Semi-Supervised Subspace Learning With L2graph

Xi Peng[a,*], Miaolong Yuan[a], Zhiding Yu[b], Wei Yun Yau[a], Lei Zhang[c]

[a] *Institute for Infocomm Research, A*STAR, Singapore 138632.*
[b] *Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA.*
[c] *Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, 610065, China.*

Abstract

Subspace learning aims to learn a projection matrix from a given training set so that a transformation of raw data to a low-dimensional representation can be obtained. In practice, the labels of some training samples are available, which can be used to improve the discrimination of low-dimensional representation. In this paper, we propose a semi-supervised learning method which is inspired by the biological observation of similar inputs having similar codes (SISC), i.e., the same collection of cortical columns of the mammal's visual cortex are always activated by the similar stimuli. More specifically, we propose a mathematical formulation of SISC which minimizes the distance among the data points with the same label while maximizing the separability between different subjects in the projection space. The proposed method, namely, semi-supervised L2graph (SeL2graph) has two advantages: 1) Unlike the classical dimension reduction methods such as principle component analysis, SeL2graph can automatically determine the dimension of feature space. This remarkably reduces the effort to find an optimal feature dimension for a good performance; and 2) It fully exploits the prior knowledge carried by the labeled samples and thus the obtained features are with higher discrimination and compactness. Extensive experiments show that the proposed method outperforms 7 subspace learning algorithms on 15 data sets with respect to classification accuracy, computational efficiency, and robustness to

[*] Corresponding author

Email addresses: pangsaai@gmail.com (Xi Peng), myuan@i2r.a-star.edu.sg (Miaolong Yuan), yzhiding@andrew.cmu.edu (Zhiding Yu), wyyau@i2r.a-star.edu.sg (Wei Yun Yau), leizhang@scu.edu.cn (Lei Zhang)

noises and disguises.

Keywords:
Bio-inspired feature learning, automatic subspace learning, dimension reduction, graph embedding.

1. Introduction

Like humans' natural sensors, diverse artificial sensors such as cameras capture huge amount of high-dimensional information every second. This information is largely redundant in dimensionality and brings so-called curse-of-dimension challenge to researches in machine intelligence. To solve this problem, various dimension reduction or feature learning techniques have been proposed, which are inspired by the working way of human' brain.

Human's perception system can efficiently and effectively perceive constancy even though the raw sensory inputs are in flux. The biological studies [**?**] characterize the working way of human's perception as *manifold learning*, i.e., the high-dimensional data probably reside on a lower dimensional manifold and the manifold corresponds to the attractors in our brain [**?** **?**]. In the community of machine intelligence, the manifold can be regarded as an invariant characteristic of data that remains unchanged in different projection spaces.

The pioneer works in manifold learning with well-defined mathematic formulations are locally linear embedding (LLE) [**?**], ISOMAP [**?**], and Laplacian Eigenmaps [**?**]. These methods map a given set of high-dimensional data points into a low-dimensional space by preserving the geometric relations among data points. The geometric relations are always described as a similarity graph of which each vertex denotes a data points and the edge weight represents the similarity between two connected data points. Thus, manifold learning is also called as graph embedding method [**?**].

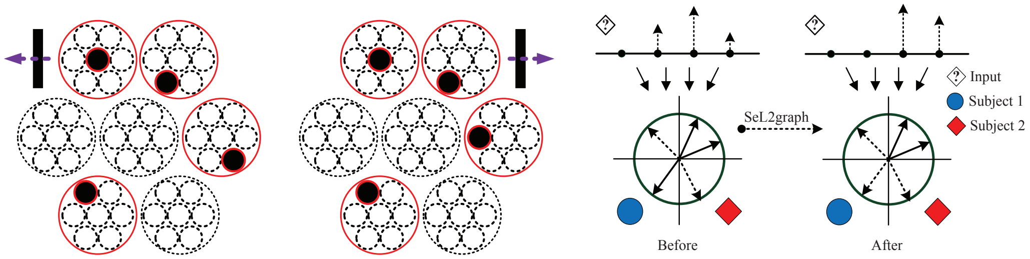
The main attraction of manifold learning is that it can handle linear inseparable data even though the data are sampled from multiple subspaces. One of the major problems is that it cannot handle the incremental data. To solve this problem, some subspace learning methods have been proposed, for example, neighborhood components analysis (NPE) [**?**], locality preserving projections (LPP) [**?**], and their extensions [**?** **?** **?**]. Unlike traditional manifold learning methods, subspace learning aims to learn a projection matrix $\mathbf{\Theta} \in \mathbb{R}^{m \times m'}$ instead of the low-dimensional features from a given data set. After obtaining the projection matrix, the low-dimensional features are obtained via $\mathbf{z} = \mathbf{\Theta}^T\mathbf{x}$, where $\mathbf{z} \in \mathbb{R}^{m'}$ is the low-dimensional feature and $\mathbf{x} \in \mathbb{R}^m$ is the raw data input.

The key of subspace learning is identifying the geometric relations among different data points, i.e., the construction of the similarity graph. A good similarity graph should only retain the connections among intra-class data points, i.e., only the data points with the same label are connected with each other. Recently, the methods using reconstruction coefficients to build a similarity graph have achieved impressive performance, including but not limited to sparsity preserving projections (SPP) [**?**], L1graph [**?**], low rank representation (LRR) [**?**], least square regression [**?**], L2graph [**? ?**], and their variants [**? ? ? ?**].

In these works, L2graph has achieved state-of-the-art performance in unsupervised subspace learning and clustering. The theoretical analysis and experimental studies [**? ?**] have shown that L2graph can achieve a good similarity graph even though the data set is grossly corrupted. Despite the advantages of L2graph, there are two disadvantages in L2graph based subspace learning. First, L2graph is a unsupervised method, which does not utilize the label information available in training data. Second, like other methods such as NPE, L2graph needs to specify the dimension of feature space. To find a optimal value for the feature dimension, most works have to search all possible values, for example, from one to the dimension of input (i.e., m). Clearly, such strategy is very time consuming, especially, in the scenario of high-dimensional data.

