Robust Structured Subspace Learning for Data Representation

Zechao Li, Jing Liu, Jinhui Tang, *Senior Member, IEEE*, and Hanqing Lu, *Senior Member, IEEE*

Abstract—To uncover an appropriate latent subspace for data representation, in this paper we propose a novel Robust Structured Subspace Learning (RSSL) algorithm by integrating image understanding and feature learning into a joint learning framework. The learned subspace is adopted as an intermediate space to reduce the semantic gap between the low-level visual features and the high-level semantics. To guarantee the subspace to be compact and discriminative, the intrinsic geometric structure of data, and the local and global structural consistencies over labels are exploited simultaneously in the proposed algorithm. Besides, we adopt the $\ell_{2,1}$-norm for the formulations of loss function and regularization respectively to make our algorithm robust to the outliers and noise. An efficient algorithm is designed to solve the proposed optimization problem. It is noted that the proposed framework is a general one which can leverage several well-known algorithms as special cases and elucidate their intrinsic relationships. To validate the effectiveness of the proposed method, extensive experiments are conducted on diversity datasets for different image understanding tasks, i.e., image tagging, clustering, and classification, and the more encouraging results are achieved compared with some state-of-the-art approaches.

Index Terms—Data representation, latent subspace, image understanding, feature learning, structure preserving

1 Introduction

For many pattern recognition and computer vision problems, images are always represented by a variety of visual features, which are often quite different from each other [1]. The dimension of data feature space is becoming increasingly large. It is inevitable to introduce noisy and/or redundant features. The effectiveness and efficiency of learning methods drop exponentially as the dimensionality increases, which is commonly referred to as the "curse of dimensionality". Therefore, it is a fundamental problem to find a suitable representation of high dimensional data [2], which can enhance the performance of numerous tasks, such as classification and multimedia analysis. To address these problems, a number of different methods have been developed, such as feature selection (i.e., select a subset of most discriminative features from the original features) [3], [4], [5] and subspace learning (i.e., transform the original features to a lower dimensional subspace) [2], [6]. In this paper, we focus on learning an appropriate representation of data by uncovering a latent subspace for the purposes of image understanding, which is referred to assigning proper high-level semantic meaning (labels or tags) to given images (normally represented by low-level visual features), including image tagging, clustering[1] and classification in this paper.

1. For clustering, the cluster indicators of samples can be deemed as the pseudo labels of samples.

- *Z. Li and J. Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P.R. China. E-mail: {zechao.li, jinhuitang}@njust.edu.cn.*
- *J. Liu and H. Lu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: {jliu, luhq}@nlpr.ia.ac.cn.*

Manuscript received 13 June 2014; revised 10 Dec. 2014; accepted 22 Jan. 2015. Date of publication 4 Feb. 2015; date of current version 4 Sept. 2015. Recommended for acceptance by M. Pantic.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2400461

Recent years have witnessed a widespread interest in subspace learning. A variety of subspace learning models and techniques have been widely used for the representation of high dimensional data, such as principal component analysis (PCA) [2], linear discriminant analysis (LDA) [2] and locality preserving projection (LPP) [6]. Despite the different motivations of these algorithms, they can be interpreted in a general graph embedding framework [7]. Some manifold learning algorithms such as ISOMAP [8], Laplacian Eigenmap (LE) [9] and locally linear embedding (LLE) [10] are designed to find a low dimensional subspace in a nonlinear manner. There are many other subspace leaning approaches to find suitable representations of data, such as nonnegative matrix factorization (NMF) [11].

Nonetheless, these methods only focus on low-level features of data, which are independent of the follow-up tasks and ignore the high-level semantic information. As is known to all, there exists the so-called semantic gap between the low-level features and the high-level semantics, which often degrades the performance [12]. To alleviate the semantic gap, we try to uncover proper representations of data by integrating image understanding and feature learning into a joint learning framework. For this purpose, in this paper we propose a novel robust structured subspace learning (RSSL) framework to discover a compact and more informative feature representation and builds a bridge between the low-level features and high-level semantics with the learned subspace by exploiting image understanding, feature learning and feature correlation simultaneously. Specifically, unlike previous subspace learning models, the proposed framework learns a latent discriminative representation of images by considering the image understanding task in the procedure of feature learning, which makes the uncovered representations well predict the labels. To guarantee that the latent subspace is more compact and discriminative, the intrinsic geometric structure of data, and the local and global structural consistencies over labels are

exploited simultaneously and incorporated into the proposed framework. To make our algorithm robust to the outliers and noise, we introduce the $\ell_{2,1}$-norm into the loss function and regularization. To solve the proposed problem, an effective and efficient iterative algorithm is proposed. Finally, we apply the proposed method to the tasks of social image tagging, unsupervised learning (i.e., clustering), semi-supervised and supervised learning (i.e., classification). Extensive experiments are conducted on social image tagging data, face data, handwritten digit data and document data to verify the effectiveness of the proposed framework. Experimental results show that compared with several representative algorithms, the proposed approach achieves outstanding performance for all the tasks.

Our key contributions are summarized as follows.

- We propose a novel data representation learning algorithm by jointly exploiting image understanding, feature learning and feature correlation. Moreover, our framework explores the intrinsic geometric structure, and the local and global structural consistencies over labels to learn a underlying subspace, which is robust to noise and outliers by using $\ell_{2,1}$-norm.
- We develop an efficient algorithm to solve the proposed formulation. The theoretical and empirical analysis demonstrate that the designed optimization algorithm is efficient and effective, and converges quickly.
- The proposed formulation is competent to the tasks of social image tagging, unsupervised clustering and semi-supervised/supervised classification, and achieves state-of-the-art results.
- The proposed framework is a general one that leverages several existing methods as special cases and their intrinsic relationships are elucidated.

The remainder of this paper is arranged as follows. We introduce the related work in Section 2. Then we elaborate our proposed formulation in Section 3 followed with its optimization algorithm in Section 4. How to apply the proposed method to various image understanding tasks is addressed in Section 5. Extensive experiments are conducted and analyzed in Section 6. We present discussions about the proposed method in Section 7. Section 8 concludes this work with future work.

2 Related Work

In this section, we briefly review the related research on feature learning including feature selection [3], [4], [5], [13] and subspace learning [2], [14].

2.1 Feature Selection

Feature selection is a process of obtaining a subset of relevant features for model construction and removing redundant or irrelevant features. According to the availability of label information, there are three broad categories: supervised [4], [13], [15], semi-supervised [16] and unsupervised feature selection methods [5], [17], [18]. Traditional feature selection usually ignores the correlations among features, such as Fisher Score [13] and Laplacian Score (LS) [17]. To this end, some sparsity-based approaches have been studied to exploit the feature correlation [1]. $\ell_{2,1}$-norm has been shown effective for sparse feature selection [4] and gains increasing interest [1], [5], [19]. Ma et al. [1] proposed a supervised feature selection algorithm to improve the image tagging performance. In [5], clustering and feature selection are incorporated into a joint framework to select a feature subset having strong discriminative power. Different from them, the proposed algorithm is an integrated framework which leverages feature learning and image understanding. In addition, several feature selection algorithms can be deemed as special cases of our framework.

2.2 Subspace Learning

Subspace learning sheds light on various tasks in computer vision and multimedia. It projects the original high-dimensional feature space to a low-dimensional subspace, wherein specific statistical properties can be well preserved. The most popular methods include PCA [2], LDA [2], LPP [6] and neighborhood preserving embedding (NPE) [20]. These approaches with different motivations can be interpreted in a general graph embedding framework [7]. The learned projections of these methods are linear combinations of all the original features. Recently, sparse subspace learning has attracted considerable interests. Sparse PCA [21] was proposed based on "Elastic Net" regularization. Cai et al. [22] proposed a unified sparse subspace learning framework based on ℓ_1-norm regularization spectral regression. Some manifold learning approaches are studied to uncover the underlying nonlinear subspace, such as ISOMAP [8], LE [9] and LLE [10]. Factor analysis is another type subspace learning algorithm, such as singular value decomposition (SVD) and NMF [11]. However, these methods only explore the visual features of images to mine the underlying subspace whereas the low-level features and the high-level semantics are not linked. Due to the semantic gap, the learned data representations can not be ensured to well predict labels. In [23], a latent semantic space is uncovered by learning a transformation to link the visual features and tags directly based on low rank approximation. Different from previous work, the proposed method learns a discriminative representation by incorporating image understanding and feature learning into a unified framework. A hidden subspace, which is an intermediate space between the low-level visual space and high-level semantic space, is uncovered to well predict labels. Besides, by introducing the row sparse model, our method is robust to outliers and noise.

3 The Proposed RSSL Framework

In this section, we introduce a novel subspace learning method for image understanding, called robust structured subspace learning, which can find a suitable representation of data.

3.1 Preliminary

Throughout this paper, we use bold uppercase characters to denote matrices, bold lowercase characters to denote vectors. For any matrix $\mathbf{A}$, $\mathbf{a}_i$ means the ith column vector of $\mathbf{A}$, $\mathbf{a}^i$ means the ith row vector of $\mathbf{A}$, A_{ij} denotes the

(i,j)-element of $\mathbf{A}$ and $\mathrm{Tr}[\mathbf{A}]$ is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. $\mathbf{A}^T$ denotes the transposed matrix of $\mathbf{A}$. We define for $q \geq 1$, the ℓ_q-norm of a vector $\mathbf{a} \in \mathbb{R}^m$ as $\|\mathbf{a}\|_q = (\sum_{i=1}^m |a_i|^q)^{1/q}$. We consider the Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$: $\|\mathbf{A}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 = \mathrm{Tr}[\mathbf{A}^{\mathrm{T}}\mathbf{A}]$. The $\ell_{2,1}$-norm for $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} A_{ij}^2} = 2\mathrm{Tr}[\mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{A}], \tag{1}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. Note that in practice, $\|\mathbf{a}^i\|_2$ could be close to zero. For this case, we can follow the traditional regularization way and define $D_{ii} = \frac{1}{\|\mathbf{a}^i\|_2+\epsilon}$, where ϵ is very small constant. When $\epsilon \to 0$, it is easy to verify that $\frac{1}{\|\mathbf{a}^i\|_2+\epsilon}$ approximates $\frac{1}{\|\mathbf{a}^i\|_2}$. Furthermore, let us use $\mathbf{I}_m$ to denote the identity matrix in $\mathbb{R}^{m\times m}$.

Consider a data set consisting of n data points $\{\mathbf{x}_i\}_{i=1}^n$ assigned with c-dimensional binary-valued label vectors $\{\mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$, $\mathbf{y_i} \in \{\mathbf{0},\mathbf{1}\}^{\mathbf{c}}$. Here c is the cardinality of the label set $\mathcal{C} = \{t_1, t_2, \ldots, t_c\}$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ denote the data matrix, in which $\mathbf{x}_i \in \mathbb{R}^d$ is the feature descriptor of the ith sample, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$ be the label matrix of size $n \times c$. The jth column vector of $\mathbf{Y}$ corresponds to a labeling configuration with respect to the tag j and $Y_{ij} = 1$ indicates that $\mathbf{x_i}$ is associated with the label j, and $Y_{ij} = 0$ otherwise. We also introduce a predicted label matrix $\mathbf{F} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$, where each row $\mathbf{f^i} \in \mathbb{R}^{\mathbf{c}}$ is the predicted label vector of the ith data $\mathbf{x_i}$. The local structure graph $\mathbf{S}$ is defined as follows [24], [25]

$$S_{ij} = \begin{cases} \exp\Big(-\frac{\|\mathbf{x_i}-\mathbf{x_j}\|^2}{\sigma^2}\Big) & \mathbf{x_i} \in \mathcal{N}_{\mathbf{k}}(\mathbf{x_j}) \text{ or } \mathbf{x_j} \in \mathcal{N}_{\mathbf{k}}(\mathbf{x_i}) \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{N}_k(\mathbf{x})$ is the set of k-nearest neighbors of $\mathbf{x}$.

