facilitate the aggregation/separation of their projections. In this way, the learned projection  $\mathbf{y}$  achieved by (5)

can well seize the hierarchical distribution characteristics of multimodal and mixmodal data.

In this setting, the generalized eigenvectors corresponding to the first  $q$  largest eigenvalues of the following generalized eigenvalue problem

$$\mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T \mathbf{y} = \lambda \mathbf{T}_1 \mathbf{y}, \quad (6)$$

determine the row vectors of the expected low-dimensional projection matrix  $\mathbf{Y} \in \mathbb{R}^{q \times a}$ . Each element  $y_{it}$  in  $\mathbf{Y}$  indicates the projection of the data sample  $\mathbf{x}_i$  onto the direction of the column vector  $\boldsymbol{\omega}_t$  in the mapping matrix  $\boldsymbol{\Omega}$  for  $t = 1, \dots, q$  and  $i = 1, \dots, a$ .

Similarly, the column vectors of the mapping matrix  $\boldsymbol{\Omega}$  can be determined by the optimization problem:

$$\max_{\boldsymbol{\omega}} \frac{\boldsymbol{\omega}^T \mathbf{S}_B \boldsymbol{\omega}}{\boldsymbol{\omega}^T \mathbf{S}_W \boldsymbol{\omega}}, \quad (7)$$

which can be solved by finding the generalized eigenvectors  $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_q$  corresponding to the first  $q$  largest eigenvalues of the following generalized eigenvalue problem:

$$\mathbf{S}_B \boldsymbol{\omega} = \lambda \mathbf{S}_W \boldsymbol{\omega}. \quad (8)$$

Usually, the generalized eigenvector solution to (8) captures the hierarchical locality of multimodal and mixmodal data with non-sparse pattern, and its relation to the low-dimensional projection matrix  $\mathbf{Y}$  yielded by (6) is as follows.

*Lemma 1:* For linear dimensionality reduction, the optimal solution to (7) equals to the largest generalized eigenvalue solution to (6).

*Proof:* The optimal solution to (7) can be determined by computing the largest generalized eigenvalue of (8). Insert (2) and (4) into (8) yields

$$\mathbf{X} \mathbf{T}_2 \mathbf{L} \mathbf{T}_2^T \mathbf{X}^T \boldsymbol{\omega} = \lambda \mathbf{X} \mathbf{T}_1 \mathbf{X}^T \boldsymbol{\omega}.$$

Recall in (3) that  $\mathbf{X}^T \boldsymbol{\omega} = \mathbf{y}$ , the conclusion of Lemma 1 thus follows.

Lemma 1 indicates that for linear dimensionality reduction, the first step of the proposed  $\alpha$ -SRHLA model, i.e., projection learning, can derive low-dimensional projections that seize the characteristics of the multimodal and mixmodal data in a *non-sparse* subspace. To find *sparse* subspaces and facilitate better interpretation for the feature extraction process, below we will take into account the structural interrelations of the data variables to learn a mapping matrix  $\boldsymbol{\Omega}$  with structurally sparse pattern.

## B. SPARSE SUBSPACE LEARNING

The sparse subspace learning procedure is to learn a sparse subspace with the desired sparse pattern by using the low-dimensional projection matrix  $\mathbf{Y} \in \mathbb{R}^{q \times a}$ . To derive structurally sparse subspaces with desired sparse patterns, it is required to take into account the structural correlations between all the original data variables in order to generate the grouping rule  $\mathcal{O} = \{O_1, \dots, O_f\}$ , with any  $O_i \subseteq \{1, \dots, p\}$

and  $\cup_{O_i \in \mathcal{O}} O_i = \{1, \dots, p\}$ . Variables correlated with each other can be set into a group and regarded as an entire structural variable. Based on the grouping, the structural regularization is defined by

$$\begin{aligned} N_{\alpha}(\boldsymbol{\omega}_t) &= \sqrt{\alpha} \sum_{O \in \mathcal{O}} \left( \sqrt{\sum_{j \in O} (d_j^O \times \omega_{jt})^2} \right)^{\alpha} \\ &= \left\| \left( \|\mathbf{d}^O \circ \boldsymbol{\omega}_t\|_2 \right)_{O \in \mathcal{O}} \right\|_{\alpha}, \end{aligned}$$

where  $0 < \alpha < 1$ ,  $\omega_{jt}$  is the  $j$ th entry of  $\boldsymbol{\omega}_t$ ,  $\mathbf{d}^O \circ \boldsymbol{\omega}_t$  means element-wise multiplication, and

$$d_j^O = \begin{cases} 1, & j \in O, \\ 0, & \text{otherwise.} \end{cases}$$

Then, we impose the  $\alpha$ -structural penalty  $N_{\alpha}(\boldsymbol{\omega}_t)$  on the subspace basis vector  $\boldsymbol{\omega}_t \in \mathbb{R}^p$  and consider the following penalized least-square problem:

$$\min_{\boldsymbol{\Omega} \in \mathbb{R}^{p \times q}} \sum_{t=1}^q \sum_{i=1}^a \left( \boldsymbol{\omega}_t^T \mathbf{x}_i - y_{it} \right)^2 + \beta \sum_{t=1}^q N_{\alpha}(\boldsymbol{\omega}_t), \quad (9)$$

where  $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_q] \in \mathbb{R}^{p \times q}$  is the mapping matrix to be determined,  $y_{it}$  is the element in the  $i$ th row and  $t$ th column of  $\mathbf{Y}$ ,  $\beta$  is the sparsity parameter. Benefiting from the regrouping, solving (9) will induce sparsity in the scale of the variable group instead of the single variable. And the loadings for the variable groups are sparse, whereas those for the variables within a group are nonsparse. This is referred to as ‘‘structural sparsity’’. In this manner, the resulting structurally sparse subspaces can be with the corresponding sparse patterns that characterize the structural correlations between the data variables. To solve (9), we introduce the following.

*Lemma 2* [30]: Suppose that  $\alpha \in (0, 2)$  and  $\gamma = \frac{\alpha}{2-\alpha}$ , for any  $\boldsymbol{\delta} \in \mathbb{R}^p$ , we have

$$\|\boldsymbol{\delta}\|_{\alpha} = \min_{\mathbf{z} \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^p \frac{\delta_i^2}{z_i} + \frac{1}{2} \|\mathbf{z}\|_{\gamma},$$

with the minimum value achieved by  $z_i = |\delta_i|^{2-\alpha} \|\boldsymbol{\delta}\|_{\alpha}^{\alpha-1}$  for  $i = 1, \dots, p$ .