To solve these two problems in L2graph, we proposed a semi-supervised subspace learning method, namely, semi-supervised L2graph (SeL2graph). The method is inspired by the biological observations in [**?**], more specifically, the layer 2/3 of rat visual cortex activates the same collection of cortical columns in response to leftward and rightward drifting gratings (see Figs. 1(a) and 1(b)). We refer to such a property as *similar inputs having similar codes (SISC)* [**?**] and propose a mathematical formulation to the SISC by using a variant of fisher criterion. Fig. 1 gives an example to illustrate our basic idea. The contributions of this paper are summarized as follows:

- We propose a semi-supervised subspace learning method by enforcing the similar inputs have the similar features. The label information is used as the metric to determine whether a pair of samples are similar. Extensive experimental results show that SeL2graph outperforms



(a) (b) (c)

Figure 1: Motivation and illustration of our basic idea. (a-b) the biological observation reported in Ohki et al.'s work [**?**]. It shows that the visual cortex of rate represents similar inputs using similar codes (i.e., SISC), where the bigger circle denotes a cortical column consisting of multiple neurons. The cortical column will be activated if one of the affiliated neurons gives a spike (with slide lines). (c) For a given sample (e.g., the *diamond*), L2graph may use the samples from another subject (e.g., the *circle*) to represent it, where the dotted lines in the circle denote unactivated columns and the solid lines denote activated columns. By utilizing the SISC property, SeL2graph could rescale the projection coefficients over the *circle* to zeroes.

L2graph and other baseline methods with a considerable performance margin;

- Unlike most dimension reduction methods such as principle component analysis (PCA) and L2graph, SeL2graph can automatically determine the dimension of feature space. We prove that the feature dimension is bounded by the number of subjects. This advantage will significantly reduce the effort to tune the parameter;
- SeL2graph is robust to the Gaussian noise, random pixel corruptions, and various disguises over the face such as sunglasses.

Notations: we use **lower-case bold letters** to represent column vectors and **UPPER-CASE BOLD LETTERS** to represent matrices. $\mathbf{A}^T$, $\mathbf{A}^\dagger$, and $\mathbf{A}^{-1}$ denote the transpose, pseudo-inverse, and inverse of the matrix $\mathbf{A}$, respectively. $\mathbf{I}$ denotes the identity matrix. $\mathbf{A}_{ij}$ denotes an entry of the matrix $\mathbf{A}$. Table 1 summarizes some notations used throughout the paper.

Table 1: Some used notations.

Notations	Definitions
$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$	Training samples
$\mathbf{X}_i \in \mathbb{R}^{m \times n_i}$	The samples belonging to the i-th class
m	The dimension of the raw data
n_i	The size of the i-th class
m'	The dimension of feature space
k	The number of subjects
$\mathbf{z}_i \in \mathbb{R}^{m'}$	The low dimensional feature of $\mathbf{x}_i$
$\mathbf{W} \in \mathbb{R}^{n \times n}$	The similarity graph
$\mathbf{\Theta} \in \mathbb{R}^{m \times m'}$	The projection matrix

2. Related works

In this section, we briefly review two related topics, i.e., unsupervised and semi-supervised subspace learning.

2.1. Unsupervised Subspace Learning

A variety of unsupervised subspace learning methods have been proposed in recent years [? ? ? ? ? ? ? ? ? ? ? ?], and most of them can be unified into the framework of graph embedding.

The most well-known dimension reduction method may be PCA which aims to find a set of mutually orthogonal basis functions. If the data are drawn from a single linear separable subspace, PCA can recovery the subspace structure and obtain a compact representation. In practice, however, such conditions are hard to satisfy since most real world data are drawn from the union of multiple subspaces and these subspaces are linear inseparable.

To overcome the shortcomings of PCA, a lot of methods have been proposed, which use the reconstruction coefficient to build a similarity graph and embed the graph from the original space into a low-dimensional one. In these methods, different object functions have been proposed, leading to different similarity graph. To build a similarity graph, LPP [?] uses the Euclidean distance with the Heat kernel, NPE [?] uses locally linear reconstruction coefficients, SPP [?] and L1graph [?] use sparse representation [? ? ? ?

?], LRR [**?**] uses low rank representation, and L2graph [**? ?**] uses the thresholding linear representation.

Unsupervised subspace learning methods only utilize the geometric information and ignore the available supervised signals. When the adopted geometric metric cannot accurately describe the relations among data points, the obtained features may be undesirable.

2.2. Semi-supervised Subspace Learning

Recently, some semi-supervised subspace learning methods have been proposed, which integrate the label information with the geometric information for achieving a discriminative representation.

The key problem of semi-supervised subspace learning is exploring how to incorporate the label information into the projection space. There are two popular criterions: label propagation [**?**] and fisher criterion [**?**]. Label propagation assumes that the data points residing on the same manifold are very likely to have the same label. Based on this assumption, label propagation first builds a similarity graph using the geometric measurement such as the Euclidean distance; and then it propagates the labels from the labeled points to the unlabeled data points and rescales the connections weights in the graph. Under the framework of label propagation, some algorithms have been proposed and impressive results are achieved [**? ?**]. Fisher criterion aims to find directions on which the data points of different subjects (i.e., inter-class data points) are far from each other while enforcing data points of the same subject (i.e., intra-class data points) to be close to each other. Due to its simplicity and effectiveness, the fisher criterion has been used in a lot of applications [**? ? ? ? ? ? ? ? ?**].

3. Semi-Supervised Subspace Learning with L2graph

In [**? ?**], we proposed L2graph for robust subspace clustering and subspace learning. The method aims to solve the following optimization problem:

$$\min_{\mathbf{c}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|_2^2, \ \text{s.t.}\ \mathbf{e}_i^T\mathbf{c}_i = 0, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^m$ is a column vector of the training samples $\mathbf{X} \in \mathbb{R}^{m\times n}$, all entries in $\mathbf{e}_i$ are zeroes except the i-th entry is one, and $\mathbf{c}_i \in \mathbb{R}^n$ denotes the self-expression of $\mathbf{x}_i$ over $\mathbf{X}$. The optimal solution of Eq. (1) is given by

$$\mathbf{c}_i^* = \mathbf{P}\left[\mathbf{X}^T\mathbf{x}_i - \frac{\mathbf{e}_i^T\mathbf{Q}\mathbf{x}_i\mathbf{e}_i}{\mathbf{e}_i^T\mathbf{P}\mathbf{e}_i}\right], \tag{2}$$

where

$$\mathbf{P} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}, \tag{3}$$

$$\mathbf{Q} = \mathbf{P}\mathbf{X}^T. \tag{4}$$

By proving that the coefficients with small values are the projection over errors, the robustness of L2graph can be improved by performing a hard thresholding operator $\mathcal{H}_{\hat{k}}(\cdot)$ over $\mathbf{c}_i^*$, i.e., $\hat{\mathbf{c}_i^*} = \mathcal{H}_{\hat{k}}(\mathbf{c}_i^*)$, where $\mathcal{H}_{\hat{k}}(\cdot)$ keeps $\hat{k}$ largest entries in $\mathbf{c}_i^*$ and returns the others to zeros.

