2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Label Propagation and Soft-similarity Measure for Graph based Constrained Semi-Supervised Learning

Zhao Zhang Mingbo Zhao Tommy W.S. Chow

Abstract— **This paper discusses a new setting of graph based semi-supervised learning (SSL) guided using pairwise constraints (PCs). Technically, we propose a novel Graph based Constrained Semi-Supervised Learning (G-CSSL) framework. In this setting, PCs are used to specify the types (intra- or inter-class) of points with labels. Because the number of labeled data is typically small in SSL setting, the core idea of this framework is to create and enrich the PCs sets using the propagated soft labels from both labeled and unlabeled data via special label propagation (SLP), and hence obtaining more supervised information for delivering enhanced learning performance. To obtain the predicted labels of unlabeled data, we calculate the sparse codes of all data vectors jointly to assign weights for SLP. To deliver enhanced inter-class separation and intra-class compactness, we also present a mixed soft-similarity measure to evaluate the similarity/dissimilarity of constrained sample pairs by using the sparse codes and outputted probabilistic values by SLP. Extensive simulations demonstrated the effectiveness of our G-CSSL for image representation and recognition, compared with other related SSL techniques.**

Keywords- Label propagation; soft-similarity measure; sparse coding; constrained semi-supervised learning; subspace learning

I. INTRODUCTION

Labeled data is always expensive to achieve and the labeling process by humans is also costly, while unlabeled data can be readily available with low expense from real world, leading to considerable interests and lots of efforts on the study of semi-supervised learning (SSL) [1][35]. The objective of SSL is to enhance the performance by using supervised information of labeled data and their relationships to unlabeled samples [1].

Based on the clustering and manifold assumptions [1][12] [17], recent years have witnessed lots of efforts on the graph based SSL (G-SSL)[6][9-11][13-19][24-25][28-29][31-35] for its validity via considering the intrinsic geometrical structure inferred from both labeled and unlabeled data. G-SSL can be broadly divided into *transductive* and *inductive*. The inductive setting is mainly for classification based (either local or global) dimensionality reduction [28-29][31-34]. But local techniques often involve the step of estimating optimal neighbor number k and kernel width, which is challenging in reality. The other setting is label propagation (LP) that propagates the labels of labeled data to unlabeled data according to the distribution of both labeled and unlabeled data [9-10][13-19]. Three popular LP methods include the harmonic function approach [16], the consistency method [17] and recent special label propagation (SLP) [19]. By comparing with the consistency and harmonic function methods, SLP can not only well detect outliers in data, but also output the labels as probabilistic values [18].

The core step of LP is to construct an edge weight matrix W to measure the similarities between vertices in a faithful graph. A graph is usually constructed by finding the neighbors using k- or ε -neighborhood [12]. One most popular weight assignment for LP is the Gaussian kernel similarity [17], but estimating an optimal kernel width is difficult [9][15]. *Linear Neighborhood Propagation* (LNP) [9] was recently proposed to first approximate the whole graph by a series of overlapped linear neighborhood patches and the edge weights in each patch are then computed by applying the neighborhood linear projection. But LNP also suffers from the issue of setting fixed neighborhood size for each vertex and there is no reliable way to determine optimal k number. More recently, to achieve adaptive neighborhood for weight construction, sparse coding (SC) based formulations have attracted many interests (such as [13-15][18]). For a given data matrix $X=[x_1,x_2,...,x_N]\in\mathbb{R}^{n\times N}$, a similar idea is to compute the (non-negative) sparse codes for each x_i individually using l_1-norm minimization with X or the pre-calculated sparse neighbors of x_i as the dictionary [13-15] [18]. To capture the global structures of data, in this paper we calculate the nonnegative sparse codes of all sample vectors *jointly* by optimizing a l_1-norm based minimization problem.

In typical SSL settings, the number of labeled data is often small, so it will be greater advantageous to apply the pairwise constraints (PCs) than the class labels to reflect supervised information of samples, since PCs can be obtained by minimal effort and can provide more supervised information if there are enough samples with labels available [28-31]. But if the labeled number is too limited, the advantages of PCs over the class labels will not exist any more. To address the insufficient data labeling problem, we construct the PCs sets based on the propagated soft labels from both labeled and unlabeled data through SLP in this paper. More specifically, we propose an adaptive neighborhood based SLP process induced pairwise constrained SSL framework, called *Graph based Constrained Semi-Supervised Learning* (G-CSSL), for feature extraction and classification. To achieve the adaptive neighborhood, the sparse codes are employed to assign the edge weights in SLP process of our framework. In addition, to deliver the enhanced inter-class separation and intra-class compactness, based on the outputted probabilistic values by SLP and the sparse codes over both labeled and unlabeled data, we also propose a voting strategy based *Mixed Soft-similarity Measure* (MSM) method for evaluating the similarity/dissimilarity of data pairs in PCs sets, with the weight values determined using the SLP process and nonnegative SC process at the same time.

Z. Zhang is now with the School of Computer Science and Technology, Soochow University, Suzhou 215006, P. R. China; also with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (email: cszzhang@gmail.com)

M. B. Zhao and Tommy W.S. Chow are with the Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR (emails: mzhao4@cityu.edu.hk, eetchow@cityu.edu.hk)

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

The paper is outlined as follows. Section II introduces the adaptive neighborhood by sparse coding for SLP. In Section III, we propose the G-CSSL framework and the MSM method. Section IV describes the settings and tests our technique using real datasets. Finally, the paper is concluded in Section V.