Clearly, Lemma 2 transforms the non-convex regularization of  $\boldsymbol{\delta}$  into an optimization problem upon another variable  $\mathbf{z}$ , which makes it possible for us to find the optimal solution to (9) by utilizing the univariate search technique [30]. According to Lemma 2, the non-convex  $\alpha$ -structural regularization imposed on  $\boldsymbol{\omega}_t$  can be reformulated as

$$N_{\alpha}(\boldsymbol{\omega}_t) = \min_{\boldsymbol{\pi}_t \in \mathbb{R}^f} \frac{1}{2} \left[ \|\boldsymbol{\pi}_t\|_{\gamma} + \sum_{O \in \mathcal{O}} \|\boldsymbol{\omega}_t \circ \mathbf{d}^O\|_2^2 (\pi_t^O)^{-1} \right], \quad (10)$$

where  $\gamma = \frac{\alpha}{2-\alpha} \in (0, 1)$ ,  $0 < \alpha < 1$ ,  $\boldsymbol{\Pi} = [\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_q] \in \mathbb{R}^{f \times q}$ , with each of its entries defined by

$$\pi_t^O = \left\| \mathbf{d}^O \circ \boldsymbol{\omega}_t \right\|_2^{2-\alpha} \cdot \left\| \left( \|\mathbf{d}^O \circ \boldsymbol{\omega}_t\|_2 \right)_{O \in \mathcal{O}} \right\|_{\alpha}^{\alpha-1}, \quad (11)$$

for  $O \in \mathcal{O}$ ,  $t = 1, \dots, q$ . This along with (10) transforms the optimization problem (9) into the following:

$$\min_{\Omega \in \mathbb{R}^{p \times q}, \Pi \in \mathbb{R}^{f \times q}} \sum_{t=1}^q \sum_{i=1}^a \left( \omega_t^T \mathbf{x}_i - y_{it} \right)^2 + \frac{\beta}{2} \sum_{t=1}^q \left( \|\pi_t\|_\gamma + (\omega_t)^T \text{diag}[\phi_t] \omega_t \right), \quad (12)$$

where each entry in  $\phi_t$  is calculated as

$$\phi_{it} = \sum_{O \in \mathcal{O}, i \in \mathcal{O}} \left( d_i^O \right)^2 \left( \pi_t^O \right)^{-1}, \quad (13)$$

and  $\text{diag}[\phi_t] \in \mathbb{R}^{p \times p}$  is a diagonal matrix that takes  $\phi_t$  as its diagonal, with the other elements set as 0. To solve the optimization problem (12), we alternatively update the variables with the following steps:

- 1) We first optimize the matrix  $\Pi$ , with  $\omega_t$ 's taking fixed values. The initial value of  $\Omega$  is set as the solution yielded from the projection learning step. By utilizing Lemma 2, the closed-form solution of  $\pi_t^O$  can be readily determined by

$$\pi_t^{O(k)} = \left\| \mathbf{d}^O \circ \omega_t^{(k-1)} \right\|_2^{2-\alpha} \times \left\| \left( \left\| \mathbf{d}^O \circ \omega_t^{(k-1)} \right\|_2 \right) \right\|_\alpha^{\alpha-1},$$

where  $O \in \{O_1, \dots, O_f\}$ ,  $t = 1, \dots, q$ , and  $k$  is the iteration round index. To avoid possible numerical instability, during the optimization process, we add a small constant  $\sigma$  ( $0 < \sigma \ll 1$ ) to  $\pi_t^O$  as follows:

$$\pi_t^{O(k)} \leftarrow \pi_t^{O(k)} + \sigma.$$

- 2) Then, we turn to optimize the mapping matrix  $\Omega$ , with  $\Pi$  taking a fixed value of the solution in the last optimization round. As the objective function in (12) is continuously differentiable with respect to  $\omega_t$ , based on the BCD method [36],  $\omega_t$  can be optimized by

$$\omega_t^{(k)} \leftarrow \left( \mathbf{X}\mathbf{X}^T + \frac{\beta}{2} \text{diag}[\phi_t^{(k)}] \right)^{-1} \times \mathbf{X}\mathbf{Y}^T \mathbf{e}_t, \quad (14)$$

where  $\phi_t$  is determined by (13),  $t = 1, 2, \dots, q$ .

Algorithm 1 provides the details of the  $\alpha$ -SRHLA optimization. For linear dimensionality reduction, to compare the results obtained from the projection learning and sparse subspace learning processes, we give the following result.

*Theorem 1:* Let the column vectors of  $\mathbf{Y}^T \in \mathbb{R}^{q \times a}$  be the generalized eigenvectors of (6) corresponding to the first  $q$  largest generalized eigenvalues, when  $\beta$  tends to 0, the solutions to the sparse subspace learning problem (9) are the generalized eigenvector solutions to (7) corresponding to the same eigenvalues.

*Proof:* For the optimization problem (12), we firstly set its derivative with respect to  $\omega_t$  as 0, and arrive at

$$\left( \mathbf{X}\mathbf{X}^T + \frac{\beta}{2} \text{diag}[\phi_t] \right) \omega_t = \mathbf{X}\mathbf{Y}^T \mu_t, \quad (15)$$

where  $\mu_t = [0, \dots, 1, \dots, 0]^T \in \mathbb{R}^q$  sets its  $t$ th entry as 1 and the others 0. If the rank of the data matrix  $\mathbf{X}$  is  $r$ , the SVD of  $\mathbf{X}$  can be written as

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T,$$

where  $\mathbf{U} \in \mathbb{R}^{p \times p}$  and  $\mathbf{V} \in \mathbb{R}^{a \times a}$  are both orthogonal matrices, the diagonal elements of  $\mathbf{S} \in \mathbb{R}^{p \times a}$  are the singular values of data matrix  $\mathbf{X}$ . Suppose that the first  $r$  diagonal elements of  $\mathbf{S}$  are positive, according to [37], the pseudo inverse of  $\mathbf{X}$  can be derived by

$$\mathbf{X}^+ = \mathbf{V}\mathbf{S}^+ \mathbf{U}^T,$$

where  $\mathbf{S}^+$  takes the reciprocals of the  $r$  positive singular values, i.e., the first  $r$  diagonal entries of  $\mathbf{X}$  in proper orders, and then be transposed into size of  $a \times p$ . From (15), it follows that

$$\left( \mathbf{X}^+ \mathbf{X}\mathbf{X}^T + \frac{\beta}{2} \mathbf{X}^+ \text{diag}[\phi_t] \right) \omega_t = \mathbf{X}^+ \mathbf{X}\mathbf{Y}^T \mu_t.$$