After obtaining the reconstruction coefficients of the whole training set, L2graph builds a similarity graph using the collection of $\hat{\mathbf{c}_i^*}$ and embeds the graph into a low-dimensional space. Note that, it is unnecessary to require the similarity graph to be symmetric in our methods.

L2graph is a unsupervised method, which only considers the geometric relations among data points and does not utilize the label information. Some recent works have shown that the integration of geometric relations and label information can significantly improve the performance of algorithms, for example, semi-supervised discriminant analysis (SDA) [**?**] is a semi-supervised extension of LPP [**?**], Discriminant SPP (DSPP) [**?**] is an extension of SPP [**?**].

To further improve the discrimination of L2graph, we propose a semi-supervised extension of L2graph. The proposed method is inspired by a biological observation, i.e., similar inputs have similar codes (see introduction). To formulate this property into our objective function, we propose the following objective function,

$$\min_{\mathbf{\Theta}} \left\|\mathbf{\Theta}^T\mathbf{X} - \mathbf{\Theta}^T\mathbf{X}\mathbf{W}\right\|_F^2 + trace\left(\beta\mathcal{S}(\mathbf{Z}) - \gamma\mathcal{J}(\mathbf{Z})\right), \tag{5}$$

where β and γ are two balanced factors with nonnegative values and $\mathbf{Z} = \mathbf{\Theta}^T\mathbf{X}$ is the low-dimension representation of $\mathbf{X}$. Note that, only β needs to be specified by users and γ can be automatically learned by our algorithm (see Theorem 1).

The objective function Eq.(5) consists three terms that play different roles: 1) the first term preserves the linear reconstruction relations among different data points based on the scheme of manifold learning; 2) the second term $\mathcal{S}(\mathbf{Z})$ aims at enforcing the data points with the same label are closed each other in the projection space, i.e, the SISC property; and 3) the third term $\mathcal{J}(\mathbf{Z})$ is a contrastive term involving the relations among different subjects. This term is crucial. Simply minimizing the first two terms over the similar samples may lead to a collapsed solution as pointed out by Lecun et al. [**?**].

Let $\mathbf{z}_j$ be the low-dimensional feature of $\mathbf{x}_j$, we formulate the SISC property as follows:

$$\mathcal{S}(\mathbf{Z}) = \sum_{i=1}^{k} \left(\sum_{\mathbf{z}_j \in \mathbf{Z}_i} (\mathbf{z}_j - \mathbf{d}_i)(\mathbf{z}_j - \mathbf{d}_i)^T \right), \tag{6}$$

and

$$\mathcal{J}(\mathbf{Z}) = \sum_{i=1}^{k} n_i (\mathbf{d}_i - \mathbf{d})(\mathbf{d}_i - \mathbf{d})^T, \tag{7}$$

where k denotes the number of subjects, $\mathbf{Z}_i$ is the collection of features that belong to the i-th subject, n_i is the number of samples belonging to the i-th subject, and $\mathbf{d}_i$ and $\mathbf{d}$ are the mean vectors of $\mathbf{Z}_i$ and $\mathbf{Z}$, respectively. Note that, the above formulations can be regarded as a variant of the fisher criterion, where the classical fisher criterion is usually defined as the minimization of the trace ratio $trace(\mathcal{S}(\mathbf{X}))/trace(\mathcal{J}(\mathbf{X}))$.

From Eqs.(6) and (7), we can find that $\mathcal{S}(\mathbf{Z})$ is the summation of the Euclidean distance among intra-class features and $\mathcal{J}(\mathbf{Z})$ is the separability among different subjects. Both $\mathcal{J}$ and $\mathcal{S}$ are defined in the feature space. In short, we aim to minimize the distance among the features with the same label and simultaneously maximize the distances among the centers of different subjects. Clearly, this formulation is based on the SISC principle, wherein any two samples are "similar" if and only if they are with the same label (i.e., $\mathbf{z}_j \in \mathbf{Z}_i$).

To solve Eq.(5), we have the following theorem:

Theorem 1. *The optimal solution to Eq.(5) consists of m' leading eigenvectors of the following generalized Eigen decomposition problem*

$$\mathbf{X}\Big((\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T + \beta(\mathbf{I} - \mathbf{\Omega})\Big)\mathbf{X}^T\mathbf{\Theta} = \gamma\mathbf{X}(\mathbf{\Omega} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}^T\mathbf{\Theta}, \tag{8}$$

where $\mathbf{\Omega}$ is a $n \times n$ matrix whose entry $\mathbf{\Omega}_{ij}$ equals $1/n_k$ if the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the k-th subject, n_k denotes the size of the k-th subject, $\mathbf{1}$ is a n-dimensional vector with elements of 1, and γ is the corresponding eigenvalue.

Proof. As $\mathbf{z}_j = \mathbf{\Theta}^T \mathbf{x}_j$, $\mathcal{S}(\mathbf{Z})$ and $\mathcal{J}(\mathbf{Z})$ can be rewritten as the functions of $\mathbf{X}$, i.e.,

$$\begin{aligned}\mathcal{S}(\mathbf{Z}) &= \mathbf{\Theta}^T \sum_{i=1}^{k} \left(\sum_{\mathbf{x}_j \in \mathbf{X}_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T \right) \mathbf{\Theta} \\ &= \mathbf{\Theta}^T \mathcal{S}(\mathbf{X}) \mathbf{\Theta} \qquad (9)\end{aligned}$$

and

$$\begin{aligned}\mathcal{J}(\mathbf{Z}) &= \mathbf{\Theta}^T \sum_{i=1}^{k} n_i (\mu_i - \mu)(\mu_i - \mu)^T \mathbf{\Theta} \\ &= \mathbf{\Theta}^T \mathcal{J}(\mathbf{X}) \mathbf{\Theta}, \qquad (10)\end{aligned}$$

where μ_i and μ are the mean vectors of $\mathbf{X}_i$ and $\mathbf{X}$, respectively. $\mathcal{S}(\mathbf{X})$ can be rewritten as follows:

$$\begin{aligned}\mathcal{S}(\mathbf{X}) &= \sum_{i=1}^{k} \left(\sum_{\mathbf{x}_j \in \mathbf{X}_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T \right) \\ &= \sum_{i=1}^{k} \Big(\sum_{\mathbf{x}_j \in \mathbf{X}_i} \mathbf{x}_j \mathbf{x}_j^T - n_i \mu_i \mu_i^T \Big) \\ &= \sum_{i=1}^{k} \mathbf{X}_i (\mathbf{I} - \frac{1}{n_i} \mathbf{1}_i \mathbf{1}_i^T) \mathbf{X}_i^T \qquad (11)\end{aligned}$$

where $\mathbf{1}_i$ is a n_i-dimensional vector with the elements of 1.