II. ADAPTIVE NEIGHBORHOOD BASED SLP

In this section, we show the l_1-norm minimization problem for calculating the sparse codes to assign weights for SLP. Then we define the PCs sets based on the predicted soft labels by SLP. Let $X=\left[X_L,X_U\right]\in\mathbb{R}^{n\times(l+u)}$, where $X_L=\left[x_1,x_2,...,x_l\right]\in\mathbb{R}^{n\times l}$ is a labeled set and $X_U=\left[x_{l+1},x_{l+2},...,x_{l+u}\right]\in\mathbb{R}^{n\times u}$ is unlabeled set. We assume that there are *c* classes and all the classes are present in the labeled set. Each point in X_L is associated with a class label $l(x_i)$ in $\{1,2,\ldots,c\}$. SLP is a transductive process that propagates label information of X_L to X_U [19].

A. Sparse Coding (SC) for Weight Assignments

Given a set of data vectors $x_i, i=1,2,...,N$, SC represents each x_i by using as few points that most compactly expresses x_i from $X=\left[x_1,x_2,...,x_N\right]\in\mathbb{R}^{n\times N}$ as possible [20-23]. By setting X itself as the dictionary, this paper calculates the nonnegative sparse codes of all vectors $\{x_i\}_{i=1}^N$ jointly from the following criterion:

$$\underset{S,E}{Min}\ \|S\|_1+\lambda\|E\|_1,\ \text{Subj}\ X=XS+E,\ Diag(S)=0,\ S>0,\ e^{\mathrm{T}}S=e^{\mathrm{T}},\quad(1)$$

where SC enforces $Diag(S)=0$ to avoid the trivial solution $S=I$ (where I is an identity matrix), $\|\bullet\|_1$ is the l^1-norm, i.e., $\|E\|_1=\sum_{i,j}|E_{i,j}|$, $e\in\mathbb{R}^N$ denotes a column vector of all ones, λ is a positive parameter, $X-XS$ identifies the errors *E*, l^1-norm is imposed on *E* to fit the random corruptions [2], and the sum-to-one constraint $e^{\mathrm{T}}S=e^{\mathrm{T}}$ can enable the solution to reflect certain local properties of data [20]. For efficiency, in this paper we use the inexact Augmented Lagrange Multiplier (ALM) method [3] to solve the above problem. Firstly, it can be transformed to the following equivalent one:

$$\begin{aligned}&(Q,S,E)=\arg\min_{Q,S,E}\|Q\|_1+\lambda\|E\|_1\\&\text{Subj}\ S=Q,\ X=XS,\ diag(S)=0,\ S>0,\ e^{\mathrm{T}}S=e^{\mathrm{T}}\end{aligned}.\quad(2)$$

The augmented Lagrangian function of the above problem can be addressed as

$$\begin{aligned}\hat{J}(Q,S,E,Y_1,Y_2,\mu)=&\|Q\|_1+\lambda\|E\|_1+\langle Y_1,X-XS\rangle+\\&\langle Y_2,S-Q\rangle+\frac{\mu}{2}\left(\|X-XS\|_F^2+\|S-Q\|_F^2\right)\end{aligned}.\quad(3)$$

We first solve *Q* by fixing *S*. When solving Q_{k+1} at the *(k+1)*-iteration, Y_2 and *S* are set to Y_2^k and S_k respectively. Thus Q_{k+1} can be inferred as

$$Q_{k+1}=\arg\min_{Q}(1/\mu_k)\|Q\|_1+(1/2)\left\|Q-\left(S_k+Y_2^k/\mu_k\right)\right\|_F^2,\quad(4)$$

which can be effectively solved by the shrinkage operator [3]. Then we can infer the solution of S_{k+1} as

$$S_{k+1}=\left(X^{\mathrm{T}}X+I\right)^{-1}\left[Q_{k+1}+X^{\mathrm{T}}X-X^{\mathrm{T}}E+\left(X^{\mathrm{T}}Y_1^k-Y_2^k\right)/\mu_k\right],\quad(5)$$

where X^{T} denotes the transpose of X and $\left(X^{\mathrm{T}}X+I\right)^{-1}$ is the inverse of matrix $X^{\mathrm{T}}X+I$. After Q_{k+1} and S_{k+1} are obtained at the *(k+1)*-iteration, the sparse error term E_{k+1} can be computed as $E_{k+1}=\arg\min_E(\lambda/\mu_k)\|E\|_1+(1/2)\left\|E-\left(X-XS_{k+1}+Y_1^k/\mu_k\right)\right\|_F^2$, which can be similarly solved by the shrinkage operator [3]. The solution $S^*=\left[s_1^*,s_2^*,...,s_N^*\right]$ is the "sparsest presentation" of the original data, where s_i^* denotes the coefficient vector to reconstruct x_i and s_i^* is naturally sparse. Note that S^* can also be used to define the edge weight matrix W (i.e., $W=S^*$) of an undirected graph to represent the sparsity of the datasets and measure the similarity between points [27], i.e., heavy weights $W_{i,j}$ will be imposed to the edges connecting "close" vertices. The intrinsic discriminant information of samples can also be preserved by S^* because of the nature of sparse representation. Due to sparse representation, W has a natural discriminating power. To make a connection to the normalized graph, we symmetrize W as $W\leftarrow\left(W+W^{\mathrm{T}}\right)/2$ or $W_{i,j}\leftarrow\left(W_{i,j}+W_{j,i}\right)/2$. Then resembling [17], we normalize W as $\widehat{W}=D^{-1/2}WD^{-1/2}$ or $\widehat{W_{i,j}}=W_{i,j}/\sqrt{D_{ii}D_{jj}}$, where D with $D_{ii}=\sum_{j=1}^{l+u}W_{i,j}$ is a diagonal matrix. Note that this normalization can help strengthen the weights in low-density region and weaken the weights in high-density region, which is useful for handling the cases that the density of dataset varies dramatically [19].