Taking the SVD of matrices  $\mathbf{X}$  and  $\mathbf{X}^+$ , it can be further obtained that

$$\left( \mathbf{V}\mathbf{S}^+ \mathbf{S}\mathbf{V}^T \mathbf{X}^T + \frac{\beta}{2} \mathbf{X}^+ \text{diag}[\phi_t] \right) \omega_t = \mathbf{V}\mathbf{S}^+ \mathbf{S}\mathbf{V}^T \mathbf{Y}^T \mu_t. \quad (16)$$

Let  $\mathbf{V} = [\mathbf{A}, \mathbf{B}]$  with  $\mathbf{A} \in \mathbb{R}^{a \times r}$  and  $\mathbf{B} \in \mathbb{R}^{a \times (a-r)}$ , it can thus be verified that  $\mathbf{A}^T \mathbf{A} = \mathbf{I}_r$  and  $\mathbf{B}^T \mathbf{B} = \mathbf{I}_{a-r}$  with  $\mathbf{I}_r$  and  $\mathbf{I}_{a-r}$  as identity matrices. This together with (16) leads to the following

$$[\mathbf{A} \ \mathbf{B}] \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{a-r} \end{bmatrix} \begin{bmatrix} \mathbf{A}^T \\ \mathbf{B}^T \end{bmatrix} \mathbf{X}^T \omega_t + \frac{\beta}{2} \mathbf{X}^+ \text{diag}[\phi_t] \omega_t = [\mathbf{A} \ \mathbf{B}] \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{a-r} \end{bmatrix} \begin{bmatrix} \mathbf{A}^T \\ \mathbf{B}^T \end{bmatrix} \mathbf{Y}^T \mu_t,$$

and hence

$$\left( \mathbf{A}\mathbf{A}^T \mathbf{X}^T + \frac{\beta}{2} \mathbf{X}^+ \text{diag}[\phi_t] \right) \omega_t = \mathbf{A}\mathbf{A}^T \mathbf{Y}^T \mu_t. \quad (17)$$

By taking  $\beta \rightarrow 0$  on both the left and the right of the equation (17), we have

$$\mathbf{A}\mathbf{A}^T \mathbf{X}^T \omega_t = \mathbf{A}\mathbf{A}^T \mathbf{Y}^T \mu_t.$$

As  $\mathbf{X}^T = \mathbf{A}\mathbf{Z}_1$  for some  $\mathbf{Z}_1 \in \mathbb{R}^{r \times p}$ , it can be yielded that

$$\mathbf{A}\mathbf{A}^T \mathbf{X}^T = \mathbf{A}\mathbf{A}^T \mathbf{A}\mathbf{Z}_1 = \mathbf{A}\mathbf{Z}_1 = \mathbf{X}^T. \quad (18)$$

Besides, for linear dimensionality reduction, since the column vectors in  $\mathbf{Y}^T$  are within the space spanned by the row vectors of  $\mathbf{X}$ , it can be derived that  $\mathbf{Y}^T = \mathbf{A}\mathbf{Z}_2$  for some  $\mathbf{Z}_2 \in \mathbb{R}^{r \times q}$  and hence  $\mathbf{A}\mathbf{A}^T \mathbf{Y}^T = \mathbf{Y}^T$ . By utilizing (17) and (18), we have

$$\mathbf{X}^T \omega_t = \mathbf{Y}^T \mu_t.$$

This along with Lemma 1 completes the proof of Theorem 1.

The computational complexity of the  $\alpha$ -SHLPA model is mainly due to the eigenvector and the inversion computation in (8) and 14. Since the rank of  $\mathbf{S}_B$  is no more than  $C - 1$ , the complexity of the eigenvector computation for (8) based on the Gram-Schmit method is  $\mathcal{O}(pC^2)$ . The inversion computation occupies  $\mathcal{O}(p^3)$  complexity.

**Algorithm 1**  $\alpha$ -SRHLA

**INPUT:** Training data matrix  $\mathbf{X}$ ,  $\sigma > 0$ ,  $\beta > 0$ , variable grouping  $\mathcal{O} = \{O_1, \dots, O_f\}$ .

Determine  $\mathbf{Y}$  by (6).

Initiate  $\mathbf{\Omega}^{(0)}$  and  $\mathbf{\Pi}^{(0)}$  respectively by Eqs.(7) and (11).

**WHILE** convergence is not obtained

**FOR**  $t = 1$  to  $q$ ,  $O = O_1$  to  $O_f$

$$\pi_t^{O(k)} = \|\mathbf{d}^O \circ \boldsymbol{\omega}_t^{(k-1)}\|_2^{2-\alpha} \left\| \|\mathbf{d}^O \circ \boldsymbol{\omega}_t^{(k-1)}\|_2, O \in \mathcal{O} \right\|_\alpha^{\alpha-1}$$

$$\pi_t^{O(k)} \leftarrow \pi_t^{O(k)} + \sigma$$

**END FOR**

**WHILE** convergence is not obtained

$$\phi_{it}^{(k)} = \sum_{O \in \mathcal{O}, i \in O} (d_i^O)^2 \left( \pi_t^{O(k)} \right)^{-1}$$

**FOR**  $t = 1$  to  $q$

$$\boldsymbol{\omega}_t^{(k)} \leftarrow \left[ \mathbf{X}\mathbf{X}^T + \frac{\beta}{2} \text{diag}[\boldsymbol{\phi}_t^{(k)}] \right]^{-1} \times \mathbf{X}\mathbf{Y}^T \boldsymbol{\mu}_t$$