Let $\mathbf{\Omega} \in \mathbb{R}^{n \times n}$ be a symmetric matrix and its entry $\mathbf{\Omega}_{ij}$ is defined as follows:

$$\mathbf{\Omega}_{ij} = \begin{cases} \dfrac{1}{n_k} & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the } k\text{-th subject} \\ 0 & \text{otherwise.} \end{cases} \qquad (12)$$

Then, we have

$$\mathcal{S}(\mathbf{X}) = \mathbf{X}(\mathbf{I} - \mathbf{\Omega})\mathbf{X}^T. \qquad (13)$$

Similarly, we have

$$\mathcal{J}(\mathbf{X}) = \mathbf{X}(\mathbf{\Omega} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}^T, \qquad (14)$$

Clearly, both $\mathcal{S}(\mathbf{X})$ and $\mathcal{J}(\mathbf{X})$ are symmetric matrices. Denoting

$$\mathcal{L} = \|\mathbf{\Theta}^T\mathbf{X} - \mathbf{\Theta}^T\mathbf{X}\mathbf{W}\|_F^2 + trace(\beta\mathcal{S}(\mathbf{Z}) - \gamma\mathcal{J}(\mathbf{Z})), \tag{15}$$

it gives that

$$\frac{\partial\mathcal{L}}{\partial\mathbf{\Theta}} = 2\mathbf{X}(\mathbf{I}-\mathbf{W})(\mathbf{I}-\mathbf{W})^T\mathbf{X}^T\mathbf{\Theta} + 2\beta\mathcal{S}(\mathbf{X})\mathbf{\Theta} - 2\gamma\mathcal{J}(\mathbf{X})\mathbf{\Theta}. \tag{16}$$

Substituting Eq.(13)–(14) into Eq.(16) and letting $\frac{\partial\mathcal{L}}{\partial\mathbf{\Theta}} = 0$, we have

$$\mathbf{X}\Big((\mathbf{I}-\mathbf{W})(\mathbf{I}-\mathbf{W})^T + \beta(\mathbf{I}-\mathbf{\Omega})\Big)\mathbf{X}^T\mathbf{\Theta} = \gamma\mathbf{X}(\mathbf{\Omega} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}^T\mathbf{\Theta} \tag{17}$$

as desired. □

Almost all the subspace learning methods have also confronted with the parameter selection problem, especially, it is hard to determine the dimension of feature space (i.e., m'). Thus, most works solve this problem by experimentally tuning all possible values of m'. Clearly, this is very time costing in practice. In this paper, we will show that our method can automatically determine the value of m' with the following theoretical result:

Corollary 1. *Suppose the data set $\mathbf{X}$ is drawn from k subspaces, the feature dimension m' is bounded by the number of subjects, i.e.,*

$$m' \leq k. \tag{18}$$

Proof. Without loss of generality, we assume that the data $\mathbf{X}$ have been preprocessed by subtracting the mean vector from all samples. As a result, $\mu = 0$. By Theorem 1, we can find that the optimal solution to Eq.(5) will consist of m' leading eigenvectors of the following generalized Eigen decomposition problem:

$$\mathbf{X}\Big((\mathbf{I}-\mathbf{W})(\mathbf{I}-\mathbf{W})^T + \beta(\mathbf{I}-\mathbf{\Omega})\Big)\mathbf{X}^T\mathbf{\Theta} = \gamma\mathbf{X}\mathbf{\Omega}\mathbf{X}^T\mathbf{\Theta}. \tag{19}$$

Moreover, we assume $\mathbf{X}$ is sorted according to its labels, i.e., $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_1, \cdots, \mathbf{X}_k]$. Thus, $\mathbf{\Omega}$ will be a block-diagonal matrix in the form of

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Omega}_k \end{bmatrix}, \tag{20}$$

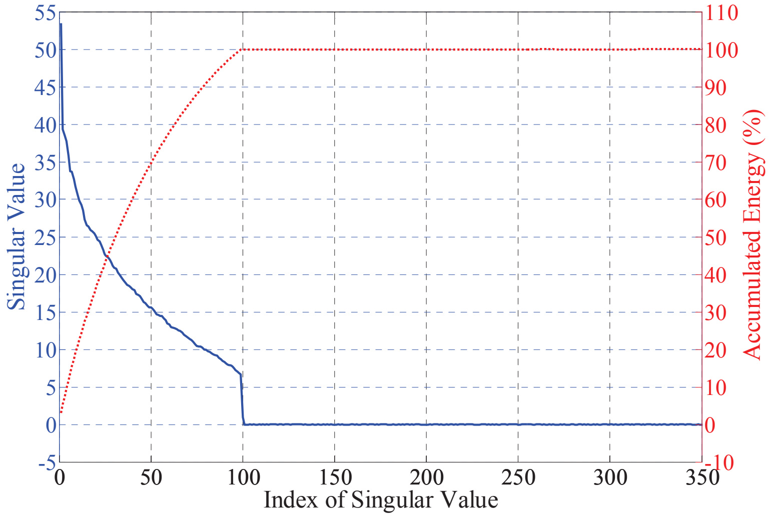


Figure 2: Eigenvalues of Eq.(10). A subset of the AR database distributed 100 subjects are used. The left y-axis is the value of the i-th eigenvalue and the right one is the accumulated energy of eigenvalues.

where the entries of the sub-matrix $\mathbf{\Omega}_i$ equal n_i and n_i is the number of the samples belonging to the i-th subject.