B. Label Propagation via SLP over Adaptive Neighborhood

Based on the normalized weight matrix $\widehat{W}=D^{-1/2}WD^{-1/2}$, we can predict the labels of unlabeled samples using SLP. Denote by $Y=\left[y_1,y_2,\ldots,y_{l+u}\right]\in\mathbb{R}^{(c+1)\times(l+u)}$ the initial labels of all the samples. For the labeled sample x_i, $y_{i,i}=1$ if x_i belongs to the *i*-th class, otherwise $y_{i,i}=0$; for unlabeled data x_i, $y_{i,i}=1$ if $i=c+1$, otherwise $y_{i,j}=0$. Note that SLP adds an additional class $c+1$ to detect outliers, so the sum of each column of Y is 1 [19]. Also let $F=\left[f_1,f_2,\ldots,f_{l+u}\right]\in\mathbb{R}^{(c+1)\times(l+u)}$ denote the predicted soft label matrix, where f_i is a column vector with the entries satisfying $0\le f_{i,i}\le 1$, and the biggest $f_{i,i}$ in each column decides the class assignment of sample x_i.

Denote a stochastic matrix $\Xi=\widehat{D}^{-1}\widehat{W}$, where $\widehat{D}$ is a diagonal matrix with each element satisfying $\widehat{D_{ii}}=\sum_{j=1}^{l+u}\widehat{W_{i,j}}$. Then, we consider an iterative process for label propagation. At each iteration, SLP expects that the class label of each sample point is partially received from its neighborhoods and the rest is from its own label. Hence the label information of samples at the *(t+1)*-th iteration can be

$$F(t+1)=F(t)\Xi I_\alpha+YI_\beta,\quad(6)$$

where $I_\alpha\in\mathbb{R}^{(l+u)\times(l+u)}$ is a diagonal matrix with each input element being α_j, $I_\beta=I-I_\alpha$, $\alpha_j\left(0\le\alpha_j\le 1\right)$ is a parameter for sample x_j to balance the initial label information of x_j and the label information received from its neighbors during the iteration. According to [19], the regularization parameter α_j for the labeled sample x_j is set to α_l, and the parameter α_j for the unlabeled sample x_j is set to α_u in the simulations. Based on the above iterative process, we can have

$$F(t)=F(0)(\Xi I_\alpha)^t+YI_\beta\sum_{k=0}^{t-1}(\Xi I_\alpha)^k. \quad (7)$$

Based on the matrix properties, namely $\lim_{t\to\infty}(\Xi I_\alpha)^t=0$ and $\lim_{t\to\infty}\sum_{k=0}^{t-1}(\Xi I_\alpha)^k=(I-\Xi I_\alpha)^{-1}$, so the iterative process of SLP can converge to

$$F=\lim_{t\to\infty}F(t)=YI_\beta(I-\Xi I_\alpha)^{-1}. \quad (8)$$

Note that it can be easily proved that the sum of each column in F is equal to 1, which indicates that the elements in F are the probability values and $f_{i,j}$ can be treated as the posterior probability of x_j belonging to the *i*-th class. If $i=c+1$, $f_{i,j}$ represents the probability of x_j belonging to outliers. Based on the SLP process, the outliers in data can be detected and the soft labels of data can be obtained at the same time [19].

III. Graph based Constrained Semi-Supervised Learning (G-CSSL) Framework

The core idea of G-CSSL is to create and enrich the PCs sets using the propagated soft labels of samples by SLP. We also address a mixed soft-similarity measure (MSM) approach for similarity measurements in this section.

A. Traditional Constrained Learning Problem

In traditional pariwise constrained problem, for a given set of labeled data samples, a *Must-link* (ML) constraint set and a *Cannot-link constraint* (CL) set are constructed as

$$ML=\{(x_i,x_j)|l(x_i)=l(x_j)\},\ CL=\{(x_i,x_j)|l(x_i)\neq l(x_j)\}, \quad (9)$$

where $l(x_i)\in\{1,2,...,c\}$ is the class label of x_i and c is the class number. Then one aims at pushing data pairs $(x_i,x_j)\in ML$ close together in the reduced space by minimizing pairwise distances between them, and separating data pairs $(x_i,x_j)\in CL$ via maximizing their pairwise distances. So we can define the following maximum margin criterion [38] based problem:

$$\begin{aligned}\underset{T\in\mathbb{R}^{n\times d}}{Max}\ &\frac{1}{2N_{CL}}\sum_{(x_i,x_j)\in CL}\left\|h(x_i)-h(x_j)\right\|^2W_{i,j}^{(CL)}\\ &-\frac{\vartheta}{2N_{ML}}\sum_{(x_i,x_j)\in ML}\left\|h(x_i)-h(x_j)\right\|^2W_{i,j}^{(ML)}\end{aligned}, \quad (10)$$

where $h(x_i)=T^{\mathrm{T}}x_i$ is the low-dimensional representation of x_i, ϑ is a control parameter, N_{ML} and N_{ML} are the number of the ML and CL constraints respectively, $W^{(ML)}$ and $W^{(CL)}$ denote the weight matrices for measuring the pairwise similarities/ dissimilarities over the ML and CL constraints. There are two popular ways (either *global* [28-29][39] or *local* [26][30-31]) to set the weights. In global setting, each data pair $(x_i,x_j)\in ML$ or $(x_i,x_j)\in CL$ is equally treated (i.e., *hard-similarity measure*). In this case, $W_{i,j}^{(ML)}=1$ for each $(x_i,x_j)\in ML$ and $W_{i,j}^{(CL)}=1$ for each $(x_i,x_j)\in CL$; otherwise $W_{i,j}^{(ML)}=W_{i,j}^{(CL)}=0$. The local setting incorporates local information of data into the definition of PCs sets, and only neighbors from the ML and CL sets are weighted with nonzero values; else zeros. In this case, either hard (e.g., *simple-minded method* [12]) or soft measure (e.g., *heat kernel* [12]) can be used. Note that the local settings also need to estimate optimal neighborhood size or kernel width.