**END FOR**

**END WHILE**

**END WHILE**

**OUTPUT:**  $\mathbf{\Omega}^{(k)}$ .

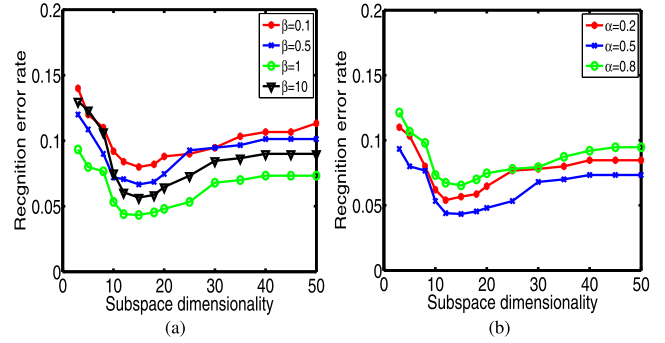
**IV. EXPERIMENTS**

In this section, in order to evaluate the proposed  $\alpha$ -SRHLA model, experiments for normal and multimodal data classification, multimodal and mixmodal digit as well as face recognition are respectively performed. For comparison, we choose the LDA model, the LPDA model, the USSL model, and the SLPDA model, where the USSL model is conducted for the supervised LDA and LPDA learning procedures respectively. In the present  $\alpha$ -SRHLA model, for  $\mathbf{W}^{(c)}$ , we use  $k^{(c)} = a_c - 1$  to denote the locality of the  $c$ th class, where  $a_c$  is the number of this class [38]. Similarly for  $\mathbf{W}$ , the locality parameter  $k_B$  is set as  $k_B = C - 1$ . For the sparsity parameter  $\gamma$ , in experiments, values from the set  $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$  will be tested alternatively and better results are reported. According to [25] that better performance of the  $l_\alpha$ -regularization is usually achieved by  $\alpha = 0.5$ ,  $\alpha$  in our  $\alpha$ -SRHLA model is thus assigned as 0.5. The parameters utilized by the other models follow from [9], [19], and [35], respectively.

**A. PARAMETER TESTING EXPERIMENTS**

In the parameter-testing experiments, digit recognition on the United States Postal Service (USPS) database are conducted. The USPS database is constituted by the automatically-scanned digit images of the envelopes from USPS and has 9,298 images of different sizes and orientations. Here we resize the training and testing images into  $16 \times 16$  pixels.

In our experiments, the training set is constituted by randomly-selected 4500 images, among which the same number of images for “1” to “9” are selected. Another 1500 images are randomly chosen for testing. We first apply the proposed models to learn subspaces and then map the testing images into the learned subspaces with low-dimensional



**FIGURE 2.** Parameter testing: Average error rates of the digit recognition obtained by different values of (a) the structural sparsity parameter and (b) regularization parameter.

representations. Then, we fulfill the recognition task by using  $k$ -nn classifier ( $k = 5$ ).

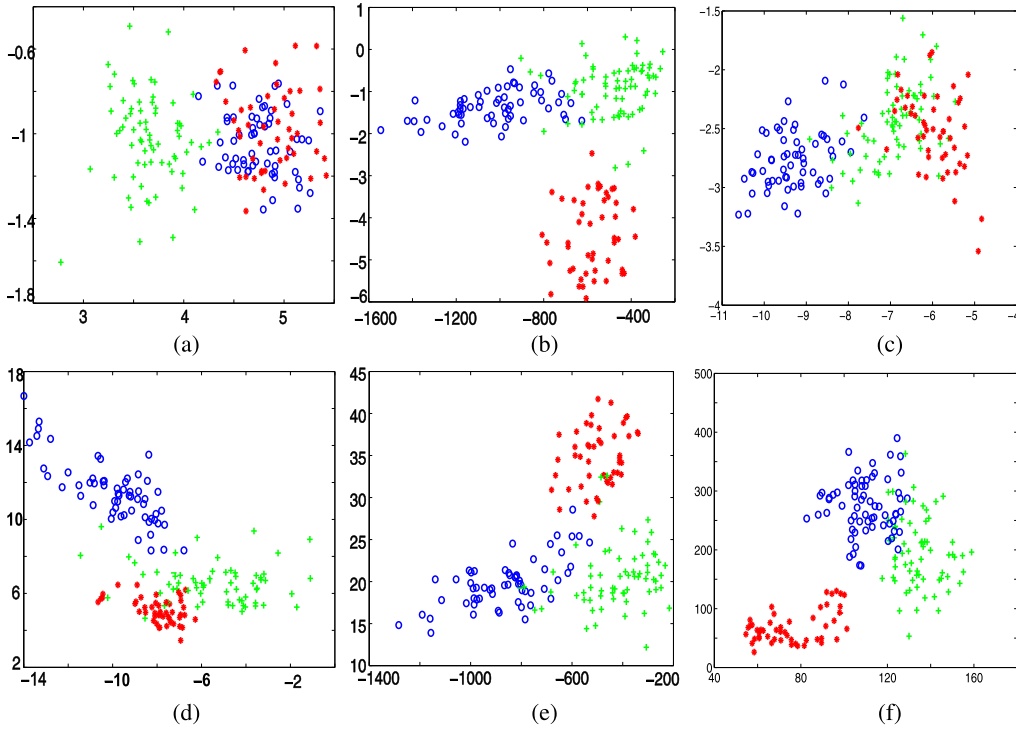
Firstly, experiments upon the structural sparsity parameter  $\beta$  are conducted, which are tested respectively by choosing values from  $\{\beta = 0.1, 0.5, 1, 10\}$  for fair comparison. The regularization parameter  $\alpha$  is set as 0.5. The obtained average results of the  $\alpha$ -SRHLA model are reported in Fig. 2a. We proceed to perform experiments upon the regularization parameter  $\{\alpha = 0.2, 0.5, 0.8\}$  and report the achieved results in Fig. 2b, with  $\beta$  set as 1. It can be seen from the results that in this digit recognition trial, better performance is achieved when the structural sparsity parameter takes the value of 1. For the regularization parameter  $\alpha$ , clearly from Fig. 2b, better results are attained at  $\alpha = 0.5$ . The results also indicate that compared with the structural sparsity parameter, the regularization parameter performs more robustly in experiments.