3.2 Formulation

Obtaining a good performance in image understanding tasks always requires to find a good data representation. Features are correlated to represent the semantic information and combinations of features are more discriminative than individual features. We present a formulation to learn appropriate representations of images embedding in a latent subspace. In this work, the underlying subspace is expected to satisfy the two following properties.

1) It should be locally smooth, i.e., the local intrinsic geometric structure should be consistent with that in the original visual space.
2) It should be discriminative to well predict the proper labels.

In light of these properties, it is reasonable to assume that the latent subspace and the original space are linked by a linear transformation $\mathbf{Q} \in \mathbb{R}^{\mathbf{d}\times\mathbf{r}}$, where r is the dimensionality of the latent subspace. For each data point $\mathbf{x_i}$, the corresponding representation in the latent subspace is $\mathbf{Q^T x_i}$. To satisfy the first property, we assume that the neighboring data in the original feature space ought to be close to each other in the latent subspace, which is analogous to the Laplace-Beltrami operator on manifolds [26]. This introduces a smooth regularization on the underlying geometric structure between samples in the latent subspace, which is formulated as

$$\min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}\left\|\frac{\mathbf{Q}^T\mathbf{x}_i}{\sqrt{E_{ii}}} - \frac{\mathbf{Q}^T\mathbf{x}_j}{\sqrt{E_{jj}}}\right\|_2^2 = \mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}], \tag{3}$$

where $\mathbf{E}$ is a diagonal matrix with $E_{ii} = \sum_{j=1}^n S_{ij}$ and $\mathbf{L} = \mathbf{E^{-1/2}}(\mathbf{E}-\mathbf{S})\mathbf{E^{-1/2}}$ is the normalized graph Laplacian matrix. Note that the orthogonal constraint $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r$ is imposed to make the problem tractable.

To fulfil the second property, we introduce predictive functions to find a proper predicted label matrix $\mathbf{F} \in \mathbb{R}^{n\times c}$ with the following attributes.

i) The predicted labels are supposed to be locally consistent. That is, the labeling information should be consistent among the nearby points.
ii) The predicted labels should be globally consistent, i.e., they should be consistent with the groundtruth labels.
iii) The predictive functions should be robust to the outliers and noise.

For simplicity, the linear function is adopted to predict the mapping relationship between the latent space and the label space, i.e.,

$$h_j(\mathbf{x}_i) = \mathbf{p}_j^T\mathbf{Q}^T\mathbf{x}_i. \tag{4}$$

Denoting $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_c]$, we obtain

$$h(\mathbf{X}) = \mathbf{P}^T\mathbf{Q}^T\mathbf{X}. \tag{5}$$

The least squares loss function is always used to learn the predictive functions. However, it is very sensitive to outliers and noise. $\ell_{2,1}$-norm has been confirmed to be robust to outliers and noise [1], [4]. Therefore, we propose the following objective function to learn the predictive functions:

$$\min_{\mathbf{P}} \|\mathbf{F} - \mathbf{X}^T\mathbf{Q}\mathbf{P}\|_{2,1} + \lambda\|\mathbf{P}\|_{2,1}. \tag{6}$$

λ is a nonnegative regularization parameter. In the above objective function, the loss function $\|\mathbf{F} - \mathbf{X}^T\mathbf{Q}\mathbf{P}\|_{2,1}$ is robust to outliers and noise. Meanwhile, the regularization term $\|\mathbf{P}\|_{2,1}$ guarantees that $\mathbf{P}$ is sparse in rows, which requires to select discriminative features in the latent subspace to predict $\mathbf{F}$.

The idea of local and global consistency, i.e., the attributes (i) and (ii), can be generalized as follows:

$$\begin{aligned} &\min_{\mathbf{F}} \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}\left\|\frac{\mathbf{f}^i}{\sqrt{E_{ii}}} - \frac{\mathbf{f}^j}{\sqrt{E_{jj}}}\right\|^2 + \sum_{i=1}^{n} U_{ii}\sum_{l=1}^{c}(F_{il}-Y_{il})^2 \\ \Leftrightarrow \quad &\min_{\mathbf{F}} \mathbf{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]. \end{aligned} \tag{7}$$

Here $\mathbf{U}$ is a diagonal matrix defined as

$$U_{ii} = \begin{cases} \zeta & \text{if } \mathbf{x_i} \text{ is tagged;} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

ζ is a large constant. The first term and the second term in the optimization problem (7) guarantee the local and global

structural consistency, respectively. Note that $\mathbf{F}$ is the predicted label matrix. It is natural and reasonable to impose a nonnegative constraint on $\mathbf{F}$, i.e., all the elements of $\mathbf{F}$ are required to be nonnegative. Consequently, the optimization problem (7) becomes:

$$\min_{\mathbf{F}\geq 0} \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]. \tag{9}$$

By jointly modelling (3), (6) and (9), we obtain

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Q},\mathbf{F}} &\|\mathbf{F}-\mathbf{X}^T\mathbf{Q}\mathbf{P}\|_{2,1} + \alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] \\ &+ \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]) + \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\ &+ \lambda\|\mathbf{P}\|_{2,1} \\ \text{s.t.}\quad &\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r, \mathbf{F}\geq 0,\end{aligned} \tag{10}$$

where α and β are two trade-off parameters. From the above objective function, we can see that the predictive functions are robust to outliers and noise and can preserve the local and global structural consistency. The third term can avoid the overfitting problem induced by sparse context links $\mathbf{Y}$, and also incorporate the content links into modeling the latent space geometry [23]. By jointly learning the predictive functions and the latent subspace with $\ell_{2,1}$-norm regularization, the proposed formulation ensures that features in the underlying subspace are combinations of original features and reflect semantic information. Thus, they are discriminative to predict labels.

4 Optimization

The optimization problem (10) involves the $\ell_{2,1}$-norm which is non-smooth and cannot have a close form solution. Consequently, we propose an iterative algorithm. To facilitate the optimization, by defining $\mathbf{W}=\mathbf{Q}\mathbf{P}\in\mathbb{R}^{d\times c}$, we rewrite the problem (10) as minimizing the following equation.

$$\begin{aligned}\mathcal{O} = &\|\mathbf{F}-\mathbf{X}^T\mathbf{W}\|_{2,1} + \alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] \\ &+ \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]) + \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\ &+ \gamma\|\mathbf{W}-\mathbf{Q}\mathbf{P}\|_F^2 + \lambda\|\mathbf{P}\|_{2,1} \\ \text{s.t.}\quad &\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r, \mathbf{F}\geq 0.\end{aligned} \tag{11}$$

In the following, we introduce the proposed update rules in brief and the elaborated inference procedure please refer to the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2015.2400461.

4.1 Update $\mathbf{P}$ As Given $\mathbf{W}$ and $\mathbf{Q}$

By setting the derivative $\partial\mathcal{O}/\partial\mathbf{P}=0$ and using the property that $\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r$, we obtain

$$\begin{aligned}&\gamma(\mathbf{Q}^T\mathbf{Q}\mathbf{P}-\mathbf{Q}^T\mathbf{W}) + \lambda\mathbf{D}\mathbf{P} = 0 \\ \Rightarrow\ &\mathbf{P} = \gamma(\gamma\mathbf{I}+\lambda\mathbf{D})^{-1}\mathbf{Q}^T\mathbf{W} = \gamma\mathbf{V}^{-1}\mathbf{Q}^T\mathbf{W}.\end{aligned} \tag{12}$$

where $\mathbf{V}=\gamma\mathbf{I}_r+\lambda\mathbf{D}$ and $\mathbf{D}$ is a diagonal matrix with $D_{ii}=\frac{1}{2\|\mathbf{p}^i\|_2}$, $i=1,\ldots,r$.

4.2 Update $\mathbf{W}$ As Given $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{F}$

Now, by substituting $\mathbf{P}$ in $\mathcal{O}$ with Eq. (12), the objective function $\mathcal{O}$ is written as follows:

$$\begin{aligned}\mathcal{O} &= \|\mathbf{Z}\|_{2,1} + f(\mathbf{F}) + g(\mathbf{Q}) + \gamma\|\mathbf{W}-\gamma\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T\mathbf{W}\|_F^2 \\ &\quad + \lambda\mathrm{Tr}[\gamma^2\mathbf{W}^T\mathbf{Q}\mathbf{V}^{-1}\mathbf{D}\mathbf{V}^{-1}\mathbf{Q}^T\mathbf{W}] \\ &= f(\mathbf{F}) + g(\mathbf{Q}) + \mathrm{Tr}[(\mathbf{F}-\mathbf{X}^T\mathbf{W})^T\bar{\mathbf{D}}(\mathbf{F}-\mathbf{X}^T\mathbf{W})] \\ &\quad + \gamma\mathrm{Tr}[\mathbf{W}^T(\mathbf{I}_d-\gamma\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T)\mathbf{W}],\end{aligned} \tag{13}$$

where $\mathbf{Z}=\mathbf{F}-\mathbf{X}^T\mathbf{W}$, $f(\mathbf{F})=\alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}]+\mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})])$ and $g(\mathbf{Q})=\beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}]$. $\bar{\mathbf{D}}$ is a diagonal matrix with $\bar{D}_{ii}=\frac{1}{2\|\mathbf{z}^i\|_2}$ and we use the property that $\|\mathbf{A}\|_F^2=\mathrm{Tr}(\mathbf{A}^T\mathbf{A})$ for any arbitrary matrix $\mathbf{A}$. By setting the derivative $\partial\mathcal{O}/\partial\mathbf{W}=0$, we get

$$\begin{aligned}&\mathbf{X}\bar{\mathbf{D}}(\mathbf{X}^T\mathbf{W}-\mathbf{F}) + \gamma(\mathbf{I}_d-\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T)\mathbf{W} = 0 \\ \Leftrightarrow\quad &(\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T + \gamma(\mathbf{I}_d-\gamma\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T))\mathbf{W} = \mathbf{X}\bar{\mathbf{D}}\mathbf{F} \\ \Leftrightarrow\quad &\mathbf{W} = (\mathbf{G}-\gamma^2\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T))^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F} \\ \Leftrightarrow\quad &\mathbf{W} = \mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F}.\end{aligned} \tag{14}$$

Here $\mathbf{G}=\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T+\gamma\mathbf{I}_d$ and $\mathbf{H}=\mathbf{G}-\gamma^2\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T$.

4.3 Update $\mathbf{Q}$ As Given $\mathbf{P}$, $\mathbf{W}$ and $\mathbf{F}$

First, we rewrite Eq. (13) as follows:

$$\begin{aligned}\mathcal{O} &= \mathrm{Tr}[(\mathbf{X}^T\mathbf{W}-\mathbf{F})^T\bar{\mathbf{D}}(\mathbf{X}^T\mathbf{W}-\mathbf{F})] + f(\mathbf{F}) \\ &\quad + g(\mathbf{Q}) + \gamma\mathrm{Tr}[\mathbf{W}^T(\mathbf{I}_d-\gamma\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T)\mathbf{W}] \\ &= \mathrm{Tr}[\mathbf{W}^T\mathbf{H}\mathbf{W}] - 2\mathrm{Tr}[\mathbf{W}^T\mathbf{X}\bar{\mathbf{D}}\mathbf{F}] \\ &\quad + \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{F}] + f(\mathbf{F}) + g(\mathbf{Q}).\end{aligned} \tag{15}$$

By substituting the expression for $\mathbf{W}$ in Eq. (14) into Eq. (15), since $\mathbf{H}=\mathbf{H}^T$, we obtain the following equation:

$$\begin{aligned}\mathcal{O} &= \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{X}^T(\mathbf{H}^{-1}\mathbf{H}\mathbf{H}^{-1}-2\mathbf{H}^{-1})\mathbf{X}\bar{\mathbf{D}}\mathbf{F}] \\ &\quad + \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{F}] + f(\mathbf{F}) + g(\mathbf{Q}) \\ &= -\mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F}] \\ &\quad + \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{F}] + f(\mathbf{F}) + g(\mathbf{Q}).\end{aligned} \tag{16}$$

By substituting Eq. (16) into the problem (11), we have the following optimization problem *w.r.t.* $\mathbf{Q}$:

$$\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F}] - g(\mathbf{Q}). \tag{17}$$