It is also noted that the above issue is usually solved under a supervised setting, with the PCs are obtained from the ground-truth labels of data. Although PCs exhibits some advantages over the class labels, if the number of labeled data is too few, the PCs guided problems will have not superiority any more and even a disadvantage in special cases. For instance, if there are only two labeled data (either intra- or inter-class), we can only derive one single ML or CL constraint. To address the insufficient labeled data sampling problem, in what follows we will propose to address the above problem under a semi-supervised setting, where the PCs sets defined based on the propagated soft labels from both labeled and unlabeled data.

B. Our Proposed G-CSSL Framework

Based on the predicted soft label matrix F (where the entries of each column f_i are probabilistic values of each data point belonging to different classes), one can easily obtain the labels of unlabeled data according to the biggest probability values in each column. Thus, the insufficient data labeling issue can be naturally addressed. Then based on the predicted soft labels, the PCs sets can be similarly constructed as Eq.9, but note that the *ML* and *CL* constraint sets are defined over the first *c* classes obtained from *F* in this framework, similarly as [8] that also use the first *c* rows of *F* for scatter matrix construction. This is mainly because the discovered novel class by SLP, i.e., the (*c+1*)-th class, mainly include outliers or ambiguous points from different classes that are difficult for classification. Since we have sufficient samples with labels now, the superiority of PCs over the class labels can be highlighted to the greatest extent possible. In what follows, we first construct two weight matrices $W^{(ML)}$ and $W^{(CL)}$ with size $N\times N$ for the similarity measurements via defining a voting strategy based mixed soft-similarity measure (MSM) before formulating the objective function of our G-CSSL framework.

Proposed mixed soft-similarity (MSM) measure

The matrices $W^{(ML)}$ and $W^{(CL)}$ are first initialized with all zeros. Note that the sparse representation S^* is naturally discriminant [20][21], that is, it selects a set of samples that most compactly expresses the given points and exclude all other less compact samples. So, if there are sufficient samples from each subject, each data point can be represented using a linear combination of samples from the same subject. In addition, the pairs, that contribute more together involving nonzero bigger values $s_{i,j}^*$, are most likely to be "neighbors". Next we first symmetrize S^* as $S^*\leftarrow(S^*+S^{*\mathrm{T}})/2$. Then, for each data pair $(x_i,x_j)\in ML$, this paper assigns the following Cosine similarity based edge weights to measure the similarities between them:

$$W_{i,j}^{(ML)}=\begin{cases}\exp(s_{i,j}^*)\times\cos(\theta), if\ (x_i,x_j)\in ML\\ 0\qquad\qquad, if\ (x_i,x_j)\notin ML\end{cases}\ \text{with}\ \cos(\theta)=\frac{\langle f_i^\dagger,f_j^\dagger\rangle}{(\|f_i^\dagger\|\bullet\|f_j^\dagger\|)}, \quad (11)$$

where $\exp(\bullet)$ is exponential function, and $s_{i,j}^*$ is the *(i,j)*-th entry of S^*. $f_i^\dagger$ denotes a column vector of the truncated version (i.e., $F^\dagger=[f_1^\dagger,f_2^\dagger,...,f_{l+u}^\dagger]\in\mathbb{R}^{c\times(l+u)}$) of the predicted

label matrix $F=\left[f_1,f_2,\ldots,f_{l+u}\right]\in\mathbb{R}^{(c+1)\times(l+u)}$, where $f_i^{\dagger}$ is the truncated f_i including the first c elements of f_i, and the entries of $f_i^{\dagger}$ satisfy $0\le f_{i,j}^{\dagger}\le 1$. The Cosine similarity ($\in[0,1]$) measures the similarity between two sample vectors through computing the cosine of the angle between them. The main idea of the above weighting approach is based on a voting strategy. Note that one major contribution of this paper is to create the pairwise constraints based on the propagated soft labels by SLP for delivering more supervised information. So ideally, if the predicted labels of samples by SLP are accurate, the data vectors $f_i^{\dagger}$ and $f_j^{\dagger}$ will be undoubtedly "close". As a result, the corresponding Cosine similarity (i.e., $\cos(\theta)$) is also higher. Note that based on this condition we also consider information from the sparse codes to make the final decision for the weight assignments, that is, a voting result is adopted. Two conditions are considered here. On one hand, if the predicted labels of data pair $(x_i,x_j)\in ML$ by SLP are accurate (i.e., $\cos(\theta)$ is bigger), and at the same time the samples x_i and x_j contribute more together (i.e., $\exp(s_{i,j}^*)$ is bigger), heavier weight $W_{i,j}^{(ML)}$ will be incurred. On the other hand, supposing that there exist two ambiguous sample points and are incorrectly predicted by SLP to be the same class. In this case, $s_{i,j}^*$ may be equal to zero or a very small number, that is, $\exp(s_{i,j}^*)=1$ or $\exp(s_{i,j}^*)\to 1$. Thus, a relatively lighter weight $W_{i,j}^{(ML)}$ will be incurred. In conclusion, the weight $W_{i,j}^{(ML)}$ will be the heaviest if and only if both $\exp(s_{i,j}^*)$ and $\cos(\theta)$ are bigger at the same time, that is both the special label propagation and sparse coding processes have reached a consensus. Hence the proposed weighting method is called voting strategy based mixed soft-similarity measure.

Based on similar idea of the voting strategy, we can define the following weights for each pair of instances $(x_i,x_j)\in CL$ predicted using SLP to measure the similarity between them:

$$W_{i,j}^{(CL)}=\begin{cases}\exp\left(1-s_{i,j}^*\right)\times\left(1-\cos(\theta)\right), & if\ (x_i,x_j)\in CL\\ 0, & if\ (x_i,x_j)\notin CL\end{cases}. \quad (12)$$

Analogously, the heaviest penalty will be imposed on the edge weights $W_{i,j}^{(CL)}$ for data pair $(x_i,x_j)\in CL$ if and only if both $\exp(1-s_{i,j}^*)$ and $1-\cos(\theta)$ are bigger at the same time. Otherwise, a lighter penalty will be incurred.