**B. MULTIMODAL DATA CLASSIFICATION**

The Wine dataset<sup>1</sup> is firstly selected to evaluate the efficiency of the proposed  $\alpha$ -SRHLA model in dealing with both normal and multimodal data classification. In the Wine dataset, there are 178 data samples dropping into 3 classes, which are from 13-dimensional space. To test the multimodal data classification performance of the considered models, preprocess upon the Wine database is taken to make it multimodal. To be specific, we gather the second and the third classes of samples together to let the samples in the gathered class share the same labels. For the proposed  $\alpha$ -SRHLA model, since the “Ash” feature is correlated with the feature of “Alcalinity of ash”, and the sixth feature also has correlations with the eighth one, we generate the following groupings for all 13 data variables in the  $\alpha$ -structural regularization:

$$\mathcal{O} = \{ \{Alcohol\}, \{Malic\ acid\}, \{Magnesium\}, \{Flavanoids\}, \{Proanthocyanins\}, \{Color\ intensity\}, \{Hue\}, \{OD280/OD315\ of\ diluted\ wines\}, \{Proline\}, \{Ash, Alcalinity\ of\ ash\}, \{Total\ phenols, Nonflavanoid\ phenols\} \}.$$

<sup>1</sup><http://cmp.felk.cvut.cz/pub/cmp/articles/Franc-TR-2004-08.pdf>



**FIGURE 3. Multimodal data classification: The recognition comparison of subspaces derived by (a) LDA, (b) LPDA, (c) USSL1, (d) USSL2, (e) SLPDA, and (f)  $\alpha$ -SRHLA.**

**TABLE 1. The ratios of average between-class distances and within-class distances for data classification.**

Model	LDA	LPDA	USSL1	USSL2	SLPDA	$\alpha$ -SRHLA
Ratio	1.53	2.52	1.68	2.55	2.86	2.89

Through the supervised learning processes conducted by the all compared models, the classification results in the learned 2-dimensional subspaces are shown in Fig.3. We further compute the ratio of the average between-class distances and the within-class distances to evaluate the classification performance of the selected models as follows:

$$ratio = \frac{\prod_c a_c \sum_{c_1, c_2 \in \{1, \dots, C\}} \sqrt{\|e_{c_1} - e_{c_2}\|^2}}{C! \sum_{c \in \{1, \dots, C\}} \sum_{i \in \{1, \dots, a_c\}} \sqrt{\|x_i^{(c)} - e_c\|^2}},$$

and compare the considered models in Table 1.

From the classification results in Table 1 and Fig. 3, it can be seen that the LDA and USSL1 (USSL sparse learning for LDA) model achieve mixed results in the multimodal case. This is because the LDA model considers the global instead of the local geometric correlations of the samples, which for the multimodal case, facilitates the aggregation of the samples from the second and third classes. As to the LPDA, USSL2 (USSL sparse learning for LPDA), and SLPDA models, it can be seen from the results that these models work well in the multimodal cases. This indicates that preserving the hierarchical local affinity information of both within-class and between-class scenarios are of benefit to deal with multimodal data. Compared with other sparse

models, the sparsity achieved by present  $\alpha$ -SRHLA model is induced in the scale of the variable group. From the results in Table 1, Fig. 3f, it can be seen that our  $\alpha$ -SRHLA model gives competitive performance in the experiments. This verifies the efficiency of incorporating “structural sparsity” with “hierarchical locality”.

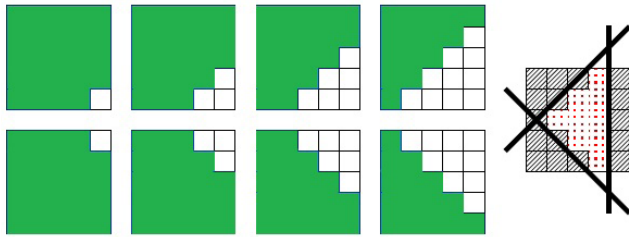
### C. MULTIMODAL AND MIXMODAL HANDWRITING RECOGNITION

In this part, handwriting recognition experiments are carried out upon the United States Postal Service (USPS) database<sup>2</sup> and MNIST database.<sup>3</sup> The USPS database is constituted by the 9298 digit images ranging from 1 to 9 on the envelopes of USPS, and the MNIST database consists of 70,000 images of handwritten digits. Different sizes and orientations are included in both databases. To test all selected models in dealing with multimodal and mixmodal data, we firstly preprocess the selected training data to be with multimodal and mixmodal properties respectively. Inspired by [29] that the  $\pm\pi/4$ -orientation grouping rules (shown in Fig. 4) for the variables helps to improve the robustness for structural regularization, we adopt such grouping method for the

<sup>2</sup><https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html>

<sup>3</sup><http://yann.lecun.com/exdb/mnist/>





**FIGURE 4.** The  $\pm\frac{\pi}{4}$ -orientational structural grouping rule (left) and the induced the diamond-shaped sparse patterns (right).

$\alpha$ -SRHLA model and generate the diamond-shaped sparse patterns shown in Fig. 4.

### 1) MULTIMODAL SITUATION

To test the considered models for multimodal data, two tasks are respectively carried out, i.e., USPS-eo (separating the even and odd digits) and USPS-sl (separating the large and small numbers). For each case, binary labels are established. Clearly, data under the same labels might follow multimodal distributions. The supervised learning process upon the USPS-eo multimodal data is illustrated by Figure 5(a).

For each of the USPS-eo and USPS-sl experiments, we randomly choose 1500 digit images from the USPS database to comprise the training set, and then select another 1500 images to test the learned subspaces. Alternatively, the training images are tackled by our  $\alpha$ -SRHLA model and the others. Then, the testing digit images are mapped into the subspaces learned by the all considered models to determine the low-dimensional projections. To classify the projections in the learned subspace, the  $k$ -nn classifier for  $k = 5$  is utilized to derive the estimated labels for testing samples. In each set of experiments, a total number of 20 trails are made to obtain the average recognition result. In Fig. 6, the average recognition error rates for USPS-eo and USPS-sl tasks are respectively shown in subfigures (a) and (b) under different subspace dimensions of 5 to 50.

Clearly from Figs. 6(a) and (b), the proposed  $\alpha$ -SRHLA model performs better in most cases when compared with the others. For instance of setting the subspace dimensionality as 10 in USPS-eo experiment, the achieved recognition error rates of the LDA, SLPDA, USSL1, USSL2, and  $\alpha$ -SRHLA models are 14.8%, 4.5%, 8.7%, 4.1%, 4.1%, and 3.3% respectively. This means that when incorporating the locality in the data-level scale with that in the class-level scale, the manifold characteristics for the considered data with multimodal and mixmodal distributions can be captured. Compared with LPDA and SLPDA, our  $\alpha$ -SRHLA model works better, which indicates that taking into account the structural correlations of variables in learning sparse subspaces facilitates the achievement of more discriminative subspaces.