To compute the matrix inverse, using the Sherman-Morrison-Woodbury formula [27]: $(\mathbf{A}+\mathbf{U}\mathbf{C}\mathbf{V})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{U}\big(\mathbf{C}^{-1}+\mathbf{V}\mathbf{A}^{-1}\mathbf{U}\big)^{-1}\mathbf{V}\mathbf{A}^{-1}$, we have

$$\begin{aligned}\mathbf{H}^{-1} &= (\mathbf{G}-\gamma^2\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}^T)^{-1} \\ &= \mathbf{G}^{-1} + \gamma^2\mathbf{G}^{-1}\mathbf{Q}(\mathbf{V}-\gamma^2\mathbf{Q}^T\mathbf{G}^{-1}\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{G}^{-1}.\end{aligned} \tag{18}$$

Thus, by using the property that $\mathrm{Tr}[\mathbf{A}\mathbf{B}]=\mathrm{Tr}[\mathbf{B}\mathbf{A}]$, the optimization problem (17) is equivalent to

$$\begin{aligned}
&\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \mathrm{Tr}[\mathbf{F}^T\bar{\mathbf{D}}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F}] - \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\
\Leftrightarrow &\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \gamma^2\mathrm{Tr}[(\gamma\mathbf{I}_r + \lambda\mathbf{D} - \gamma^2\mathbf{Q}^T\mathbf{G}^{-1}\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{T}\mathbf{Q}] \\
&\qquad - \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\
\Leftrightarrow &\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \gamma^2\mathrm{Tr}[\mathbf{Q}^T(\gamma\mathbf{I}_d + \lambda\mathbf{Q}\mathbf{D}\mathbf{Q}^T - \gamma^2\mathbf{G}^{-1})^{-1}\mathbf{T}\mathbf{Q}] \\
&\qquad - \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\
\Leftrightarrow &\max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_r} \mathrm{Tr}[\mathbf{Q}^T\mathbf{N}^{-1}\mathbf{Q}],
\end{aligned} \tag{19}$$

where $\mathbf{T} = \mathbf{G}^{-1}\mathbf{X}\bar{\mathbf{D}}\mathbf{F}\mathbf{F}^T\bar{\mathbf{D}}\mathbf{X}^T\mathbf{G}^{-1}$ and $\mathbf{N} = (\frac{1}{\gamma}\mathbf{I}_d + \frac{\lambda}{\gamma^2}\mathbf{Q}\mathbf{D}\mathbf{Q}^T - \mathbf{G}^{-1})^{-1}\mathbf{T} - \beta\mathbf{X}\mathbf{L}\mathbf{X}^T$. $\mathbf{Q}$ can be relaxedly obtained by the eigen-decomposition of $\mathbf{N}^{-1}$. Note that although $\mathbf{N}$ needs $\mathbf{Q}$ as input, the above solution is effective and empirically validated since our algorithm converges very quickly to make $\mathbf{Q}$ stable.

Algorithm 1. The Proposed RSSL Algorithm

Input:
Data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and Tag matrix $\mathbf{Y} \in \mathbb{R}^{n\times c}$;
Parameters α, β, γ, λ, k and r

Output:
Converged $\mathbf{F}$, $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{W}$.

Construct the k-nn graph and calculate $\mathbf{L}$;
Compute the selection matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$;
Set $t = 0$; Initialize $\mathbf{F}_0 \in \mathbb{R}^{n\times c}$ and $\mathbf{Q}_0 \in \mathbb{R}^{d\times r}$, and set $\bar{\mathbf{D}}_0 \in \mathbb{R}^{n\times n}$ and $\mathbf{D}_0 \in \mathbb{R}^{r\times r}$ as identity matrices;
repeat
  $\mathbf{G}_t = \mathbf{X}\bar{\mathbf{D}}_t\mathbf{X}^T + \gamma\mathbf{I}_d$;
  $\mathbf{T}_t = \mathbf{G}_t^{-1}\mathbf{X}\bar{\mathbf{D}}_t\mathbf{F}_t\mathbf{F}_t^T\bar{\mathbf{D}}_t\mathbf{X}^T\mathbf{G}_t^{-1}$;
  $\mathbf{N}_t = (\frac{1}{\gamma}\mathbf{I}_d + \frac{\lambda}{\gamma^2}\mathbf{Q}_t\mathbf{D}_t\mathbf{Q}_t^T - \mathbf{G}_t^{-1})^{-1}\mathbf{T}_t - \beta\mathbf{X}\mathbf{L}\mathbf{X}^T$;
  Get $\mathbf{Q}_{t+1}$ by the eigen-decomposition of $\mathbf{N}_t^{-1}$;
  $\mathbf{H}_t = \mathbf{G}_t - \gamma^2\mathbf{Q}_{t+1}(\gamma\mathbf{I}_r + \lambda\mathbf{D}_t)^{-1}\mathbf{Q}_{t+1}^T$;
  $\mathbf{M}_t = \bar{\mathbf{D}}_t + \alpha\mathbf{L} - \bar{\mathbf{D}}_t\mathbf{X}^T\mathbf{H}_t^{-1}\mathbf{X}\bar{\mathbf{D}}_t$;
  $(\mathbf{F}_{t+1})_{ij} = (\mathbf{F}_t)_{ij}\frac{(\alpha\mathbf{U}\mathbf{Y}_t)_{ij}}{(\mathbf{M}_t\mathbf{F}_t+\alpha\mathbf{U}\mathbf{F}_t)_{ij}}$;
  $\mathbf{W}_{t+1} = \mathbf{H}_t^{-1}\mathbf{X}\bar{\mathbf{D}}_t\mathbf{F}_{t+1}$;
  $\mathbf{Z}_{t+1} = \mathbf{F}_{t+1} - \mathbf{X}^T\mathbf{W}_{t+1}$;
  $$\bar{\mathbf{D}}_{t+1} = \begin{bmatrix} 2\|\mathbf{z}_{t+1}^1\|_2 & & \\ & \cdots & \\ & & \frac{1}{2\|\mathbf{z}_{t+1}^n\|_2} \end{bmatrix};$$
  $\mathbf{P}_{t+1} = \gamma(\gamma\mathbf{I}_r + \lambda\mathbf{D}_t)^{-1}\mathbf{Q}_{t+1}^T\mathbf{W}_{t+1}$;
  $$\mathbf{D}_{t+1} = \begin{bmatrix} 2\|\mathbf{p}_{t+1}^1\|_2 & & \\ & \cdots & \\ & & \frac{1}{2\|\mathbf{p}_{t+1}^r\|_2} \end{bmatrix};$$
  t = t + 1;
until Convergence criterion satisfied

4.4 Update $\mathbf{F}$ As Given $\mathbf{P}$, $\mathbf{W}$ and $\mathbf{Q}$

We substitute the expression in Eq. (16) into Eq. (11) and get the following optimization problem *w.s.t.* $\mathbf{F}$.

$$\begin{aligned}
&\min_{\mathbf{F}\geq 0} \mathrm{Tr}[\mathbf{F}^T(\bar{\mathbf{D}} + \alpha\mathbf{L} - \bar{\mathbf{D}}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}})\mathbf{F}] \\
&\quad + \alpha\mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})].
\end{aligned} \tag{20}$$

Letting $\mathbf{M} = \bar{\mathbf{D}} + \alpha\mathbf{L} - \bar{\mathbf{D}}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\bar{\mathbf{D}}$, ϕ_{ij} be the Lagrange multiplier for constraint $F_{ij} \geq 0$ and $\Phi = [\phi_{ij}]$, the Lagrange function is

$$\mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \alpha\mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})] + \mathrm{Tr}(\Phi\mathbf{F}^T). \tag{21}$$

Setting its derivative with respect to F_{ij} to 0 and using the Karush-Kuhn-Tuckre (KKT) condition [28] $\phi_{ij}F_{ij} = 0$, we obtain the updating rules:

$$\begin{aligned}
&\mathbf{M}\mathbf{F} + \alpha\mathbf{U}(\mathbf{F}-\mathbf{Y}) + \Phi = 0 \\
\Rightarrow \quad &F_{ij} \leftarrow F_{ij}\frac{(\alpha\mathbf{U}\mathbf{Y})_{ij}}{(\mathbf{M}\mathbf{F}+\alpha\mathbf{U}\mathbf{F})_{ij}}.
\end{aligned} \tag{22}$$

From the above analysis, we can see that $\mathbf{D}$ and $\bar{\mathbf{D}}$ related to $\mathbf{W}$ is required to solve $\mathbf{Q}$ and $\mathbf{F}$ and it is still not straightforward to obtain $\mathbf{W}$, $\mathbf{Q}$ and $\mathbf{F}$. To this end, we design an iterative algorithm to solve the proposed formulation, which is summarized in Algorithm 1. The convergence criterion used in our experiments is that the number of iterations is more than 20 or $|\mathcal{O}_{t-1} - \mathcal{O}_t|/\mathcal{O}_{t-1} < 0.001$, where $\mathcal{O}_t$ is the value of the objective function in the tth iteration. Once $\mathbf{Q}$ and $\mathbf{W}$ are obtained, given a testing image $\mathbf{x}$, its latent representation and label prediction vector are computed by $\mathbf{b} = \mathbf{Q}^T\mathbf{x}$ and $\mathbf{f} = [f_1, \ldots, f_c]^T = \mathbf{W}^T\mathbf{x}$, respectively. Besides, $\mathbf{P}$ can be deemed as a feature selection matrix in the latent subspace.

4.5 Computational Complexity Analysis

In this section, we discuss the computational cost of the proposed method. The common way to express the complexity of one algorithm is using big O notation

As stated in Algorithm 1, the k-nn graph is first constructed based on the euclidean distance in the original space. The corresponding cost is $O(dn^2)$, where n is the number of images and d is the dimension of features. Then, the proposed optimization problem is solved iteratively. In each iteration, $\mathbf{G}$ is computed with the cost of $O(dn^2 + d^2n)$ while the complexity to obtain $\mathbf{G}^{-1}$ is $O(d^3)$. The cost for computing $\mathbf{Q}$ is $O(d^3 + d^2n + dn^2 + dnc)$, in which the time complexities to compute $\mathbf{T}$, $\mathbf{N}$ and the eigen-decomposition are $O(d^2n + dn^2 + dnc + d^3)$, $O(d^3 + d^2n + dn^2)$ and $O(d^3)$, respectively, where c is the number of tags. Since $\gamma\mathbf{I}_r + \lambda\mathbf{D}_t$ is a diagonal matrix and its inversion cost $O(r)$, it needs $O(dr^2 + d^2r)$ to obtain $\mathbf{H}$, where r is the dimension of the subspace. Then $\mathbf{H}^{-1}$ is computed with the cost of $O(d^3)$. With $\mathbf{H}^{-1}$ obtained, it costs $O(d^2n + dn^2)$, $O(n^2c)$ and $O(d^2n + dn + dnc)$ to calculate $\mathbf{M}$, $\mathbf{F}$ and $\mathbf{W}$, respectively. Finally, $\mathbf{Z}$, $\bar{\mathbf{D}}$, $\mathbf{P}$ and $\mathbf{D}$ are got with the time complexity of $O(dnc)$, $O(nc + n)$, $O(r + dr^2 + drc)$ and $O(rc + r)$, respectively. Thus the total cost of our method is $O(T(d^2n + dn^2 + d^3 + dnc + dr^2 + d^2r + n^2c + drc) + n^2d)$, where T is the number of iterations in our algorithm. Since $r \ll n$ and $r \ll d$, the whole cost of the proposed method is $O(T(dn^2 + d^2n + d^3 + dnc + n^2c) + dn^2)$.

5 Image Understanding Tasks

We now elaborate how to apply the proposed learning framework to different image understanding tasks, i.e., social image tagging, clustering and classification.