The objective function of G-CSSL

After the weight matrices $W^{(ML)}$ and $W^{(CL)}$ are constructed, we can define the following objective function for our G-CSSL approach to compute a projection matrix $T\in\mathbb{R}^{n\times d}$ onto which enhanced inter-class separation and intra-class compactness can be obtained at the same time:

$$\underset{T\in\mathbb{R}^{n\times d}}{Max}\ \frac{1}{2}\sum_{i,j=1}^{N}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2\widehat{W_{i,j}}+\frac{1}{2N_{CL}}\sum_{(x_i,x_j)\in CL}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2W_{i,j}^{(CL)}$$
$$-\ \frac{\vartheta}{2N_{ML}}\sum_{(x_i,x_j)\in ML}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2W_{i,j}^{(ML)},\ \text{Subj}\ T^{\mathrm{T}}T=I \quad (13)$$

where $\widehat{x}_i=Xs_i^*$ denotes the reconstructed sample by the sparse coefficient vector s_i^*, $h(\widehat{x}_i)=T^{\mathrm{T}}\widehat{x}_i$ represents the transformed low-dimensional representation of $\widehat{x}_i$, and $\widehat{W_{i,j}}=1/N$ for each pair of instances (i.e., $\sum_{i,j=1}^{l+u}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2\widehat{W_{i,j}}$ is the *principal component analysis* (PCA) operator [5]), so this regularization $\sum_{i,j=1}^{l+u}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2\widehat{W_{i,j}}$ is mainly added to preserve the global covariance structures of all samples including both labeled and unlabeled data, especially useful for the extreme case such that labels of all unlabeled data are incorrectly predicted by SLP to be c+1 class. In this extreme case, the number of constraints obtained from the labeled data is few, thus the motivation for exploiting unlabeled data is to combine them for enhancing performance. Note that the above modeling is similar to the formulation of [28], but we used the reconstructed data $\widehat{x}_i$ and our mixed soft-similarity weights in the optimizations. We can have a concise form for Eq.13:

$$\underset{T^{\mathrm{T}}T=I}{Max}\ \frac{1}{2}\sum_{i,j=1}^{N}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2\Omega_{i,j}^{(CL)}-\frac{\vartheta}{2N_{ML}}\sum_{(x_i,x_j)\in ML}\left\|h(\widehat{x}_i)-h(\widehat{x}_j)\right\|^2W_{i,j}^{(ML)}$$
$$where\ \ \Omega_{i,j}^{(CL)}=\begin{cases}\widehat{W_{i,j}}+(1/N_{CL})W_{i,j}^{(CL)} & if\ (x_i,x_j)\in CL\\ \widehat{W_{i,j}} & if\ (x_i,x_j)\notin CL\end{cases}, \quad (14)$$

Because $\left\|T^{\mathrm{T}}\widehat{x}_i-T^{\mathrm{T}}\widehat{x}_j\right\|^2=tr\left[\left(T^{\mathrm{T}}\widehat{x}_i-T^{\mathrm{T}}\widehat{x}_j\right)\left(T^{\mathrm{T}}\widehat{x}_i-T^{\mathrm{T}}\widehat{x}_j\right)^{\mathrm{T}}\right]$, based on the matrix interpretation, Eq.14b can be converted into

$$\underset{T\in\mathbb{R}^{n\times d}}{Max}\ tr\left[T^{\mathrm{T}}\widehat{X}\left(L^{(CL)}-\vartheta L^{(ML)}\right)\widehat{X}^{\mathrm{T}}T\right],\ \text{Subj}\ T^{\mathrm{T}}T=I, \quad (15)$$

where $L^{(ML)}=D^{(ML)}-W^{(ML)}$ and $L^{(CL)}=\Theta^{(CL)}-\Omega^{(CL)}$ denote the graph Laplacian matrices, $W_{i,j}^{(ML)}=(1/N_{ML})W_{i,j}^{(ML)}$, $D^{(ML)}$ and $\Theta^{(CL)}$ are diagonal matrices with $D_{ii}^{(ML)}=\sum_j\widehat{W_{i,j}^{(ML)}}$ and $\Theta_{i,j}^{(CL)}=\sum_j\Omega_{i,j}^{(CL)}$, $tr(\bullet)$ is trace operator and $\widehat{X}=\left[\widehat{x_1},\widehat{x_2},\ldots,\widehat{x_N}\right]$. ϑ is a parameter for trading-off terms $tr\left(\widehat{X}L^{(CL)}\widehat{X}^{\mathrm{T}}\right)$ and $tr\left(\widehat{X}L^{(ML)}\widehat{X}^{\mathrm{T}}\right)$, where $tr\left(\widehat{X}L^{(CL)}\widehat{X}^{\mathrm{T}}\right)$ can measure the separation degree of data points and $tr\left(\widehat{X}L^{(ML)}\widehat{X}^{\mathrm{T}}\right)$ measures the compactness degree of data. From the above problem, the projection matrix $T\in\mathbb{R}^{n\times d}$ can be obtained including the orthogonal eigenvectors according to leading d eigenvalues of the following eigenvalue problem: $\widehat{X}\left(L^{(CL)}-\vartheta L^{(ML)}\right)\widehat{X}^{\mathrm{T}}\psi_j=\widehat{\lambda}_j\psi_j$. After T is obtained, dimension reduction of X can be performed in the form of $T^{\mathrm{T}}X$ and T can be used for embedding new data in classification. More specifically, when a new test data is input, its low-dimensional embedding can be obtained by projecting it onto the projection axes. For complete presentation of the method, we summarize our G-CSSL framework in Algorithm 1. Note that a detailed version of this work appeared in [6] that also presented a two-stage SC to gain adaptive neighborhood for SLP and conducts a thorough simulation evaluation on classification.