Therefore, the rank of $\mathbf{\Omega}$ is k and thus Eq.(19) has k nonzero eigenvalues at most, which gives the result $m' \leq k$. □

To verify the correctness of our theoretical result, we carry out experiments by using 700 clean AR images [**?**] sampled from 100 individuals. Figure 2 shows all the eigenvalues of Eq.(19). One can find that there are only 100 nonzero eigenvalues, which is matching with the ground truth and our theoretical result.

Algorithm 1 summarizes the proposed method. In Steps 1–4, SeL2graph computes the thresholding linear reconstruction coefficients as the geometric information. In Steps 5–6, the method learns a projection matrix by enforcing the low-dimensional features to satisfy the SISC property.

4. Experimental Results

In this section, we investigate the performance of SeL2graph using 15 different data sets, 7 subspace learning algorithms, and 3 different classifiers. We mainly consider the performance of SeL2graph with respect to

Algorithm 1 Semi-supervised Subspace Learning with L2-Graph

```
Input: A given data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, a new coming datum $\mathbf{y} \in span\{\mathbf{X}\}$,
    the regularization parameter $\lambda$, the thresholding parameter $\hat{k}$, and the
    discriminative parameter $\beta = 0.1$.
Calculate $\mathbf{P}$ and $\mathbf{Q}$ as in (3) and (4), and store them.
For each point $\mathbf{x}_i$, obtain its representation $\mathbf{c}_i$ via (2).
For each $\mathbf{c}_i$, eliminate the effect of errors in the projection space via
    $\hat{\mathbf{c}}_i = \mathcal{H}_{\hat{k}}(\mathbf{c}_i)$, where the hard thresholding operator $\mathcal{H}_{\hat{k}}(\mathbf{c}_i)$ keeps $\hat{k}$ largest
    entries in $\mathbf{c}_i$ and zeroizes the others.
Construct a similarity graph by $\mathbf{W}_{ij} = |\hat{\mathbf{c}_{ij}}| + |\hat{\mathbf{c}_{ji}}|$ and normalize each
    column of $\mathbf{W}$ to have a unit $\ell_2$-norm, where $\hat{\mathbf{c}_{ij}}$ is the $j$th entry of $\hat{\mathbf{c}}_i$.
Build the matrix $\mathbf{\Omega} \in \mathbb{R}^{n \times n}$ based on the label information as in Eq.(12).

Calculate the projection matrix $\mathbf{\Theta} \in \mathbb{R}^{m \times k}$ that consists of $k$ leading
    eigenvectors of Eq.(8) or Eq.(19).
Output: The projection matrix $\mathbf{\Theta}$ and the low-dimensional representation
    of $\mathbf{y}$ (i.e., $\mathbf{z} = \mathbf{\Theta}^T\mathbf{y}$).
```

three aspects: 1) accuracy, 2) computational efficiency, and 3) robustness to noises and real disguises. The code of our method can be downloaded from http://www.machineilab.org/users/pengxi/.

4.1. Experimental Setting and The Used Data Sets

We compare SeL2graph with seven well-known subspace learning methods including LPP [**?**], NPE [**?**], SPP [**?**] or called L1graph [**?**], LRR [**?**], L2graph [**?** **?**], SDA [**?**], and DSPP [**?**]. We download the codes of all the baseline algorithms except SPP and DSPP from the authors' websites. SPP and DSPP are based on sparse representation, and the original codes are based the *CVX* package [**?**] which is very inefficient. To accelerate the computing speed of SPP and DSPP, we implement them using a faster ℓ_1-solver (i.e., the Homotopy algorithm [**?**]).

After extracting the low-dimensional features using the evaluated subspace learning methods, we perform classifications over the features with three classifiers, i.e., *sparse representation based classifier* (SRC) [**?**], *support vector machine with linear kernel* (SVM) [**?**], and *the nearest neighbor classifier* (NN).

For fair comparisons, we tune the parameters of all the evaluated methods

Table 2: The used databases. s and n_i denote the number of subjects and the number of images for each group.

Databases	k	n_i	Original Size	Downsize
AR1	100	14	165 × 120	55 × 40
AR2	100	12	165 × 120	55 × 40
AR3	100	12	165 × 120	55 × 40
Yale	15	11	32 × 32	32 × 32
ExYaleB	38	58	192 × 168	54 × 48
MPIE-S1	249	14	100 × 82	55 × 40
MPIE-S2	203	10	100 × 82	55 × 40
MPIE-S3	164	10	100 × 82	55 × 40
MPIE-S4	176	10	100 × 82	55 × 40
USPS	10	1100	16 × 16	16 × 16
FERET	200	7	80 × 80	80 × 80

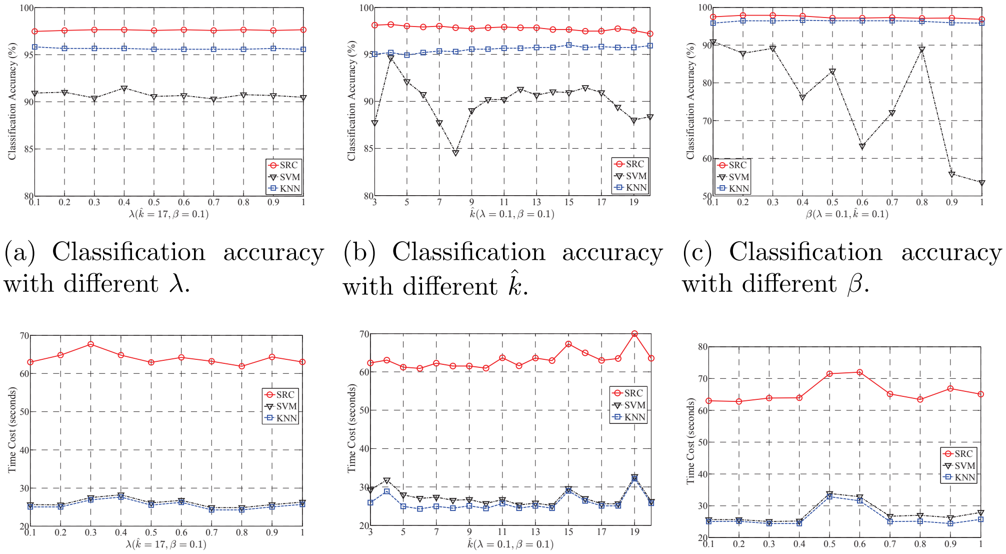
for achieving their best performance and report the mean accuracy and time cost in 10 repeating tests. The unsupervised subspace learning methods and its semi-supervised extensions have shared the same tuned parameters, for example, the parameter λ of SeL2graph and L2graph are with the same value. Moreover, we fix the dimension of feature space using the number of subjects. This strategy not only avoids the time cost to identify the optimal feature dimension, but also provides a fair measurement to the compactness of features.