5.1 Social Image Tagging

For social image tagging, the raw correspondences between social images and their associated tags are available to define $\mathbf{Y}$, but it is possibly imprecise since the community contributed tags annotated by web users could be noisy, irrelevant, and often incomplete for describing the image contents [29], [30]. As a consequence, the goal of social image tagging is to remove noisy and irrelevant tags, complement relevant tags, and add tags to untagged images. It is observed from the aforementioned analysis that the proposed formulation is adaptive to the social image tagging task in nature. The $\ell_{2,1}$ norm in the loss function makes it enable to refine the raw tags. That is, for the learning data, *i.e.* in-sample, the learned $\mathbf{F}$ is used to refine the raw tags. For the new image $\mathbf{x}$, *i.e.*, out-of-sample, its tag vector is computed by $\mathbf{f} = \mathbf{W}^T\mathbf{x}$.

5.2 Clustering

The proposed formulation can be applied to clustering. It enables to find an appropriate representation of data to improve the performance of clustering.

For unsupervised learning, all the diagonal elements of the selection matrix $\mathbf{U}$ are all zeros since there exists no labeled data. The term $\mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]$ in the problem (11) is equal to 0. In the clustering task, each sample is assigned to one cluster. The matrix $\mathbf{F}$ is deemed to be the scaled cluster indicator matrix and required to be orthogonal [5], [19]. Consequently, the problem (11) becomes

$$\begin{aligned} \min & \|\mathbf{Z}\|_{2,1} + \alpha\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] \\ & + \gamma\|\mathbf{W}-\mathbf{Q}\mathbf{P}\|_F^2 + \lambda\|\mathbf{P}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \tag{23}$$

We can see that it is nature and reasonable to impose orthogonal constraint on $\mathbf{F}$ for clustering, which makes the learned $\mathbf{F}$ more accurate. When both nonnegative and orthogonal constraints are satisfied, there is only one element in each row of $\mathbf{F}$ is greater than zero and all of the others are zeros. The solutions of $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{W}$ are consistent with those in Section 4. To obtain the solution of $\mathbf{F}$, we relax the orthogonal constraint and the Lagrange function (21) becomes

$$\min_{\mathbf{F}\geq 0} \mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\eta}{2}\left\|\mathbf{F}^T\mathbf{F}-\mathbf{I}_c\right\|_F^2 + \mathrm{Tr}(\Phi\mathbf{F}^T). \tag{24}$$

$\eta > 0$ is a parameter to control the orthogonality condition. In practice, η should be large enough to insure the orthogonality satisfied. Setting its derivative with respect to F_{ij} to 0 and using the KKT condition [28] $\phi_{ij}F_{ij} = 0$, we obtain the updating rules:

$$\begin{aligned} & \mathbf{M}\mathbf{F} + \eta\mathbf{F}(\mathbf{F}^T\mathbf{F}-\mathbf{I}_c) + \Phi = 0 \\ \Rightarrow \quad & F_{ij} \leftarrow F_{ij}\frac{(\eta\mathbf{F})_{ij}}{(\mathbf{M}\mathbf{F}+\eta\mathbf{F}\mathbf{F}^T\mathbf{F})_{ij}}. \end{aligned} \tag{25}$$

Then we normalize $\mathbf{F}$ with $(\mathbf{F}^T\mathbf{F})_{ii} = 1, i = 1, \ldots, c$. The clustering results are obtained from the learned $\mathbf{F}$. In addition, the proposed method can be treated as an unsupervised feature selection method. We first map data into the latent subspace by $\mathbf{Q}$ and then select features in the subspace using $\mathbf{P}$.

5.3 Classification

Similar to the above analysis, the proposed formulation is also adaptive to semi-supervised and supervised classification. For semi-supervised and supervised classification, we still use the definition of the selection matrix $\mathbf{U}$ in Eq. (8), that is, U_{ii} is set to a large enough constant if $\mathbf{x}_i$ is labeled. The changes in the extension to clustering are still applied to classification. That is, we impose orthogonal constraint on $\mathbf{F}$. As a consequence, we obtain the following problem:

$$\begin{aligned} \min & \|\mathbf{Z}\|_{2,1} + \alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]) \\ & + \beta\mathrm{Tr}[\mathbf{Q}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{Q}] + \gamma\|\mathbf{W}-\mathbf{Q}\mathbf{P}\|_F^2 + \lambda\|\mathbf{P}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \tag{26}$$

Note that for consistency to $\mathbf{F}$, $\mathbf{Y}$ is made orthogonal by $\mathbf{Y} \leftarrow \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$.

The solutions of $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{W}$ are consistent with those in Section 4. To obtain the solution of $\mathbf{F}$, by relaxing the orthogonal constraint and using the Lagrange function and the KKT condition, we obtain

$$F_{ij} \leftarrow F_{ij}\frac{(\eta\mathbf{F}+\alpha\mathbf{U}\mathbf{Y})_{ij}}{(\mathbf{M}\mathbf{F}+\alpha\mathbf{U}\mathbf{F}+\eta\mathbf{F}\mathbf{F}^T\mathbf{F})_{ij}}. \tag{27}$$

Note that in semi-supervised classification, the unlabeled learning data are labeled by the learned $\mathbf{F}$. The testing data are labeled using the learned matrix $\mathbf{W}$ in semi-supervised and supervised classification.

6 Experimental Validation

We present extensive experiments to validate the effectiveness of our method for image understanding tasks, including image tagging, clustering and classification. The experiments are discussed in details in terms of image tagging and briefly analyzed for clustering and classification. Statistical significance test is also performed with a significance level of 0.05. The student t-test is employed in our experiments.

6.1 Image Tagging

6.1.1 Data Set

Social photo sharing sites allow users to post, tag and comment on images. Therefore, social images cover almost all the concepts people are interested in, which makes researchers collect images to build social image data sets for experimental purpose. In this work, we conduct our experiments on two large scale publicly-available social image data sets: MIRFlickr [31] and NUS-WIDE [32]. Table 1 summarizes some statistics of these data sets.

MIRFlickr [31]. It contains 25,000 images from Flickr. It contains 1,386 tags and provides the ground-truth annotation of 38 concepts. We kept the tags that appear at least 50 times, resulting in a vocabulary of 457 tags, which only contains 18 concepts of those 38 concepts. Thus, we adopt these 18 concepts to validate the performance. We adopt two types of global image descriptors: Gist features and color histograms with 16 bins in each color channel for LAB and HSV representations and one type of local feature:

TABLE 1
Statistics of the Data Sets with Image and Tag Counts in the Format Mean/Maximum

	MIRFlickr	NUS-WIDE
Tag size	457	2,892
Concept size	18	81
Image size	25,000	55,615
Tags per img.	2.7/45	9.4/199
Concepts per img.	4.7/17	4.2/13
Img. per tag	145.4/1,483	180.9/9,208
Img. per concept	3,102.8/10,373	2,891.5/38,098

SIFT feature. The features are available at http://lear.inrialpes.fr/data/.

NUS-WIDE [32]. It contains $55{,}615$ images from Flickr associated with $5{,}018$ tags annotated by amateur users. The data set provides the ground-truth annotations of 81 concepts, which are used to evaluate the performance. Note that these 81 concepts are different from the user tags with much irrelevant noise information while the ground-truth labels are manually labeled. To reduce too noisy tags, we removed tags whose occurrence numbers are below 25 and obtained $2{,}892$ unique tags. We download five types of features: 144D color correlation, 73D edge direction histogram, 128D wavelet texture, 64D color histogram and 225D block-wise color moments.

6.1.2 Experimental Setting

In our experiments, data are partitioned into two groups: the learning data and the testing data. The learning data is used for model estimation and evaluate the performance of noisy tagged data while the testing data is utilized to test the performance of new data. We randomly select n samples as learning data and the remaining samples are used as testing data. In our experiments, we set $n = 5{,}000$, $n = 10{,}000$, respectively, and report all of the results. During the partition process, each label is guaranteed to be associated with at least one images. To alleviate the instability introduced by the randomly selected training data, we independently repeat experiments 10 times to generate different learning and testing data, and report the average results. The results on the noisy tagged learning data and the testing data are both reported.

6.1.3 Compared Scheme

To validate the effectiveness of RSSL, we compare it with one baseline and a number of related state-of-the-art approaches, which are enumerated as follows. The parameters of these methods are tuned within the candidate set $[10^{-6}, 10^{-3}, 1, 10^{3}, 10^{6}]$.

1) *Baseline*. The rigid regression is utilized as the baseline algorithm.
2) *ASO* [33]. It learns predictive structures from multiple tasks and unlabeled data.
3) *LapRLS* [34]. With the manifold assumption, it uses the least square loss to seek a decision function which is smooth over the whole data distribution according to the graph Laplacian.
4) *SDA* [35]. It reduces the dimension of the input visual features and then rigid regression is performed as a classifier.
5) *MPMF* [36]. Multiple correlations are jointly exploited by multi-correlation probabilistic matrix factorization algorithm for image annotation.
6) *SSLF* [37]. It discovers the correlation information among multiple labels by a low-dimensional subspace learning framework.
7) *LSCCA* [38]. The least-squares formulation of canonical correlation analysis is used to predict labels for samples.
8) *SFSS* [39]. It predicts labels by considering both label consistence with the training data labels and manifold fitting on the data structure.
9) *SFUS* [1]. It annotates images by uncovering the shared subspace of original features based on a sparsity-based model.
10) *LMGE* [40]. Images are annotated by integrating shared structure learning and graph-based learning into a joint framework.
11) *C2MR* [23]. The underlying latent semantic space is learned by mining both context and content links in social media networks.

6.1.4 Evaluation Metric

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is currently considered to be the standard method for model comparison. It is well known that it is a more faithful criterion used in many applications. Following [37], [40], in our experiments we adopt area under cures (AUC) as evaluation metric. As done in [40], both the microaveraging and macroaveraging measures are utilized to evaluate both the global performance across multiple concepts and the average performance of all the concepts. To calculate the microaveraging result, we first concatenate the concept indicator vectors of all concepts as a long vector and then compute the average AUC. For the macroaveraging result, we first compute the mean AUC of each concept and then average the mean AUC values of all the concepts.

Besides, the performance for image tagging is also evaluated by using F1 measure, which is defined as: $\mathrm{F1} = \frac{2 \times \mathbf{R} \times \mathbf{P}}{\mathbf{R}+\mathbf{P}}$, where $\boldsymbol{R} = \frac{|N_c|}{|N_g|}$, and $\boldsymbol{P} = \frac{|N_c|}{|N_t|}$. Here $|N_g|$ be the number of images tagged with one concept w in the ground truth, $|N_t|$ be the number of images tagged with w of our algorithm, and $|N_c|$ be the number of correct tagged images with w by our algorithm. The mean F1 over concepts is presented. Note that in our experiments we annotate each image with five concepts. Furthermore, mean average precision (MAP) is utilized to measure the performance for image retrieval.

6.1.5 Experimental Results

The mean MicroAUC, mean MacroAUC and mean F1 of 10 times independent experiments with standard deviation of all the algorithms on the MIRFlickr and NUS-WIDE data sets are presented in Tables 2 and 3, respectively. Results that are significantly better than others are indicated in boldface. From the results, we have the following observations.