IV. Simulation Results and Analysis

This section examines our G-CSSL method for image feature extraction and representation, along with illustrating results. Because G-CSSL is a SLP process induced PCs based SSL algorithm, its classification performance is mainly compared with *Semi-Supervised Dimensionality Reduction* (SSDR) [28], *Semi-Supervised Metric Learning* (SSML) [29], *Marginal*

Semi-Supervised Sub-Manifold Projections (MS^{3}MP) [26], *orthogonal MS^{3}MP* (OMS^{3}MP) [26] and the *Semi-supervised Orthogonal Discriminant Analysis* (SODA) [8]. It is noted that SSDR, SSML, MS^{3}MP and OMS^{3}MP are pairwise constrained SSL algorithms, while the SODA technique performs SSL via label propagation. SSML, MS^{3}MP, OMS^{3}MP and SODA have a common parameter (i.e., neighborhood size k) to estimate. In addition, SLP uses the Gaussian kernel to assign edge weights, so it has a parameter (i.e., kernel width δ) to estimate. To provide a reasonable estimation for δ, the Gaussian kernel width is defined as $\delta = \hat{\delta} / \varpi$, $\hat{\delta} = \sum_{i,j} \|x_i - x_j\|^2 / \left(N^2 - N\right)$ with a carefully chosen ϖ, similarly as [7][19]. For the k-neighbor search based methods, the number of k is carefully tuned from $\{5,7,9,11\}$ and the best classification performance is reported. We perform all simulations on a PC with Intel (R) Core (TM)2 Quad CPU Q9550 @ 2.83GHz 2.83 GHz.

Algorithm 1: Graph based Constrained Semi-Supervised Learning

Inputs:

Data matrix $X \in \mathbb{R}^{n \times N}$ including labeled X_L and unlabeled X_U;

The reduced dimensionality $d \le n$.

Output: The transformation matrix $T \in \mathbb{R}^{n \times d}$.

1. Predict the soft labels of data by using the adaptive neighborhood based SLP process;
2. Construct the PCs sets based on the propagated labels and define the MSM approach;
3. Solve the eigen-value problem: $\widehat{X}\left(L^{(CL)} - \vartheta L^{(ML)}\right)\widehat{X}^{\mathrm{T}}\psi_j = \widehat{\lambda}_j \psi_j$, where $T \leftarrow [\psi_1, \psi_2, ..., \psi_d]$ according to d leading eigenvalues $\{\widehat{\lambda}_j\}_{j=1}^{d}$.

The recognition process of our G-CSSL and other methods are described as follows. Each dataset is randomly split into a training set X_{Tr} and a test set X_{Te}. The training set including labeled data X_L and unlabeled data X_U is used to train a learner. Prior to subspace learning, PCA is used to eliminate the null space of training set. The data X_{Te} is then embedded onto the reduced space with the projection matrix learned from training data. Finally, the learner is used to evaluate the test accuracies. The one-nearest-neighbor (1NN) classifier with Euclidean metric is used for classification due to its simplicity. Note that we show the whole procedures of applying our G-CSSL for recognition in Figure 1. In this study, one synthetic set and two real problems are tested. The first one is a "two moon" dataset; the second one is COIL-20 database (available from *http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php*); the third one is the COIL-100 database (available from *http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php*). As is common practice, all images are resized to 20×20 pixels due to the computational consideration, so each image corresponds to a point in a 400-dimensional space.

A. Object Recognition on COIL-20

The COIL-20 database has a total of 1440 gray object images with black background for 20 different subjects (objects), with 72 images from each subject. In the experiments, the PCs sets in SSDR and SSML are created based on whether class labels of samples in X_L are the same or different [28][29], while the PCs sets are obtained relying on whether the class labels of neighboring points in X_L are the same or not in MS^{3}MP and OMS^{3}MP [26]. For fair comparison, the labels of unlabeled data are predicted by SLP for SODA and our method. In the simulations, the parameters ϑ in our proposed G-CSSL and λ in the SC are carefully chosen from $\{10^i \mid i = -6, -5, ..., 6\}$ for fair comparison and the best classification results will be reported. In the simulations below, the regularization factor α_l is set to 0 and α_u is carefully tuned from $\{1 - 10^{-i} \mid i = 3, 5, ..., 15\}$ for SLP.

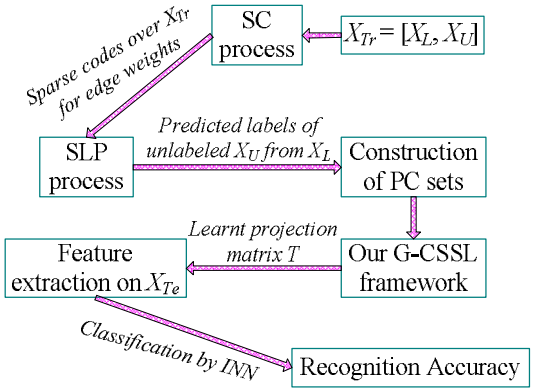


Figure 1: Illustration of the recognition procedures using our proposed G-CSSL algorithm.

Object recognition results

In this study, three experimental settings over various numbers of labeled data points (i.e. 5, 10 and 15 labeled respectively) randomly selected from each object class are tested. For each case, the number of unlabeled samples is double the number of labeled samples. For each setting, we regulate the numbers of the reduced dimensions from 3 to 60 with interval 3, and the test results are averaged over first 15 best records based on 20 realizations of training/test sets. We report the mean accuracy, best record and the optimal image subspace (i.e., *Dim*), where the optimal subspace corresponds to the highest recognition accuracy of each method in each setting, in Table 1. For fair comparison, SSDR, SSML, MS^{3}MP and OMS^{3}MP also use all available constraints to learn the projections. We have the following similar observations. First, the accuracies of all the algorithms are improved when the number of training samples increases. Second, our proposed G-CSSL can always achieve comparable and even better accuracies than other methods. Most importantly, our proposed G-CSSL criterion is capable of delivering the best results using smaller number of reduced dimensions in each case. The major reason may owe to the adaptive neighborhood and noise removal by applying the SC process. Third, MS^{3}MP are comparable to OMS^{3}MP, and both are highly competitive with SODA for recognizing the objects. SSDR is the worst method for this dataset.