The used data sets include three subsets of AR facial images [?], the Yale facial database (Yale) [?], the Extended Yale facial database B (ExYaleB) [?], four sessions of CMU Multiple PIE (MPIE) [?], the USPS handwritten digital database[1], and the FERET facial images [?]. Table 2 gives an overview on the used data sets of which some data sets are downsized for computational efficiency.

4.2. The Influence of Parameters

In this section, we investigate the performance of SeL2graph using a subset of the Extended Yale database B. SeL2graph has three user specified parameters, i.e., the regularization parameter λ, the thresholding parameter $\hat{k}$, and the discriminant term parameter β. In each test, we change the value of one of these three parameters and report the mean classification accuracy

[1] http://www.cs.nyu.edu/ roweis/data.html.



(a) Classification accuracy with different λ.

(b) Classification accuracy with different $\hat{k}$.

(c) Classification accuracy with different β.

(d) Time cost with different λ.

(e) Time cost with different $\hat{k}$.

(f) Time cost with different β.

Figure 3: The influence of different parameters of SeL2graph on the Extend Yale database B, where the training samples and the testing samples are with the equal size.

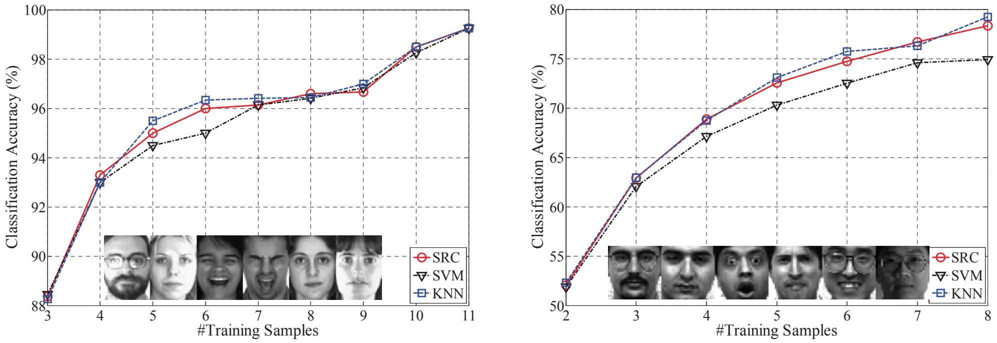
and time cost.

Figure 3 reports the results from which we have the following observations:

- The parameter λ has little influence on the classification accuracy. The recognition rates of SRC, SVM, and NN almost remain unchanged. The accuracy of SRC ranges from 97.46% to 97.64%, that of SVM ranges from 90.29% to 91.47%, and that of NN ranges from 95.55% to 95.83%;
- With the increase of $\hat{k}$ and β, SVM achieves a lower accuracy. In most cases, SRC achieves the highest recognition rates, but it is the most inefficient. In general, SVM and NN are two times at least faster than SRC;
- The NN classifier finds a good balance between the computational efficiency and recognition accuracy.

4.3. Performance with Increasing Training Data

In this section, we report the performance of SeL2graph with increasing labeled samples. In the experiment, we use two facial image databases, i.e.,



(a) The AR1 data set.

(b) The Yale data set.

Figure 4: The performance of SeL2graph with increasing training samples. On the x-axis, some sample images are illustrated.

AR1 and Yale.

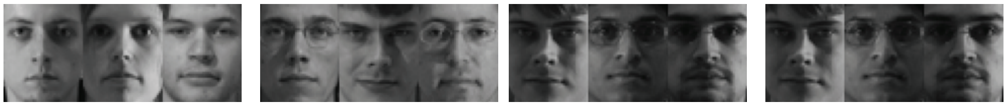
AR1 consists of 1400 clean images that are uniformly sampled from 100 individuals (50 males and 50 females). In the experiment, we randomly select 3, 4, 5, 6, 7, 8, 9, 10, and 11 images from each subject for training and use the rest of images for testing.

The Yale data set contains of 165 images over 15 subjects. Each subject includes 11 samples that are different in facial expression or configuration: center-light, wearing glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. We follow Cai et al.'s testing protocol by using their data partitions. More specifically, they randomly selected p samples as training set and used the rest of database as the testing set, where p increases from 2 to 8 with an interval of 1. For each given p, there are 50 randomly splits.

Fig.4 shows that SeL2graph with three classifiers perform better when more labeled sample are available. Moreover, the SRC and the NN are superior to the SVM in the tests.

4.4. Performance with Different Classifiers

In this section, we compare SeL2graph with seven popular subspace learning methods using four subsets of MPIE that captured under four different sessions. In the experiments, we use all the frontal faces with 14 illumina-



(a) MPIE-S1 (b) MPIE-S2 (c) MPIE-S3 (d) MPIE-S4

Figure 5: Sample images from four sessions of MPIE.

tions[2], where Fig. 5 shows some samples. On each data set, we randomly split it into two parts with the equal size for training and testing. Tables 3–7 report the results, from which we can find that:

- On all the used data sets, SeL2graph achieves the highest recognition rates, while keeping a competitive computational efficiency. For example, SeL2graph is about 8.69 times faster than DSPP on the MPIE-S1, and its accuracy is 1.34% higher than that of DSPP when the SRC is used;

- With different subspace learning methods, the SRC classifier usually achieves the highest recognition rate and the NN is the second best classifier. However, the SRC is significantly slower than the NN and the SVM. On MPIE-S2 and MPIE-S4, SDA, DSPP, and SeL2graph achieve the accuracy of 100% with the SRC;

- The label information remarkably improves the performance of feature learning methods. For example, SeL2graph (semi-supervised version) outperforms L2graph (unsupervised version) with a performance margin of 27.48%, 2.86%, 4.52%, and 1.48% on these four data sets when the NN is used.

- The computational efficiency of L2graph and SeL2graph are very close. Although these two methods solve two different Eigen decomposition problems, the computational complexity of these two problems are the same.

Table 3: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with three classifiers on **the MPIE-S1 data set**, where $m' = 249$.