TABLE 2
Experimental Results (Mean Microauc ± Standard Deviation, Mean Macroauc ± Standard Deviation and Mean F1 ± Standard Deviation) on the MIRFlickr Data Set

Method		$n = 5{,}000$			$n = 10{,}000$		
		MicroAUC	MacroAUC	F1	MicroAUC	MacroAUC	F1
Baseline	learn	0.555 ± 0.001	0.544 ± 0.001	0.195 ± 0.002	0.582 ± 0.003	0.588 ± 0.004	0.230 ± 0.004
ASO	learn	0.590 ± 0.004	0.579 ± 0.002	0.229 ± 0.004	0.634 ± 0.005	0.613 ± 0.003	0.231 ± 0.002
LapRLS	learn	0.578 ± 0.003	0.566 ± 0.003	0.211 ± 0.003	0.598 ± 0.005	0.590 ± 0.003	0.242 ± 0.006
SDA	learn	0.561 ± 0.005	0.553 ± 0.006	0.204 ± 0.006	0.571 ± 0.004	0.550 ± 0.001	0.201 ± 0.003
MPMF	learn	0.623 ± 0.003	0.624 ± 0.007	0.289 ± 0.008	0.592 ± 0.001	0.631 ± 0.001	0.292 ± 0.001
SSLF	learn	0.667 ± 0.004	0.638 ± 0.003	0.369 ± 0.004	0.685 ± 0.002	0.649 ± 0.001	0.297 ± 0.002
LSCCA	learn	0.574 ± 0.007	0.553 ± 0.010	0.247 ± 0.018	0.586 ± 0.006	0.566 ± 0.004	0.259 ± 0.003
SFSS	learn	0.676 ± 0.005	0.658 ± 0.007	0.370 ± 0.014	0.708 ± 0.001	0.668 ± 0.001	0.380 ± 0.004
SFUS	learn	0.677 ± 0.004	0.655 ± 0.003	0.369 ± 0.004	0.701 ± 0.003	0.674 ± 0.003	0.380 ± 0.003
LGME	learn	0.688 ± 0.004	0.658 ± 0.005	0.369 ± 0.005	0.703 ± 0.002	0.673 ± 0.006	0.381 ± 0.002
C2MR	learn	0.636 ± 0.002	0.616 ± 0.003	0.261 ± 0.002	0.671 ± 0.014	0.647 ± 0.005	0.286 ± 0.007
RSSL	learn	$\mathbf{0.703 \pm 0.001}$	$\mathbf{0.675 \pm 0.002}$	$\mathbf{0.498 \pm 0.006}$	$\mathbf{0.724 \pm 0.002}$	$\mathbf{0.685 \pm 0.002}$	$\mathbf{0.512 \pm 0.001}$
Baseline	test	0.524 ± 0.002	0.511 ± 0.003	0.162 ± 0.003	0.550 ± 0.003	0.536 ± 0.001	0.185 ± 0.001
ASO	test	0.547 ± 0.003	0.5311 ± 0.002	0.188 ± 0.002	0.566 ± 0.005	0.562 ± 0.003	0.194 ± 0.002
LapRLS	test	0.561 ± 0.002	0.540 ± 0.001	0.189 ± 0.008	0.570 ± 0.007	0.538 ± 0.001	0.189 ± 0.001
SDA	test	0.557 ± 0.002	0.541 ± 0.001	0.192 ± 0.001	0.571 ± 0.003	0.541 ± 0.001	0.189 ± 0.002
MPMF	test	-	-	-	-	-	-
SSLF	test	0.634 ± 0.004	0.594 ± 0.002	0.191 ± 0.003	0.646 ± 0.002	0.623 ± 0.002	0.205 ± 0.002
LSCCA	test	0.561 ± 0.003	0.516 ± 0.002	0.175 ± 0.006	0.563 ± 0.005	0.521 ± 0.002	0.179 ± 0.002
SFSS	test	0.658 ± 0.013	0.603 ± 0.001	0.215 ± 0.002	0.681 ± 0.003	0.622 ± 0.003	0.218 ± 0.008
SFUS	test	0.643 ± 0.001	0.617 ± 0.001	0.214 ± 0.002	0.674 ± 0.002	0.619 ± 0.002	0.218 ± 0.003
LGME	test	0.644 ± 0.011	0.627 ± 0.002	0.221 ± 0.004	0.677 ± 0.002	0.613 ± 0.003	0.223 ± 0.002
C2MR	test	0.630 ± 0.001	0.607 ± 0.003	0.193 ± 0.002	0.658 ± 0.014	0.642 ± 0.007	0.197 ± 0.009
RSSL	test	$\mathbf{0.673 \pm 0.005}$	$\mathbf{0.642 \pm 0.001}$	$\mathbf{0.255 \pm 0.003}$	$\mathbf{0.697 \pm 0.002}$	$\mathbf{0.653 \pm 0.001}$	$\mathbf{0.267 \pm 0.002}$

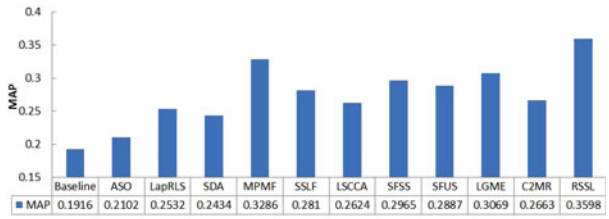
The best results are highlighted in bold.

TABLE 3
Experimental Results (Mean Microauc ± Standard Deviation, Mean Macroauc ± Standard Deviation and Mean F1 ± Standard Deviation) on the NUS-WIDE Data Set

Method		$n = 5{,}000$			$n = 10{,}000$		
		MicroAUC	MacroAUC	F1	MicroAUC	MacroAUC	F1
Baseline	learn	0.677 ± 0.003	0.594 ± 0.002	0.225 ± 0.010	0.697 ± 0.002	0.612 ± 0.002	0.316 ± 0.005
ASO	learn	0.709 ± 0.001	0.636 ± 0.001	0.327 ± 0.003	0.724 ± 0.002	0.641 ± 0.001	0.343 ± 0.008
LapRLS	learn	0.708 ± 0.001	0.663 ± 0.004	0.354 ± 0.004	0.712 ± 0.005	0.664 ± 0.007	0.333 ± 0.006
SDA	learn	0.709 ± 0.001	0.642 ± 0.003	0.310 ± 0.001	0.722 ± 0.002	0.644 ± 0.006	0.347 ± 0.009
MPMF	learn	0.665 ± 0.003	0.719 ± 0.002	0.358 ± 0.006	0.674 ± 0.001	0.775 ± 0.001	0.381 ± 0.005
SSLF	learn	0.708 ± 0.002	0.677 ± 0.009	0.326 ± 0.010	0.738 ± 0.002	0.685 ± 0.003	0.341 ± 0.007
LSCCA	learn	0.618 ± 0.002	0.724 ± 0.003	0.258 ± 0.002	0.632 ± 0.002	0.732 ± 0.003	0.264 ± 0.002
SFSS	learn	0.757 ± 0.009	0.722 ± 0.004	0.480 ± 0.007	0.786 ± 0.006	0.736 ± 0.004	0.485 ± 0.011
SFUS	learn	0.753 ± 0.006	0.727 ± 0.001	0.390 ± 0.009	0.785 ± 0.004	0.725 ± 0.003	0.407 ± 0.008
LGME	learn	0.761 ± 0.003	0.739 ± 0.005	0.453 ± 0.007	0.780 ± 0.006	0.744 ± 0.004	0.471 ± 0.04
C2MR	learn	0.689 ± 0.003	0.621 ± 0.003	0.301 ± 0.009	0.770 ± 0.002	0.655 ± 0.002	0.351 ± 0.003
RSSL	learn	$\mathbf{0.835 \pm 0.008}$	$\mathbf{0.768 \pm 0.006}$	$\mathbf{0.576 \pm 0.005}$	$\mathbf{0.844 \pm 0.002}$	$\mathbf{0.791 \pm 0.009}$	$\mathbf{0.589 \pm 0.003}$
Baseline	test	0.658 ± 0.002	0.550 ± 0.002	0.131 ± 0.003	0.674 ± 0.001	0.560 ± 0.001	0.152 ± 0.004
ASO	test	0.673 ± 0.002	0.559 ± 0.004	0.157 ± 0.002	0.675 ± 0.001	0.574 ± 0.004	0.159 ± 0.003
LapRLS	test	0.677 ± 0.001	0.554 ± 0.009	0.140 ± 0.003	0.691 ± 0.007	0.561 ± 0.008	0.151 ± 0.009
SDA	test	0.672 ± 0.002	0.613 ± 0.0021	0.165 ± 0.001	0.682 ± 0.008	0.633 ± 0.009	0.160 ± 0.008
MPMF	test	-	-	-	-	-	-
SSLF	test	0.683 ± 0.002	0.630 ± 0.002	0.173 ± 0.007	0.696 ± 0.001	0.653 ± 0.001	0.189 ± 0.002
LSCCA	test	0.588 ± 0.001	0.613 ± 0.001	0.158 ± 0.002	0.599 ± 0.003	0.621 ± 0.006	0.192 ± 0.004
SFSS	test	0.706 ± 0.001	0.630 ± 0.002	0.228 ± 0.004	0.731 ± 0.002	0.653 ± 0.002	0.239 ± 0.005
SFUS	test	0.708 ± 0.001	0.633 ± 0.002	0.238 ± 0.004	0.735 ± 0.001	0.662 ± 0.001	0.247 ± 0.003
LGME	test	0.711 ± 0.001	0.641 ± 0.001	0.243 ± 0.001	0.729 ± 0.001	0.657 ± 0.001	0.248 ± 0.003
C2MR	test	0.642 ± 0.004	0.620 ± 0.002	0.220 ± 0.001	0.767 ± 0.002	0.641 ± 0.001	0.225 ± 0.001
RSSL	test	$\mathbf{0.773 \pm 0.001}$	$\mathbf{0.703 \pm 0.004}$	$\mathbf{0.269 \pm 0.006}$	$\mathbf{0.795 \pm 0.005}$	$\mathbf{0.731 \pm 0.006}$	$\mathbf{0.272 \pm 0.007}$

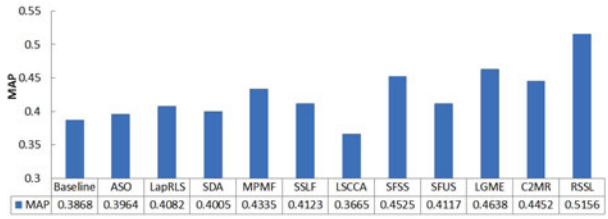
The best results are highlighted in bold.

First, from the results in Tables 2 and 3, we can see that the proposed method gains the best performances among all of the compared algorithms in terms of mean MicroAUC, MacroAUC and F1 on both the MIRFlicrk and NUS-WIDE data sets. This indicates that the proposed RSSL enables to effectively learn a robust structured subspace from data.



	Baseline	ASO	LapRLS	SDA	MPMF	SSLF	LSCCA	SFSS	SFUS	LGME	C2MR	RSSL
MAP	0.1916	0.2102	0.2532	0.2434	0.3286	0.281	0.2624	0.2965	0.2887	0.3069	0.2663	0.3598

(a) MIRFlickr



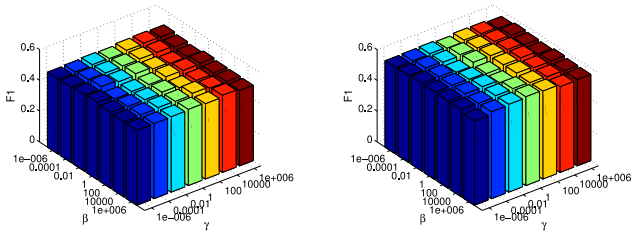
	Baseline	ASO	LapRLS	SDA	MPMF	SSLF	LSCCA	SFSS	SFUS	LGME	C2MR	RSSL
MAP	0.3868	0.3964	0.4082	0.4005	0.4335	0.4123	0.3665	0.4525	0.4117	0.4638	0.4452	0.5156

(b) NUS-WIDE

Fig. 1. Comparison of different algorithms on MIFlickr and NUS-WIDE learning data sets with $n = 5{,}000$ in terms of MAP.