Object recognition against pixel corruptions

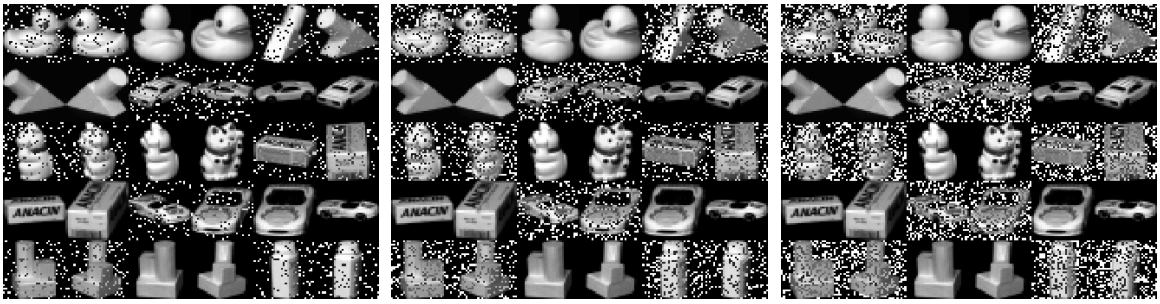
We also address an experiment to examine the robustness of our G-CSSL in recognizing the objects under various degrees of random pixel corruptions. This simulation considers three

settings over different levels of corruptions: one is with 10% pixels corrupted, one is with 20% pixels corrupted and the last one is with 40% pixels corrupted. For each pixel selected to be corrupted, its pixel value ξ is replaced by its inverse pixel, i.e., subtracting ξ from the biggest pixel value of images. We show typical samples in the training set, including the original images and corrupted images under various levels in Figure 2. The number of labeled data per object is fixed to 10 and the number of unlabeled data is still double the labeled number. To investigate the robustness of each method against pixel corruptions, half of the labeled set, half of the unlabeled set and the whole test set are corrupted. In each setting, the first half of the labeled data (and unlabeled data) per object class are chosen to be corrupted. For each method, the training set, including labeled and unlabeled data, is applied to train the classifier and the test set is used for performance evaluation.

We describe the averaged results of the cases handing pixel corruptions in Table 2. We find that: (1) the increasing level of pixel corruptions can decrease the recognition power of each approach. Specifically, SSDR and SSML are more sensitive to the pixel corruptions in the images, because their accuracies decreased faster than other methods. SODA, MS^{3}MP and our G-CSSL are more robust against corruptions in all cases due to their reasonable motivations and formulations. OMS^{3}MP works well in the first two cases, but when the level of pixel corruptions is increased to 40%, the performance of OMS^{3}MP is significantly weakened. (2) Our G-CSSL can outperform the other methods in delivering the boosted accuracies in most cases. SODA obtains comparative highest records with our G-CSSL technique in most cases. Note that SSDR and SSML deliver the worst results in all cases. (3) Our G-CSSL method achieves the highest records with smaller number of reduced dimensions involved in most cases.

Table 1: Performance comparison of the algorithms on the COIL-20 object database.

Result / Method	COIL-20 (5 labeled)			COIL-20 (10 labeled)			COIL-20 (15 labeled)		
	Mean	Best	Dim	Mean	Best	Dim	Mean	Best	Dim
SSDR	0.7597	0.7886	51	0.8683	0.8881	36	0.9090	0.9259	24
SSML	0.7886	0.8193	36	0.8993	0.9161	15	0.9201	0.9364	18
MS^{3}MP	0.8352	0.8538	21	0.9176	0.9312	12	0.9397	0.9537	18
OMS^{3}MP	0.8320	0.8515	12	0.9121	0.9248	33	0.9405	0.9540	24
SODA	0.7943	0.8529	60	0.9077	0.9378	60	0.9432	0.9578	57
SC based G-CSSL	0.8476	0.8578	12	0.9253	0.9361	12	0.9528	0.9627	15



(a) 10% pixels corrupted (b) 20% pixels corrupted (c) 40% pixels corrupted

Figure 2: Typical samples of the original images and corrupted images under various levels.

Table 2: Performance comparison of the algorithms on the COIL-20 dataset with pixel corruptions.

Result / Method	COIL-20 (10% corrupted)			COIL-20 (20% corrupted)			COIL-20 (40% corrupted)		
	Mean	Best	Dim	Mean	Best	Dim	Mean	Best	Dim
SSDR	0.7829	0.8085	21	0.6961	0.7512	60	0.3991	0.5054	60
SSML	0.7104	0.7625	60	0.5890	0.6509	60	0.4229	0.4664	51
MS^{3}MP	0.8505	0.8857	12	0.7949	0.8303	24	0.6347	0.6852	24
OMS^{3}MP	0.8649	0.8930	21	0.8066	0.8305	18	0.5633	0.6560	12
SODA	0.8159	0.8705	60	0.7339	0.8299	60	0.5539	0.6505	60
SC based G-CSSL	0.8753	0.9018	15	0.8315	0.8510	15	0.6606	0.7129	12

B. Object Recognition on COIL-100 Database

This study examines the recognition capability of our G-CSSL on the COIL-100 database. This database has 7200 images of 100 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects were taken at pose intervals of 5 degrees, corresponding to 72 different poses per object. We show some typical sample images of the database in Figure 3.