### 2) MIXMODAL SITUATION

We proceed to carry out experiments for mixmodal data, which is generated by assigning the digit images from the

same database by the same labels. To be specific, we firstly randomly choose 900 images including the same amount of digits for “1” to “9” from the USPS database, and label these samples as the same. Then, we choose another database, i.e., MNIST database, and select the same amount (900 images) of digit images by the same way as the above. Labels for the samples from the MNIST database are the same. In this manner, binary labels are generated. Although images for the same digits from different databases might have closer Euclidean distance in the original image space, these samples are with different labels, and thus follow mixmodal distribution characteristics. The graphical description of the supervised learning task upon USPS-MNIST mixmodal data is illustrated by Fig. 5(b).

The proposed  $\alpha$ -SRHLA model and the other compared models are applied alternatively for the mixmodal data. By utilizing the  $k$ -nn ( $k = 5$ ) classifier, each set of experiment is conducted for 20 times to achieve the average recognition performance. The results are reported in Fig. 6(c), which indicates that in most cases, the present  $\alpha$ -SRHLA model outperforms than the others. For example, in the USPS-MNIST experiment, the average recognition error rate achieved by our  $\alpha$ -SRHLA model is 3.5% when the subspace dimensionality is 10, whereas those of the LDA, SLPDA, USSL1, and USSL2 are 6.5%, 5.75%, 6.0%, and 4.9% respectively.

### D. MULTIMODAL AND MIXMODAL FACE RECOGNITION

In this section, to verify the present model in dealing with multimodal and mixmodal face recognition tasks, experiments for both multimodal and mixmodal cases are made. As introduced before, the multimodal and mixmodal properties are frequently encountered in face recognition tasks, such as recognizing face images of different genders, different regions, or different ages. On the other hand, between-class mixmodality might happen when recognizing face images with noises such as wearing glasses, having decorations, etc.

To test the proposed  $\alpha$ -SRHLA model, multimodal and mixmodal face recognition experiments are carried out upon two databases including Yale [35], which is constituted by the face images of 15 persons, and the Extended Yale-B [39] face database that is comprised by 16128 images for 28 persons. For the Extended Yale-B database, all the front-pose images are selected for experiments.

To simulate multimodal face recognition tasks, for the selected database, we generate several multimodal classes by aggregating together the face images for different persons into a single class. For example, for the Extended Yale-B database, we randomly select 16 from the original 28 individuals and combine the images of any two from the selected 16 individuals into a class, generating 8 multimodal classes. Similarly for the Yale database, 5 multimodal classes are established. Upon the generated multimodal classes in the preprocessed database, 50% images will be abandoned in order to guarantee that the size of all classes are the same. For the preprocessed database, various proportions of the face images are selected for training. Then, multimodal and

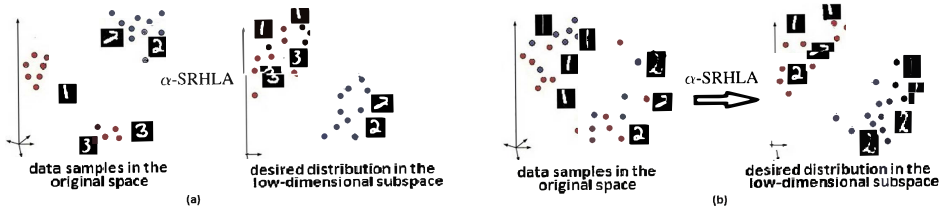


FIGURE 5. The supervised learning procedures of (a) USPS-eo multimodal and (b) USPS-MNIST mixmodal tasks.

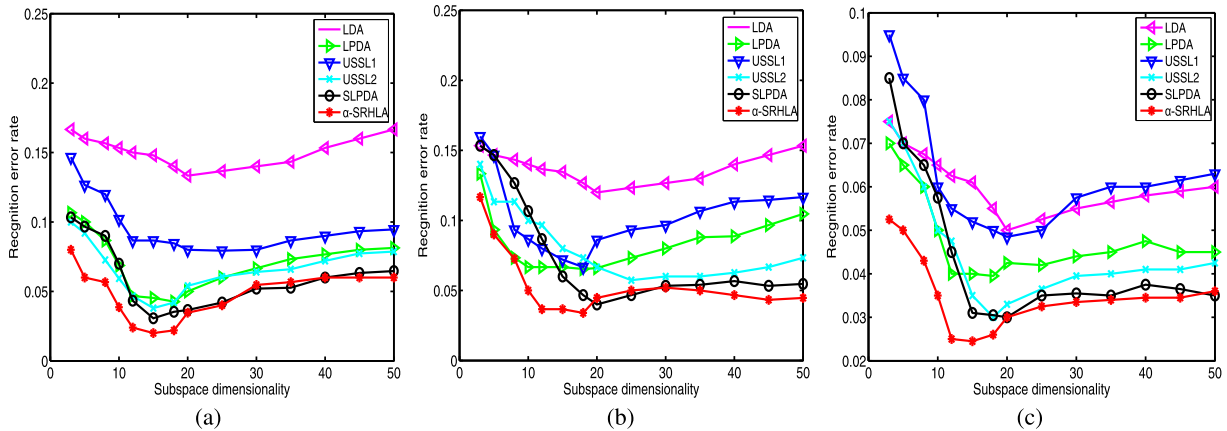


FIGURE 6. Handwriting recognition: Comparisons of the different models for (a) USPS-eo multimodal case, (b) USPS-sl multimodal case, (c) USPS-MNIST mixmodal case.

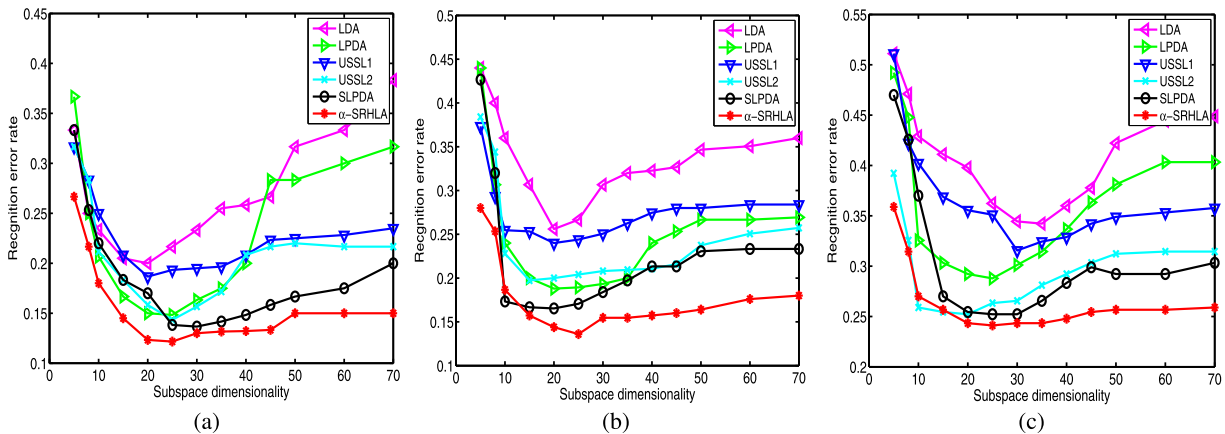


FIGURE 7. Multimodal and mixmodal face recognition for Yale database: Comparisons of different models for (a) 64% training, (b) 55% training, and (c) 45% training scenarios.

mixmodal face recognition experiments are conducted upon the rest of the face images.