Second, compared with the related approaches, i.e., ASO, SSLF, LapRLS, SFSS, SFUS and LGME, the proposed RSSL achieves significant improvements, which demonstrates the necessity and advantage of the introduced $\ell_{2,1}$-norm and smooth regularization. $\ell_{2,1}$-norm makes RSSL robust to noise and outliers. By mining the local geometric structure, RSSL maps visually similar images close to each other in the hidden subspace. Thus they have consistent feature representations in the hidden subspace, which makes much easier to assign concepts to images in the subspace. Third, RSSL, LGME and SFSS are superior to other methods in general, such as ASO, SDA, SSLF and SFUS. It indicates the effectiveness of the local consistent constraint over tags, which gives an intuitive interpretation of better performance of the proposed algorithm since visually similar images can implicitly share common tags. On the other hand, the noise in tags can also be somewhat alleviated by exploiting visual geometric structure. Fourthly, by exploring the feature combinations in the prediction process, RSSL, LGME, SFUS, SSLF and ASO are better than Baseline. Fifth, to jointly mining the image tagging information and visual content information, MPMF factorizes the image-tag relation matrix, image correlation matrix and tag correlation matrix simultaneously and it cannot map new images into the learned model. On the contrary, the proposed model learns a robust latent subspace by exploiting the image tagging information and visual content information simultaneously and can easily assign tags to new images. The proposed RSSL is better than MPMF since RSSL can learn more robust and compact latent subspace. Sixthly, RSSL outperforms LSCCA by jointly uncovering the image tagging information and image content. The proposed framework can reduce the noise-induced uncertainty. In addition, RSSL achieves better results than C2MR, which indicates that it is better to jointly explore the visual geometric structure and the tag local and global consistency.

Finally, we present the performance of image retrieval in terms of MAP of all the compared methods on the learning data sets with $n = 5{,}000$ of MIRFlcirk and NUS-WIDE data sets in Fig. 1. It is observed that RSSL performs significantly better than other methods for image retrieval, which shows that our method can rank the relevant tags at the top positions.



(a) MIRFlickr (b) NUS-WIDE

Fig. 2. The performance in terms of F1 by varying the parameters β and γ on MIFlickr and NUS-WIDE learning data sets with $n = 5{,}000$.

6.1.6 Parameter Sensitiveness

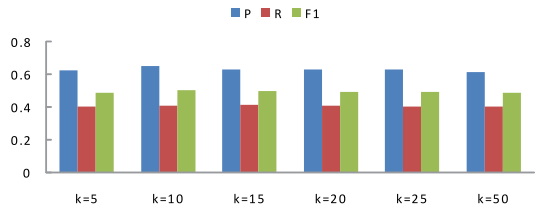
In the proposed framework, there are several parameters to be tuned. In our experiments, it is observed that the performance is not sensitive to the dimensionality r of the latent subspace and we set $r = 5 \times \lfloor \frac{m-1}{5} \rfloor$, where m is the number of concepts and $\lfloor c \rfloor$ denotes the largest integer not greater than c. When α is larger than 10, the performance of RSSL is good for the two image data sets. Besides, we observe that RSSL is not sensitive to λ when it is in the range of $[10^{-3}, 10^{3}]$. To validate how the rest parameters affect the performance, we conduct experiments to evaluate their sensitivity. The MIRFlickr and NUS-WIDE data sets with $n = 5{,}000$ learning data are used.

Fig. 2 shows the performance variance *w.r.t.* β and γ in terms of F1 on the two data sets. It is observed that the tagging performance varies corresponding to different values of the parameters β and γ. The performance is good when the parameter β is not too large or small, which demonstrates the necessity of the smooth regularization on the underlying geometric structure between samples in the latent subspace. However, due to images represented by the low-level visual features, large β may introduce inaccurate information, which degrades the performance. For the parameter γ, we can see that it should not be small. Large γ makes the learned $\mathbf{W}$ satisfy the expected properties, which guarantees that better results are achieved.

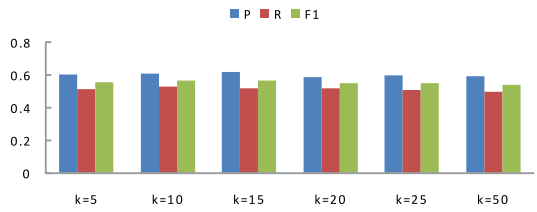
The underlying geometric structure preservation in the latent subspace is dependent on the neighbor number k to compute the Laplcian matrix. In this experiments we tune k within the range of $\{5, 10, 15, 20, 25, 50\}$. The performances in terms of recall, precision and F1 by varying k on the two databases are presented in Fig. 3. We observe that the performance of the proposed RSSL varies slightly with varying k when k is not large. Thus, in our experiments, we fix $k = 15$ for both the MIRFlickr and NUS-WIDE data sets.

6.1.7 Convergence Study

To solve the proposed formulation, we develop an iterative update algorithm. Now we experimentally validate its convergence and study the speed of convergence. Following the above experiments, the corresponding experiments are conducted on the MIRFlickr and NUS-WIDE datasets with $n = 5{,}000$ learning data. The convergence rates are shown in Fig. 4. From these figures, we can see that the value of our objective function monotonically decreases when the iteration



(a) MIRFlickr



(b) NUS-WIDE

Fig. 3. The performance in terms of precision, recall and F1 by varying k to compute $\mathbf{L}$ on MIFlickr and NUS-WIDE learning data sets with $n = 5{,}000$.

round increases and changes only a little bit after five iterations for these two data sets, demonstrating that the proposed optimization algorithm is effective and converges quickly.

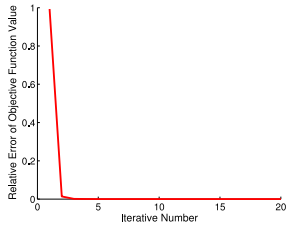
6.2 Clustering

In this section, we evaluate the performance of the proposed formulation for unsupervised clustering. The performance in terms of clustering is measured by two widely used evaluation metrics, i.e., Accuracy (ACC) and normalized mutual information (NMI) [18], [19]. The experiments are conducted on seven publicly available datasets, including three face image data sets (i.e., UMIST [41], JAFFE [42] and Pointing4 [43]), two handwritten digit data sets (*i.e.*, a subset of USPS [41] and Binary Alphabet (BA) [41]), and two text data sets (i.e., tr11 [44] and oh15 [44]). Data sets from different areas serve as a good test bed for a comprehensive evaluation. Table 4 summarizes the details of these seven benchmark data sets.

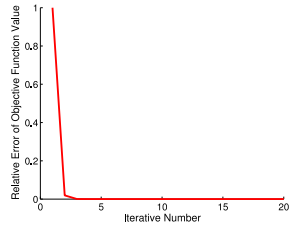
As stated above, the goal of the proposed method is to learn a suitable representation and our method is suitable for clustering. Therefore, we adopt the learned $\mathbf{F}$ by the proposed method for clustering and denote it as RSSL. Besides, to validate the performance of the proposed method for representation learning, the proposed method is treated as a feature selection algorithm, denoted as RSSL-FS. That is, we first map data into the underlying subspace and then select

TABLE 4
Dataset Description

Domain	Dataset	n	d	c
Image, Face	UMIST	575	644	20
	JAFFE	213	1,024	10
	Poingting4	2,790	1,120	15
Image, Handwritten Digits	USPS	400	256	10
	BA	1,404	320	36
Text	tr11	414	6,429	9
	oh15	913	3,100	10



(a) MIRFlickr



(b) NUS-WIDE

Fig. 4. Convergence curves of the relative error of the objective function value of the proposed algorithm.

features in the subspace by $\mathbf{P}$ to represent data. To demonstrate the performance of our method, we take all the original features for clustering as baseline and compare it with a various of representative algorithms, including spectral clustering (SC) [45], PCA [2], LPP [6] and several unsupervised feature selection methods, i.e., MaxVar (Features corresponding to the maximum variance are selected.), Laplacian Score [17], SPEC [15], MCFS [18], UDFS [19], NDFS [46] and LLCFS [47]. For feature selection methods, the numbers of features used for clustering are set as $\{5, 10, 20, 50, 100, 150, 200, 250, 300\}$ for all the datasets except USPS. Because the total feature number of USPS is 256, we set the number of selected features as $\{5, 10, 20, 50, 80, 110, 140, 170, 200\}$. In our experiments, we adopt Kmeans to cluster samples based on the selected features. The performance of Kmeans depends on initialization. We repeat the clustering 20 times with random initialization for each setup and report the average results. The average results with standard deviation (std) are presented in Tables 5 and 6.

It can be observed that our method is significantly superior to other algorithms, which verifies that the proposed RSSL enables to uncover more informative representations. And our method is the only one which has consistently high performance on all seven data sets. Intuitively, this indicates that it is necessary and useful to exploit the latent subspace to find the discriminative representation of data.

6.3 Classification

In this section we apply our method to the problem of classification including semi-supervised classification and supervised classification. To validate the performance, we conduct experiments on four datasets, i.e., UMIST, JAFFE, USPS and tr11. The performance is measured by the widely used evaluation metric, i.e., classification ACcuracy.

For semi-supervised classification, we randomly choose 50 percent of samples as training data and the rest as testing data. We further randomly sample s percent of training data as labeled data. In our experiments, we set s as 5, 10, 20 and 50, respectively, and report all of the results. The proposed algorithm is compared with several semi-supervised methods, including LapRLS [34], SDA [35], LSR [48] and SFSS [39]. The experiments are independently repeated 20 times to generate different training and testing data, and the average results are reported. The compared results are presented in Tables 7 and 8. We can see that the performance of all the algorithms improve as the number of labeled samples increases. Furthermore, it also is observed that our framework significantly outperforms all of the compared semi-supervised classification approaches, which

TABLE 5
Clustering Results (ACC% ± Std) of Different Algorithms for Clustering on Different Datasets

Dataset	Face			Handwritten Digits		Text	
	UMIST	JAFFE	Pointing4	USPS	BA	tr11	oh15
Baseline	41.8 ± 1.7	72.5 ± 2.2	35.9 ± 1.2	62.6 ± 2.3	40.3 ± 2.0	30.9 ± 2.0	30.4 ± 1.0
SC	59.6 ± 2.3	75.8 ± 1.8	64.6 ± 2.2	71.5 ± 2.3	44.4 ± 2.2	31.1 ± 1.1	39.4 ± 2.7
PCA	43.3 ± 2.2	78.2 ± 1.5	36.5 ± 2.1	65.5 ± 2.5	43.4 ± 2.4	32.4 ± 2.4	34.9 ± 0.8
LPP	56.5 ± 2.4	69.7 ± 2.4	47.1 ± 2.0	64.0 ± 1.8	44.1 ± 1.8	48.2 ± 3.2	33.4 ± 1.6
MaxVar	45.8 ± 2.8	67.3 ± 0.8	44.0 ± 1.8	63.8 ± 1.3	40.7 ± 1.7	30.5 ± 1.0	32.9 ± 2.2
LS	45.9 ± 1.9	74.0 ± 1.6	37.1 ± 1.6	64.9 ± 2.1	42.1 ± 1.7	31.6 ± 1.6	33.8 ± 1.7
SPEC	47.9 ± 2.0	76.9 ± 1.2	38.6 ± 1.2	65.5 ± 1.8	42.2 ± 2.2	35.1 ± 1.8	33.8 ± 2.1
MCFS	46.3 ± 1.6	78.8 ± 2.1	46.2 ± 1.9	64.4 ± 1.1	41.5 ± 1.8	36.3 ± 2.1	33.8 ± 1.8
UDFS	48.6 ± 2.7	76.7 ± 1.1	45.1 ± 1.4	66.2 ± 1.7	42.7 ± 1.8	35.9 ± 1.5	32.9 ± 1.3
NDFS	51.3 ± 1.9	81.2 ± 1.1	48.9 ± 2.2	67.3 ± 0.7	43.4 ± 2.0	37.4 ± 2.5	34.8 ± 0.9
LLCFS	52.6 ± 2.2	81.6 ± 3.5	65.7 ± 2.6	68.9 ± 1.4	41.8 ± 1.8	38.0 ± 1.4	35.8 ± 1.8
RSSL-FS	$\mathbf{65.2 \pm 0.3}$	$\mathbf{91.5 \pm 0.4}$	$\mathbf{69.1 \pm 1.4}$	$\mathbf{76.5 \pm 1.8}$	$\mathbf{47.6 \pm 0.6}$	$\mathbf{57.8 \pm 1.5}$	$\mathbf{48.2 \pm 1.4}$
RSSL	$\mathbf{67.9 \pm 0.8}$	$\mathbf{97.4 \pm 1.2}$	$\mathbf{76.9 \pm 1.1}$	$\mathbf{79.6 \pm 0.4}$	$\mathbf{49.2 \pm 0.7}$	$\mathbf{58.4 \pm 0.6}$	$\mathbf{54.0 \pm 0.7}$

The best results are highlighted in bold.