Classifiers	SRC		SVM		NN	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	29.43	59.78	17.97	13.48	25.40	4.46
NPE	28.86	61.15	15.16	25.34	23.24	9.53
SPP	56.58	1101.88	16.71	875.77	52.13	822.43
LRR	61.44	62.38	25.59	532.86	61.25	9.52
L2graph	70.28	97.51	25.07	63.57	71.60	47.48
SDA	99.13	62.11	99.20	7.72	97.31	4.79
DSPP	98.60	822.98	81.81	615.05	97.02	611.27
SemiL2graph	**99.94**	94.66	**99.66**	49.31	**99.08**	45.11

Table 4: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with three classifiers on **the MPIE-S2 data set**, where $m' = 203$.

Classifiers	SRC		SVM		NN	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	43.15	29.17	19.11	5.09	26.11	1.31
NPE	43.35	27.12	12.02	7.82	27.49	2.73
SPP	77.59	546.89	56.80	420.38	68.87	372.39
LRR	85.52	26.03	71.38	131.72	89.67	2.46
L2graph	99.21	36.06	91.72	15.43	97.14	11.89
SDA	100.00	29.91	98.41	4.61	99.80	3.67
DSPP	100.00	349.70	94.95	247.61	96.95	246.45
SemiL2graph	**100.00**	34.72	**100.00**	14.88	**100.00**	13.69

4.5. Subspace Learning on Clean Images

In this section, we investigate the performance of the tested subspace learning methods on three facial images data sets and one handwritten digital data set (Fig. 6 illustrates some sample images). For all the used data sets except the FERET, we use a half of samples per subject for training and the rest of the samples for testing. In other words, both the training sets and testing sets include 700 AR1 samples, 1102 ExYaleB samples, or 5500 USPS images. For FERET, we use 800 samples for training and 600 samples for testing. Table 7 reports our results from which we can find that:

- SeL2graph outperforms the other testing methods on all the tests. It finds a good tradeoff between recognition rate and computational efficiency. The difference in accuracy between SeL2graph and the baseline algorithms ranges from +1.43% to +50% on AR1, +2.64% to +62.71%

[2] illuminations: 0,1,3,4,6,7,8,11,13,14,16,17,18,19.

Table 5: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with three classifiers on **the MPIE-S3 data set**, where $m' = 164$.

Classifiers	SRC		SVM		NN	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	61.83	26.35	31.95	2.77	32.68	0.94
NPE	76.10	32.24	84.02	10.37	53.78	9.67
SPP	83.66	462.11	76.59	313.12	67.68	292.64
LRR	90.49	18.68	81.10	71.30	83.78	1.47
L2graph	99.02	24.87	90.98	9.55	95.24	6.88
SDA	99.01	23.07	98.88	3.76	98.39	3.29
DSPP	98.97	268.28	93.29	171.09	94.63	170.34
SemiL2graph	**99.88**	24.92	**99.88**	10.11	**99.76**	9.33

Table 6: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with three classifiers on **the MPIE-S4 data set**, where $m' = 176$.

Classifiers	SRC		SVM		NN	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	50.80	26.20	20.68	3.46	30.57	0.96
NPE	48.75	23.56	20.45	5.23	34.43	2.07
SPP	66.36	660.30	46.02	421.50	51.59	388.92
LRR	88.86	19.75	81.93	87.49	82.50	1.90
L2graph	99.77	27.03	92.61	10.12	98.52	8.12
SDA	100.00	23.84	98.93	3.90	99.31	3.25
DSPP	100.00	281.90	98.41	191.83	98.07	191.04
SemiL2graph	**100.00**	29.22	**100.00**	12.33	**100.00**	11.45

on the extended Yale database B, +1.56% to +61.67% , and +12.50% to +24.00% on FERET;

- The integration of label information and geometric information remarkably improves the discrimination of L2graph. More specifically, the performance gain in recognition rate between SeL2graph and L2graph on these four data sets are +21.43%, +31.58, +1.52%, and +14.17%, respectively. Note that, L2graph can achieve higher accuracy if we use more features (e.g., 300) as shown in our previous work [**?**].

4.6. Subspace Learning on Corrupted Facial Images

In this section, we investigate the robustness of SeL2graph against two noises using the ExYaleB data set. The noises include the white Gaussian noise (additive noise) and random pixel corruption (non-additive noise) [**?**]. Fig. 7 shows some sample images.

In the experiments, we randomly choose a half of samples (29 images per subject) to corrupt using these two noises. Specifically, we add the white

Table 7: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with the nearest neighbor classifier. m' denotes the feature dimension.

Data sets	AR1 ($m' = 100$)		ExYaleB ($m' = 38$)		USPS ($m' = 10$)		FERET ($m' = 200$)	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	46.86	0.75	33.03	1.88	89.87	2.43	14.67	2.75
NPE	74.71	9.38	85.66	26.66	89.80	7.74	17.83	3.07
SPP	53.71	250.09	21.51	486.02	89.71	185.93	22.17	668.41
LRR	60.57	12.61	40.83	3.37	61.67	0.88	23.33	1.69
L2graph	75.43	2.45	64.16	18.40	90.17	610.21	24.50	7.05
SDA	95.43	3.01	93.10	4.88	75.62	3.08	18.83	45.08
SemiSPP	89.00	148.26	92.20	354.01	90.13	101.23	26.17	668.41
SeL2graph	**96.86**	10.52	**95.74**	24.60	**91.69**	618.09	**38.67**	63.95

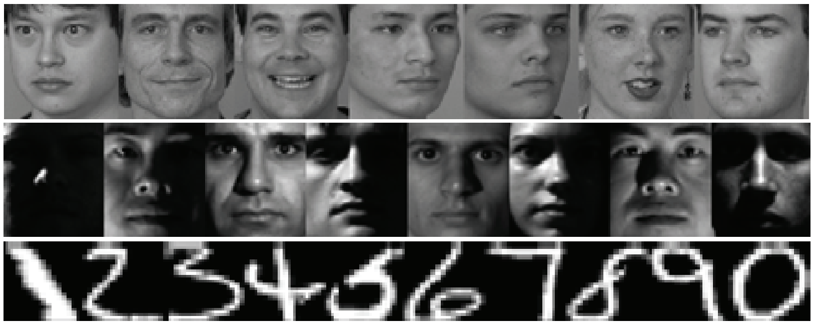


Figure 6: The used clean image databases. From top to bottom: FERET, the extended Yale database B, and USPS.