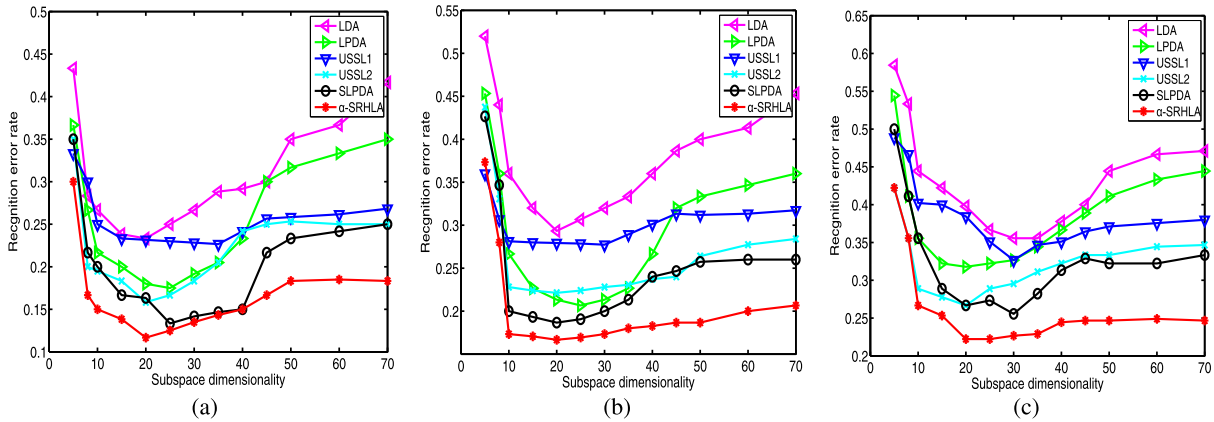
1) FACE RECOGNITION ON THE YALE DATABASE

We first carry out the face recognition experiments upon the preprocessed Yale database, testing the efficiency of the proposed  $\alpha$ -SRHLA model in dealing with multimodal and mixmodal data. Note that the randomly-chosen face images are cropped into  $32 \times 32$  pixels. From the regrouped images, we respectively choose 6, 7, and 8 images of each class for training in a random way and utilize the rest of the images for testing. For each model in each case, the supervised learning process are conducted for 20 times to derive the average

recognition accuracies upon subspace dimensionality of 5 to 50. The  $k$ -nn classifier for  $k = 5$  is adopted to finish the recognition task. For the present  $\alpha$ -SRHLA model, the  $\pm\pi/4$  orientational grouping rule shown in Fig. 4 is applied to induce the “diamond-shaped” sparse patterns. In Fig. 7, the obtained results for the proposed  $\alpha$ -SRHLA model and the other models under different training proportions are shown.

2) FACE RECOGNITION ON THE EXTENDED YALE-B DATABASE

To be specific, 50%, 60%, and 70% images in both multimodal and normal classes are randomly chosen to supervisely



**FIGURE 8. Multimodal and mixmodal face recognition for extended Yale-B database: Comparisons of the different models for (a) 70%, (b) 60%, and (c) 50% training scenarios.**

learn the low-dimensional subspaces. The other images are utilized for testing based on the  $k$ -nn classifier with  $k = 5$ . In each scenario, 20 trails are made to obtain the average results. In the proposed  $\alpha$ -SRHLA model, the  $\pm\pi/4$  orientational grouping rule is adopted. Fig. 8 provides the achieved results of all the models for different subspace dimensionalities.

It can be concluded from Figs. 7 and 8 that the present  $\alpha$ -SRHLA gives more accurate recognition results in most experiments for multimodal and mixmodal face recognition. For example for the Extended Yale-B Database, when the training proportion is 70%, the best average recognition result for the  $\alpha$ -SRHLA model is 88.3%, whereas those corresponding of the LDA, LPDA, USSL1, USSL2, and SLPDA models are 76.7%, 79.2%, 77.3%, 84.2%, 86.7% respectively. This verifies that incorporating the “structural sparsity” with “hierarchical locality” facilitates the achievement of more discriminative subspaces, as well as the improvement of the robustness for the multimodal and mixmodal face recognition tasks.

## V. CONCLUSION

This paper proposes an  $\alpha$ -SRHLA model to learn structurally sparse subspaces upon data with complex distribution properties, i.e., multimodal and mixmodal. The merit of our model, compared with most existing data studies, is incorporating the “hierarchical locality” with “structural sparsity”, which can better extract the hierarchical distribution features of multimodal and mixmodal data in sparse subspace learning scenarios. Experimental evaluations including normal/multimodal data classification as well as multimodal/mixmodal digit and face recognition verify the validity of such incorporation.

The proposed optimization approach considers both data-level and class-level local geometric information, facilitating the maintaining of the local correlations for multimodal samples, as well as the separation of mixmodal ones. In particular, the non-convex  $\alpha$ -structural regularization method is effective in preserving the structural correlations between

data variables and can successfully yield the optimal structurally sparse subspace with desired sparse pattern.