TABLE 6
Clustering Results (NMI% ± Std) of Different Algorithms for Clustering on Different Datasets

Dataset	Face			Handwritten Digits		Text	
	UMIST	JAFFE	Pointing4	USPS	BA	tr11	oh15
Baseline	62.3 ± 2.3	80.0 ± 2.7	41.7 ± 1.4	62.6 ± 3.3	40.3 ± 2.0	7.0 ± 1.4	17.7 ± 3.0
SC	77.2 ± 3.2	85.6 ± 1.8	73.8 ± 2.0	69.2 ± 3.0	60.3 ± 0.9	2.5 ± 0.3	30.0 ± 1.1
PCA	63.8 ± 2.6	83.4 ± 1.9	47.4 ± 1.4	58.7 ± 2.0	58.4 ± 1.2	7.7 ± 2.4	26.0 ± 2.4
LPP	74.2 ± 3.0	80.8 ± 3.3	56.0 ± 2.1	62.4 ± 1.5	59.5 ± 0.9	40.2 ± 8.1	24.0 ± 2.8
MaxVar	63.5 ± 1.5	70.3 ± 1.2	50.8 ± 1.8	58.1 ± 1.7	56.9 ± 1.3	7.6 ± 1.1	22.1 ± 2.7
LS	63.9 ± 1.8	79.4 ± 4.0	42.7 ± 1.2	58.7 ± 1.0	57.3 ± 0.8	8.0 ± 2.0	23.2 ± 2.8
SPEC	65.2 ± 2.0	82.8 ± 1.8	40.5 ± 1.0	59.5 ± 1.1	57.9 ± 1.1	11.5 ± 2.9	23.6 ± 2.2
MCFS	66.7 ± 1.9	83.4 ± 2.0	53.1 ± 1.1	59.3 ± 0.9	57.5 ± 0.8	13.5 ± 3.3	23.1 ± 3.0
UDFS	67.3 ± 3.0	82.3 ± 3.5	52.4 ± 1.7	60.1 ± 2.3	58.1 ± 1.0	13.7 ± 1.9	21.8 ± 2.0
NDFS	69.7 ± 2.3	86.3 ± 4.1	56.4 ± 1.3	61.3 ± 2.3	58.8 ± 0.8	14.2 ± 3.0	24.2 ± 2.7
LLCFS	71.2 ± 2.5	85.8 ± 4.4	73.2 ± 2.9	65.8 ± 1.0	57.3 ± 1.0	31.9 ± 2.1	24.9 ± 1.5
RSSL-FS	$\mathbf{80.5 \pm 2.5}$	$\mathbf{94.7 \pm 2.1}$	$\mathbf{77.9 \pm 2.8}$	$\mathbf{72.2 \pm 0.5}$	$\mathbf{62.3 \pm 0.9}$	$\mathbf{41.9 \pm 2.2}$	$\mathbf{36.8 \pm 1.9}$
RSSL	$\mathbf{81.4 \pm 1.2}$	$\mathbf{96.2 \pm 0.7}$	$\mathbf{82.4 \pm 2.8}$	$\mathbf{73.4 \pm 0.3}$	$\mathbf{62.7 \pm 0.3}$	$\mathbf{43.2 \pm 1.0}$	$\mathbf{39.3 \pm 1.1}$

The best results are highlighted in bold.

TABLE 7
Semi-Supervised Classification Results (AC% ± Std) of Different Algorithms on UMIST and JAFFE Datasets

Dataset		UMIST				JAFFE			
		$s = 5$	$s = 10$	$s = 20$	$s = 50$	$s = 5$	$s = 10$	$s = 20$	$s = 50$
LapRLS	semi	51.9 ± 1.7	56.1 ± 1.9	59.3 ± 1.6	65.8 ± 1.4	81.9 ± 2.7	85.3 ± 1.7	90.1 ± 1.9	91.8 ± 1.3
SDA	semi	34.4 ± 3.1	39.0 ± 2.0	42.1 ± 1.8	46.3 ± 1.4	72.3 ± 3.3	75.9 ± 2.6	82.0 ± 1.9	89.2 ± 0.6
LSR	semi	30.9 ± 2.7	37.7 ± 2.1	46.3 ± 1.4	65.2 ± 1.3	77.8 ± 2.9	85.7 ± 2.1	87.9 ± 1.9	90.6 ± 2.0
SFSS	semi	41.2 ± 1.7	59.3 ± 2.6	72.0 ± 1.8	89.0 ± 1.4	76.7 ± 3.6	87.9 ± 1.1	96.5 ± 0.7	99.9 ± 0.4
RSSL	semi	$\mathbf{60.5 \pm 1.6}$	$\mathbf{71.8 \pm 0.5}$	$\mathbf{82.9 \pm 0.6}$	$\mathbf{93.4 \pm 0.3}$	$\mathbf{92.7 \pm 0.9}$	$\mathbf{97.6 \pm 0.7}$	$\mathbf{99.9 \pm 0.2}$	$\mathbf{100.0 \pm 0.0}$
LapRLS	test	50.4 ± 2.7	56.4 ± 1.6	58.5 ± 1.2	63.7 ± 1.5	82.7 ± 1.8	84.1 ± 1.8	90.0 ± 1.8	90.2 ± 0.7
SDA	test	32.4 ± 1.8	37.1 ± 1.0	40.8 ± 1.1	45.5 ± 1.0	70.2 ± 3.8	75.6 ± 2.3	80.6 ± 1.2	85.7 ± 0.8
LSR	test	–	–	–	–	–	–	–	–
SFSS	test	40.2 ± 1.2	58.9 ± 2.2	71.6 ± 2.4	88.2 ± 1.0	76.8 ± 3.2	88.1 ± 1.5	96.0 ± 1.6	99.5 ± 0.7
RSSL	test	$\mathbf{59.7 \pm 1.6}$	$\mathbf{71.2 \pm 0.9}$	$\mathbf{82.2 \pm 1.3}$	$\mathbf{93.9 \pm 1.0}$	$\mathbf{89.5 \pm 1.2}$	$\mathbf{96.3 \pm 1.1}$	$\mathbf{99.6 \pm 0.4}$	$\mathbf{100.0 \pm 0.0}$

The best results are highlighted in bold.

indicates that RSSL can effectively learn a representation of data and classifiers from the labeled and unlabeled data.

The performance of the proposed RSSL for supervised classification is also validated on these four datasets. We randomly choose c percent of samples as training data and the rest as testing data. In our experiments, we set c as 5, 10, 20 and 50, respectively. The experiments are independently repeated 20 times to generate different training and testing data, and the average results are reported. To demonstrate the superiority of the proposed RSSL, we compared it with several state-of-the-art methods, i.e., ASO [33], FSNM [4], SSLF [37], LSCCA [38], SFUS [1] and SFSS [39]. The

TABLE 8
Semi-Supervised Classification Results (AC% $\pm$ Std) of Different Algorithms on USPS and tr11 Datasets

Dataset		USPS				tr11			
		$s=5$	$s=10$	$s=20$	$s=50$	$s=5$	$s=10$	$s=20$	$s=50$
LapRLS	semi	56.5 ± 1.3	63.6 ± 2.2	68.9 ± 1.7	71.3 ± 1.9	40.7 ± 2.7	45.9 ± 2.3	51.3 ± 1.5	56.6 ± 1.2
SDA	semi	44.7 ± 2.4	50.4 ± 2.3	54.4 ± 3.0	61.0 ± 1.3	37.2 ± 2.0	45.5 ± 2.1	52.6 ± 2.2	57.2 ± 1.4
LSR	semi	50.2 ± 2.3	65.5 ± 2.9	70.7 ± 3.0	78.4 ± 2.2	52.0 ± 3.4	62.1 ± 2.6	67.6 ± 2.2	72.9 ± 2.5
SFSS	semi	50.3 ± 2.8	62.1 ± 2.5	72.3 ± 2.2	80.9 ± 2.3	46.6 ± 2.7	57.4 ± 3.2	69.6 ± 2.7	74.9 ± 1.7
RSSL	semi	$\mathbf{68.8 \pm 1.5}$	$\mathbf{76.4 \pm 1.1}$	$\mathbf{83.4 \pm 0.67}$	$\mathbf{91.8 \pm 0.6}$	$\mathbf{64.7 \pm 0.96}$	$\mathbf{75.4 \pm 0.5}$	$\mathbf{81.4 \pm 0.6}$	$\mathbf{86.7 \pm 0.3}$
LapRLS	test	56.0 ± 2.3	63.0 ± 2.9	68.5 ± 2.2	70.7 ± 1.4	40.1 ± 2.8	45.4 ± 1.4	49.3 ± 1.5	56.4 ± 1.0
SDA	test	44.2 ± 2.6	47.3 ± 1.8	53.0 ± 2.1	60.6 ± 1.6	37.0 ± 3.0	46.7 ± 2.4	50.3 ± 1.6	57.3 ± 1.2
LSR	test	–	–	–	–	–	–	–	–
SFSS	test	49.6 ± 2.6	60.8 ± 1.9	71.1 ± 2.0	80.4 ± 1.0	44.3 ± 3.1	55.5 ± 3.4	69.4 ± 1.5	73.1 ± 0.7
RSSL	test	$\mathbf{67.4 \pm 0.9}$	$\mathbf{74.7 \pm 1.5}$	$\mathbf{81.5 \pm 1.2}$	$\mathbf{89.5 \pm 1.1}$	$\mathbf{63.5 \pm 0.9}$	$\mathbf{73.3 \pm 0.8}$	$\mathbf{80.3 \pm 0.6}$	$\mathbf{84.5 \pm 0.4}$

The best results are highlighted in bold.

TABLE 9
Classification Results (AC% $\pm$ Std) of Different Algorithms for Supervised Classification on UMIST and JAFFE Datasets

Dataset	UMIST				JAFFE			
	$c=5$	$c=10$	$c=20$	$c=50$	$c=5$	$c=10$	$c=20$	$c=50$
ASO	60.4 ± 2.8	71.8 ± 3.3	85.3 ± 2.7	94.8 ± 1.1	90.6 ± 1.4	95.4 ± 1.7	98.8 ± 0.7	$\mathbf{100.0 \pm 0.0}$
FSNM	46.3 ± 1.8	69.0 ± 2.5	85.6 ± 2.0	94.6 ± 1.5	78.7 ± 3.0	92.5 ± 1.3	99.0 ± 0.8	99.9 ± 0.3
SSLF	47.5 ± 1.9	72.4 ± 2.5	87.4 ± 1.6	96.5 ± 1.3	87.3 ± 2.3	95.6 ± 1.5	99.4 ± 0.6	$\mathbf{100.0 \pm 0.0}$
LSCCA	47.6 ± 2.3	70.1 ± 2.2	86.1 ± 1.9	94.8 ± 1.4	86.3 ± 3.3	95.9 ± 1.9	99.1 ± 0.7	99.9 ± 0.3
SFUS	48.4 ± 2.0	72.3 ± 2.8	87.3 ± 2.1	95.6 ± 1.1	80.9 ± 2.2	96.9 ± 1.6	99.5 ± 0.5	$\mathbf{100.0 \pm 0.0}$
SFSS	48.3 ± 2.1	72.2 ± 1.6	87.2 ± 1.6	96.9 ± 0.9	83.9 ± 1.8	95.2 ± 1.5	99.2 ± 0.7	$\mathbf{100.0 \pm 0.0}$
RSSL	$\mathbf{65.2 \pm 0.9}$	$\mathbf{83.5 \pm 0.8}$	$\mathbf{93.6 \pm 0.7}$	$\mathbf{98.8 \pm 0.4}$	$\mathbf{96.8 \pm 1.0}$	$\mathbf{99.7 \pm 0.3}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$

The best results are highlighted in bold.