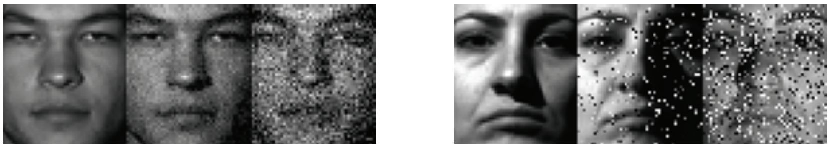
Gaussian noise into the sampled data $\mathbf{y}$ via $\tilde{\mathbf{y}} = \mathbf{y} + \rho\mathbf{e}$, where $\tilde{\mathbf{y}} \in [0\ 255]$, ρ is the corruption ratio, and $\mathbf{e}$ denotes the noise following the standard normal distribution. For random pixel corruption, we replace the value of a percentage of pixels randomly selected from the image with the values following a uniform distribution over $[0,\ p_{max}]$, where p_{max} is the largest pixel value of $\mathbf{y}$. After adding the noises into the images, we randomly divide the data into training and testing sets with equal size. In other words, both the training data and testing data probably contain corrupted samples. Table 8 shows that:

- SeL2graph is more robust than all the evaluated methods by a considerable performance margin. It is 9.25%, 9.98%, 11.07%, and 0.27% higher than the second best algorithms on these four data sets;

- The semi-supervised methods (i.e., SDA, DSPP, and SeL2graph) are more robust than the unsupervised methods (i.e., LPP, NPE, SPP,

Table 8: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with the nearest neighbor classifier, where 38 features are extracted. RPC is the short of random pixel corruption. The numbers in the parentheses are the values of the noise level ρ.

Corruptions	Gaussian Noise (10%)		Gaussian Noise (30%)		RPC (10%)		RPC (30%)	
Algorithms	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
LPP	39.47	1.30	28.58	1.60	28.58	1.60	19.69	1.60
NPE	30.22	2.91	24.05	2.81	24.05	2.81	17.70	2.75
SPP	29.05	549.16	16.52	526.11	17.60	512.81	14.79	490.78
LRR	33.97	1.63	23.97	1.63	28.13	2.48	12.75	2.40
L2graph	46.82	18.18	34.94	18.20	31.58	18.20	28.31	18.20
SDA	79.13	3.26	57.26	2.90	60.25	3.09	49.82	3.19
DSPP	77.59	339.47	66.15	337.73	59.26	348.61	53.45	343.67
SeL2graph	**88.38**	28.10	**76.13**	27.45	**71.32**	26.46	**53.72**	5.88



(a) The Gaussian Corruption.

(b) Random Pixel Corruption.

Figure 7: Sample images from the corrupted ExYaleB images. From left to right in each subfigure, the corruption rates are 0%, 10%, and 30%, respectively.

LRR, and L2graph);

- The non-additive noise (i.e., the random pixel corruption) is more challenging than the additive noise (i.e., the Gaussian noise). All the tested methods perform better in the former corruption. Moreover, with the increase of the corruption level, all the evaluated methods achieve a lower accuracy.

4.7. Subspace Learning on Disguised Facial Images

In this section, we evaluate the robustness of SeL2graph to the real disguises using two subsets of the AR database, i.e., AR2 and AR3 (see Fig. 8). Both AR2 and AR3 consist of 600 clean and 600 disguised facial images. The difference between them is that the AR2 images are disguised by sunglasses and the AR3 images are disguised by scarves. The occluded rates of these two different disguises are about 20% and 40%, respectively. Similar to the tests on the corrupted images, both the training and testing set are

Table 9: The classification accuracy (%) and time cost (seconds) of different subspace learning algorithms with the nearest neighbor classifier, where 100 features are extracted.

Disguises	AR2 (Glasses)		AR3 (Scarves)	
Algorithms	Accuracy	Time	Accuracy	
LPP	24.50	0.54	19.83	0.56
NPE	38.00	8.15	27.00	9.29
SPP	14.79	490.78	12.50	187.40
LRR	49.04	2.45	50.17	0.88
L2graph	56.33	2.85	57.83	2.84
SDA	88.50	3.01	87.83	3.32
DSPP	76.83	116.50	80.17	117.03
SeL2graph	**91.00**	8.77	**92.00**	10.88

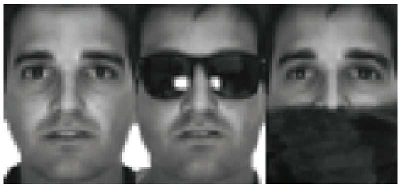


Figure 8: Some sample images from AR2 (disguised by sunglasses) and AR3 (disguised by scarves).

with equal size and probably contain disguised samples. Table 9 reports the comparisons from which we have the following observations:

- SeL2graph is 2.5% and 4.17% at least higher than the other tested methods on AR2 and AR3, respectively. Moreover, the recognition rates achieved by all the tested methods are very close even though the occluded rates are largely different in these two data sets. This verifies a claim in face recognition, i.e., eye and chin are with different discrimination;

- The semi-supervised methods again show their effectiveness. The worst semi-supervised method (DSPP) outperforms the best unsupervised method (L2graph) with the accuracy gain of 20.50% and 22.34% on these two different disguises;

- SPP and DSPP are very inefficient because they both require solving an ℓ_1-minimization problem whose computational complexity is proportional to the cube of the data size at least.

5. Conclusion

In this paper, we proposed a semi-supervised subspace learning method, i.e., SeL2graph. The method is motivated by the biological observation of similar inputs having similar codes. In our objective function, this property is mathematically formulated by minimizing the Euclidean distance among intra-class data points and maximizing the Euclidean distance among the centers different subjects. Extensive experimental studies have shown the effectiveness, robustness, and computational efficiency of our method.

There are no perfect algorithm and the SeL2graph is not an exception. The objective function of SeL2graph consists of three terms of which the last two terms are with a good biological interpretation. However, the first term is designed based on the scheme in machine learning and computer vision, which reconstructs the input using the linear combination of a given data set. Although such scheme has shown the effectiveness in practice and good intuition in mathematics, it lacks interpretation in biology. Therefore, some works may be conducted to fill this gap in future.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that significantly improve the quality of this paper. This work was supported by the A-STAR SERC under grant No.1225100002 and National NFSC under grant No.61432012, No. 61322203, and No.61332002