## REFERENCES

- [1] L. Jin, M. Chen, Y. Jiang, and H. Xia, “Multi-traffic scene perception based on supervised learning,” *IEEE Access*, vol. 6, pp. 4287–4296, 2018.
- [2] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, “Where should I walk? Predicting terrain properties from images via self-supervised learning,” *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1509–1516, Apr. 2019.
- [3] J. Liu, C. Li, and W. Yang, “Supervised learning via unsupervised sparse autoencoder,” *IEEE Access*, vol. 6, pp. 73802–73814, 2018.
- [4] S. Zhang, J. Li, M. Jiang, P. Yuan, and B. Zhang, “Scalable discrete supervised multimedia hash learning with clustering,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2716–2729, Oct. 2017.
- [5] L. Zhang, Z. Wu, and J. Cao, “Detecting spammer groups from product reviews: A partially supervised learning model,” *IEEE Access*, vol. 6, pp. 2559–2568, 2017.
- [6] K. Lenc, E. Elsen, T. Schaul, and K. Simonyan, “Non-differentiable supervised learning with evolution strategies and hybrid methods,” 2019, *arXiv:1906.03139*. [Online]. Available: <http://arxiv.org/abs/1906.03139>
- [7] J. Zhang, Y. Zhu, and Z. Chen, “Evolutionary game dynamics of multi-agent systems on multiple community networks,” *IEEE Trans. Syst., Man, Cybern. Syst.*, to be published.
- [8] M. K. Müller, M. Tremer, C. Bodenstein, and R. P. Würtz, “Learning invariant face recognition from examples,” *Neural Netw.*, vol. 41, pp. 137–146, May 2013.
- [9] Q. Zhang and T. Chu, “Learning in multimodal and mixmodal data: Locality preserving discriminant analysis with kernel and sparse representation techniques,” *Multimedia Tools Appl.*, vol. 76, no. 14, pp. 15465–15489, Jul. 2017.
- [10] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” *J. Multivariate Anal.*, vol. 99, no. 6, pp. 1015–1034, Jul. 2008.
- [11] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comp. Rendus Mathématique*, vol. 346, nos. 9–10, pp. 589–592, May 2008.
- [12] Z. Yang, C. Zhang, and L. Xie, “On phase transition of compressed sensing in the complex domain,” *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 47–50, Jan. 2012.
- [13] W. Zhao, Z. Liu, Z. Guan, B. Lin, and D. Cai, “Orthogonal projective sparse coding for image representation,” *Neurocomputing*, vol. 173, pp. 270–277, Jan. 2015.
- [14] H. Zhang, Z. Liu, G.-B. Huang, and Z. Wang, “Novel weighting-delay-based stability criteria for recurrent neural networks with time-varying delay,” *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 91–106, Jan. 2010.
- [15] Y. Huang, Y. Quan, and T. Liu, “Supervised sparse coding with decision forest,” *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 327–331, Feb. 2019.

- [16] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 83–126, 2019.
- [17] X. Pei, J. Zou, and W. Chen, "Graph learning via edge constrained sparse representation for image analysis," *IEEE Access*, vol. 7, pp. 42408–42417, 2019.
- [18] P. Ablin, T. Moreau, M. Massias, and A. Gramfort, "Learning step sizes for unfolded sparse coding," 2019, *arXiv:1905.11071*. [Online]. Available: <http://arxiv.org/abs/1905.11071>
- [19] D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, Oct. 2007, pp. 73–82.
- [20] F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on L1-norm maximization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2065–2074, Nov. 2014.
- [21] Z. Zheng, X. Huang, Z. Chen, X. He, H. Liu, and J. Yang, "Regression analysis of locality preserving projections via sparse penalty," *Inf. Sci.*, vol. 303, pp. 1–14, May 2015.
- [22] Z. Lai, "Sparse local discriminant projections for discriminant knowledge extraction and classification," *IET Comput. Vis.*, vol. 6, no. 6, pp. 551–559, Nov. 2012.
- [23] T. Song, D. Li, Z. Liu, and W. Yang, "Online ADMM-based extreme learning machine for sparse supervised learning," *IEEE Access*, vol. 7, pp. 64533–64544, 2019.
- [24] Q. Zhang, K. Deng, and T. Chu, "Sparsity induced locality preserving projection approaches for dimensionality reduction," *Neurocomputing*, vol. 200, pp. 35–46, Aug. 2016.
- [25] Z. Xu, X. Chang, F. Xu, and H. Zhang, " $\ell_{1/2}$  regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.
- [26] Q. Zhang and T. Chu, "Semi-supervised discriminant analysis based on sparse-coding theory," in *Proc. 35th Chin. Control Conf. (CCC)*, Jul. 2016, pp. 7082–7087.
- [27] T. Zhang, "Multistage convex relaxation for learning with sparse regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1929–1936.
- [28] J. Zhang, Y. Zhu, Q. Li, and Z. Chen, "Promoting cooperation by setting a ceiling payoff for defectors under three-strategy public good games," *Int. J. Syst. Sci.*, vol. 49, no. 10, pp. 2267–2286, Jul. 2018.
- [29] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Feb. 2011.
- [30] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 366–373.
- [31] J. Sang, D. Peng, and Y. Sang, "A general approach for achieving supervised subspace learning in sparse representation," *IEEE Access*, vol. 7, pp. 74017–74028, 2019.
- [32] B. Zu, K. Xia, T. Li, Z. He, Y. Li, J. Hou, and W. Du, "SLIC superpixel-based  $\ell_{2,1}$ -norm robust principal component analysis for hyperspectral image classification," *Sensors*, vol. 19, no. 3, p. 479, 2019.
- [33] D. Wang, X. Zhang, M. Fan, and X. Ye, "Semi-supervised dictionary learning via structural sparse preserving," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2137–2144.
- [34] 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.
- [35] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
- [36] P. Tseng and S. Yun, "A coordinate gradient descent method for non-smooth separable minimization," *Math. Program.*, vol. 117, nos. 1–2, pp. 387–423, Mar. 2009.
- [37] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Ind. Appl. Math. B, Numer. Anal.*, vol. 2, no. 2, pp. 205–224, Jan. 1965.
- [38] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, Jan. 2010.
- [39] A. S. Georghiadis, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.



**QI ZHANG** received the Ph.D. degree in dynamics and control from Peking University, Beijing, China, in 2017. From September 2014 to September 2015, she was a Visiting Ph.D. student with the Yale Institute of Network Science, Yale University. She is currently an Assistant Professor with the School of Information Technology and Management, University of International Business and Economics. Her research interests include machine learning, data mining, and social networks.



**CHAOYI SHI** received the B.S. degree in mechanics and engineering from Zhengzhou University, Zhengzhou, China, in 2011. He is currently pursuing the Ph.D. degree with the College of Engineering, Peking University. His current research interests include machine learning and network science.



**TIANGUANG CHU** received the Ph.D. degree from Tsinghua University, Beijing, China, in 1993. He was a Visiting Research Fellow with The University of Melbourne, in 2001. He is currently a Professor with the College of Engineering, Peking University, Beijing. His research interests include nonlinear dynamics and control, multi-agent systems, evolutionary dynamics, and learning systems.

• • •