TABLE 10
Classification Results (AC% $\pm$ Std) of Different Algorithms for Supervised Classification on USPS and tr11 Datasets

Dataset	USPS				tr11			
	$c=5$	$c=10$	$c=20$	$c=50$	$c=5$	$c=10$	$c=20$	$c=50$
ASO	62.2 ± 2.7	69.3 ± 2.1	74.2 ± 2.2	82.2 ± 1.3	54.4 ± 3.7	69.7 ± 2.3	75.9 ± 1.3	82.9 ± 1.8
FSNM	60.8 ± 2.3	64.4 ± 2.8	71.9 ± 1.4	81.5 ± 1.2	50.8 ± 3.3	63.6 ± 2.7	76.1 ± 2.9	83.3 ± 2.1
SSLF	62.9 ± 2.7	70.3 ± 1.6	74.1 ± 1.7	83.4 ± 0.9	56.0 ± 4.7	73.0 ± 2.1	80.1 ± 1.5	83.7 ± 1.9
LSCCA	61.9 ± 2.1	63.3 ± 2.7	64.9 ± 2.7	66.1 ± 3.0	42.4 ± 5.0	54.2 ± 3.0	67.3 ± 2.0	81.0 ± 1.4
SFUS	64.5 ± 2.7	70.0 ± 1.9	77.8 ± 2.2	83.3 ± 2.2	60.1 ± 2.4	73.5 ± 2.2	80.4 ± 2.6	84.8 ± 1.6
SFSS	64.9 ± 2.5	72.2 ± 2.3	78.6 ± 1.2	84.3 ± 1.5	58.8 ± 4.3	72.5 ± 2.3	81.0 ± 1.8	85.5 ± 0.8
RSSL	$\mathbf{73.9 \pm 1.4}$	$\mathbf{82.9 \pm 1.1}$	$\mathbf{88.5 \pm 0.9}$	$\mathbf{92.6 \pm 0.2}$	$\mathbf{71.7 \pm 1.1}$	$\mathbf{82.3 \pm 1.4}$	$\mathbf{89.9 \pm 0.6}$	$\mathbf{94.8 \pm 0.2}$

The best results are highlighted in bold.

results measured by classification accuracy are reported in Tables 9 and 10. From the results, we can see that the proposed RSSL obviously achieves the best performance.

From these experiments, we conclude that our method is well suited to classification problems, including semi-supervised and supervised ones.

7 Discussion

The proposed formulation (11) is a general one, and can be used to explain several existing algorithms as special cases.

Connection with feature selection algorithms. By now many feature selection methods have been studied, such as NDFS [46] and SFUS [1].

First, in unsupervised scenarios there is no labeled data. If we set $r = d$ and $\mathbf{Q} = \mathbf{I}_d$, we have $\mathbf{W} = \mathbf{P}$ and the objective function (23) is changed to

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{W}} \|\mathbf{F} - \mathbf{X}^T\mathbf{W}\|_{2,1} + \alpha \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \lambda\|\mathbf{W}\|_{2,1} \\ &\text{s.t.} \qquad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{F} \geq 0. \end{aligned} \tag{28}$$

If we use the least-squares loss function, the above formulation is same to the formulation of NDFS [46]. Our formulation is robust to the outliers using the $\ell_{2,1}$-norm based loss function.

Second, for supervised learning, we set $\zeta \to \infty$ to make $\mathbf{F} = \mathbf{Y}$ since all the data are labeled. If we set $\beta = 0$ and use the regularization term $\|\mathbf{W}\|_{2,1}$ instead of $\|\mathbf{P}\|_{2,1}$ to facilitate feature selection, our formulation leads to

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{Q},\mathbf{W}} \|\mathbf{Y} - \mathbf{X}^T\mathbf{W}\|_{2,1} + \gamma\|\mathbf{W} - \mathbf{Q}\mathbf{P}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \\ &\text{s.t.} \qquad \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r, \end{aligned} \tag{29}$$

which is the formulation of SFUS [1].

Connection with semi-supervised learning (SSL) algorithms. We now discuss the relationship between the proposed formulation and several Semi-supervised Learning algorithms.

First, a SSL algorithm is proposed for multitask learning in [33], which is closed to the proposed formulation. When $\zeta \to \infty$, we have $\mathbf{F}_{labeled} = \mathbf{Y}_{labeled}$. If we set $\beta = 0$ and $\lambda = 0$ and adopt the same loss function, the proposed formulation reduces to the one in [33] in the special case where the input data are the same for all tasks. Besides, if we further set $r = c$ and $\mathbf{P} = \mathbf{I}_c$, we have $\mathbf{W} = \mathbf{Q}$ and the proposed framework leads to LapRLS [34] with linear predictive function.

Second, in [39], a semi-supervised feature analyzing framework for multimedia data understanding is proposed, which is formulated as

$$\begin{aligned}\min_{\mathbf{F},\mathbf{W}} & \alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]) \\ & + \|\mathbf{F} - \mathbf{X}^T\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}.\end{aligned} \tag{30}$$

If we set $r = d$ and $\mathbf{Q} = \mathbf{I}_d$, we have $\mathbf{W} = \mathbf{P}$. Then using the least-squares loss function, our formulation leads to the above formulation.

Besides, if we use Frobenius norm rather than $\ell_{2,1}$-norm and impose a regularization term on $\mathbf{W}$ instead of $\mathbf{P}$, by setting $\beta = 0$, our formulation reduces to

$$\begin{aligned}\min_{\mathbf{F},\mathbf{P},\mathbf{Q},\mathbf{W}} & \alpha(\mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \mathrm{Tr}[(\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})]) \\ & + \|\mathbf{F} - \mathbf{X}^T\mathbf{W}\|_F^2 + \gamma\|\mathbf{W} - \mathbf{Q}\mathbf{P}\|_F^2 + \lambda\|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r.\end{aligned} \tag{31}$$

It is same to the objective function in [40]. In [40], the relationships to dimensionality reduction algorithms, transductive classification and traditional graph regularization have been discussed. Thus, the relationships analyzed in [40] are appropriate to ours.

Connection with multi-label classification. When all data are labeled and $\zeta \to \infty$, we we have $\mathbf{F} = \mathbf{Y}$. Then when the regularization parameter $\beta = 0$, the proposed formulation reduces to the one in [37] in the special case where we employ Frobenius norm rather than $\ell_{2,1}$-norm. In [37], it discusses its connections with several algorithms, such as the classical ridge regression. Similarly, we can build the corresponding relationships between those algorithms and ours.

8 Conclusion

In this paper, we propose a subspace learning framework which can learn an appropriate representation for data by incorporating image understanding and feature learning into a unified framework. It exploits the visual geometric structure and the local and global consistencies over labels simultaneously to uncover a underlying subspace robust to the outliers and noise. We formulate the subspace learning problem into an optimization problem and develop an iterative algorithm. Then we apply the proposed framework to several image understanding tasks, i.e., image tagging, clustering and classification. And the proposed formulation is a general framework that can include several well-known formulations as special cases. For evaluation, we conduct extensive experiments to compare the proposed algorithm with related methods for different image understanding tasks. Extensive experiments over diverse public data sets show that the proposed algorithm is quite effective to uncover a latent subspace for image understanding tasks.

Acknowledgments

This work was partially supported by the 973 Program (Project No. 2014CB347600), the National Natural Science Foundation of China (Grant No. 61402228, 61472422, 61332016, 61103059), Natural Science Fund for Distinguished Young Scholars of Jiangsu Province under Grant BK2012033 and the Open Projects Program of National Laboratory of Pattern Recognition. Jing Liu is the corresponding author.

References

[1] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, and A. Hauptmann, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[2] P. N. Belhumeur, J. P. Hespanha, X. Wu, 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.

[3] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[4] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[5] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 26, pp. 2138–2150, Sep. 2014.

[6] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 153–160.

[7] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[8] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 22, pp. 2319–2323, 2000.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 29, no. 5500, pp. 2323–2326, 2000.

[11] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Recognition*, 2nd ed. New York, USA: Wiley, 2001.

[14] F. D. la Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1–3, pp. 117–142, 2003.

[15] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. Int. Conf. Mach. Learning*, 2007, pp. 1151–1157.

[16] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, Jun. 2011.

[17] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," presented at the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2005.

[18] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.

[19] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.

[20] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1208–1213.

[21] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 262–286, 2006.

[22] D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *Proc. IEEE 7th Int. Conf. Data Mining*, 2007, pp. 73–82.

[23] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.

[24] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Image annotation via graph learning," *Pattern Recognit.*, vol. 42, no. 2, pp. 218–228, 2009.

[25] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[26] F. R. K. Chung, *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*, Amer. Math. Soc., Providence, RI, USA, 1997.

[27] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed. Manchester, England: SIAM, 2002.

[28] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proc. Berkeley Symp. Math. Statist. Probabilistics*, 1951, pp. 481–492.

[29] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, p. 14, 2011.

[30] Z. Li, J. Liu, J. Tang, and H. Lu, "Projective matrix factorization with unified embedding for social image tagging," *Comput. Vis. Image Understanding*, vol. 124, pp. 71–78, 2014.

[31] M. Huiskes and M. Lew, "The MIR flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.

[32] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from Nation University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.

[33] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeld data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.

[34] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[35] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–7.

[36] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1187–1190.

[37] S. Ji, L. Zhang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 2, pp. 1817–1853, 2010.

[38] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 33, pp. 2194–200, Jan. 2011.

[39] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, and A. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Dec. 2012.

[40] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. Hauptmann, "Web & personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[41] [Online]. Available: http://cs.nyu.edu/~roweis/data.html

[42] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[43] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. ICPR Workshop Vis. Observation Deictic Gestures*, 2004, pp. 1–9.

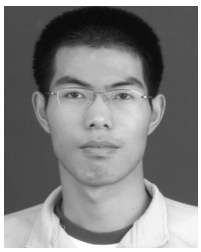
[44] [Online]. Available: http://tunedit.org/repo/Data/Text-wc

[45] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 568–586, Mar. 2011.

[46] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. Nat. Conf. Artif. Intell.*, 2012, pp. 1026–1032.

[47] H. Zeng and Y. ming Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[48] S. Xiang, F. Nie, and C. Zhang, "Semi-supervised classification via local spline regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2039–2053, Nov. 2010.



Zechao Li received the BE degree from the University of Science and Technology of China (USTC), Anhui Province, China, in 2008, and the PhD degree in pattern recognition and intelligent system from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2013. He is currently an assistant professor in the School of Computer Science, Nanjing University of Science and Technology, China. His research interests include machine learning, subspace learning, multimedia understanding, etc. He received the 2013 President Scholarship of Chinese Academy of Science.

Jing Liu received the BE and ME degrees in 2001 and 2004, respectively, from Shandong University, and the PhD degree from Institute of Automation, Chinese Academy of Sciences in 2008. Currently, she is an associate professor in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include machine learning, image content analysis and classification, multimedia information indexing and retrieval, etc.

Jinhui Tang received the BE and PhD degrees in July 2003 and July 2008, respectively, both from the University of Science and Technology of China (USTC). He is currently a professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. From July 2008 to December 2010, he was a research fellow in School of Computing, National University of Singapore. During that period, he visited School of Information and Computer Science, UC Irvine, from January 2010 to April 2010, as a visiting research scientist. From September 2011 to March 2012, he visited Microsoft Research Asia, as a visiting researcher. His current research interests include large-scale multimedia search, social media mining, and computer vision. He has authored more than 80 journal and conference papers in these areas. He serves as a editorial board member of *Pattern Analysis and Applications*, *Multimedia Tools and Applications*, *Information Sciences*, *Neurocomputing*, a Technical Committee member for about 30 international conferences, and a reviewer for about 30 prestigious international journals. He co-received the Best Paper Award in ACM Multimedia 2007, PCM 2011, and ICIMCS 2011. He is a member of the ACM and a senior member of the IEEE.

Hanqing Lu (M'05-SM'06) received the BE and ME degrees in 1982 and 1985, respectively, from Harbin Institute of Technology, and the PhD degree from Huazhong University of Sciences and Technology, Wuhan, China, in 1992. Currently, he is a professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, object tracking, recognition and image retrieval, etc. He has published more than 300 papers in those areas. He is a senior member of the IEEE.