In this simulation, the first 40 objects (totally 2880 images) of the database are selected. We prepare three settings under

different numbers of labeled samples (i.e. 5, 10 and 15 labeled respectively) that are randomly selected from each object class. For each setting, the number of unlabeled data is double the labeled number and the numbers of reduced dimensions are regulated from 3 to 60 with interval 3 for each fixed training size. Table 3 summarizes the mean and the highest accuracies under various numbers of reduced dimensions, where we also describe the experimental setting of training set. We have the following findings. First, the increasing numbers of training samples significantly boost the performance of each algorithm. Second, the mean and highest accuracies of the MS^{3}MP and OMS^{3}MP methods are comparable to G-CSSL in most cases. SSDR outperforms SSML in each case. SODA delivers the comparable results to SSDR and SSML in most cases, and SODA obtains close results to MS^{3}MP, OMS^{3}MP and our G-CSSL if more reduced dimensions are used in each setting.

Figure 3: Typical sample images of first 50 objects of the COIL-100 database.

Table 3: Performance comparison of the algorithms on the COIL-100 object database.

Method \ Result	COIL-100 (5 labeled)			COIL-100 (10 labeled)			COIL-100 (15 labeled)		
	Mean	Best	Dim	Mean	Best	Dim	Mean	Best	Dim
SSDR	0.7298	0.7508	36	0.8306	0.8529	33	0.8741	0.9048	48
SSML	0.6956	0.7241	30	0.7983	0.8230	21	0.8272	0.8614	51
MS^{3}MP	0.8024	0.8220	24	0.8824	0.8997	27	0.9086	0.9261	30
OMS^{3}MP	0.8026	0.8147	27	0.8786	0.8926	27	0.9046	0.9173	27
SODA	0.7222	0.7870	60	0.8394	0.9069	60	0.8852	0.9399	60
SC based G-CSSL	0.8118	0.8289	12	0.8932	0.9063	27	0.9260	0.9381	54

V. Concluding Remarks

This paper has introduced a novel mechanism to achieve more supervised information of samples in graph based constrained semi-supervised learning through creating and enriching the pairwise constraint sets based on the propagated soft labels by SLP. In order to improve the performance by enhancing intra-class compactness and inter-class separation, a voting strategy guided mixed soft-similarity measure approach built based on the propagated outputs and the sparse codes is also proposed. Finally, we propose a novel graph based constrained semi-supervised learning framework, called G-CSSL, to reduce the dimensionality of data and embed new points for classification. The orthogonal projection matrix of G-CSSL can be obtained by eigen-decomposition analytically.

This paper mainly tests G-CSSL for image representation. Although promising results are delivered by our algorithm, the following future directions are still worth investigating. First, exploring how to speed up the sparse coding process with the effectiveness ensured is required. Second, when there are no sufficient clean data X available, the robustness of SC to noise and outliers may be greatly weakened by setting the matrix X itself as the dictionary [4], thus it is important to explore the effective approach of designing an optimal clean informative dictionary for the sparse coding process.

VI. Acknowledgements

This present work is partially supported by the Major Program of National Natural Science Foundation of China (Grant No. 61033013), and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-Supervised Learning," Cambridge: MIT Press, 2006.

[2] G.C Liu, Z. C Lin, S. C Yan, J. Sun, Y. Yu, and Y. Ma, "Robust Recovery of Subspace Structures by Low-Rank Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.85, no.1, pp.663-670, 2012.

[3] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Technical report, UILU-ENG-09-2215*, 2009.

[4] J. Wright, Y. Y. Tao, Z. C. Lin, Y. Ma, and H. Y. Shum, "Classification via minimum incremental coding length (MICL). In: *Neural Information Processing Systems (NIPS)*, 2008.

[5] I. Jolliffe, "Principal Component Analysis," Springer-Verlag, 1986.

[6] Z. Zhang, M.B. Zhao, and Tommy W. S. Chow, "Graph based Constrained Semi-Supervised Learning Framework via Label Propagation over Adaptive Neighborhood," *IEEE Transactions on Knowledge and Data Engineering*, Dec 2013. To appear.

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

[8] F. P. Nie, S. M. Xiang, Y.Q. Jia, C. S. Zhang, "Semi-Supervised Orthogonal Discriminant Analysis via Label Propagation," *Pattern Recognition*, vol.42, no.11, pp.2615-2627, 2009.

[9] F. Wang, and C. S. Zhang, "Label propagation through linear Neighborhoods," *IEEE Transactions on Knowledge and Data Engineering*, vol.20, no.1, pp.55-67, 2008.

[10] Z. Tian, R. Kuang, "Global Linear Neighborhoods for Efficient Label Propagation," In: *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2012.

[11] Z. Zhang, and N. Ye, "Learning a Tensor Subspace for Semi-Supervised Dimensionality Reduction," *Soft Computing*, vol.15, no.2, pp.383-395, 2011.

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

[13] N. Yang, Y. Sang, R. He, and X. Wang, "Label propagation algorithm based on non-negative sparse representation," In K. Li, L. Jia, X. Sun, M. Fei, and G. Irwin, editors, *Life System Modeling and Intelligent Computing*, volume 6330 of Lecture Notes in Computer Science, pages 348-357, 2010.

[14] F. Zang, and J. S. Zhang, "Label propagation through sparse neighborhood and its applications," *Neurocomputing*, vol.97, pp.267–277, 2012.

[15] H. Cheng, Z. Liu, and J. Yang, "Sparsity induced similarity measure for label propagation," In: *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 317-324, 2009.

[16] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions. In: *the International Conference on Machine Leaning*, 2003.

[17] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. S cholkopf, "Learning with local and global consistency," In: *NIPS*, 2004.

[18] Y. Liu, F. P. Nie, J. G. Wu, and L. H. Chen, "Semi-supervised feature selection based on label propagation and subset selection," In: *Proceedings of the International Conference on Computer and Information Application*, 2010.

[19] F. P. Nie, S. M. Xiang, Y. Liu, C.S. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing Applications*, vol.19, no.4, pp.549-555, 2010.

[20] L. S. Qiao, S. C. Chen, and X. Y. Tan, "Sparsity Preserving Projections with Applications to Face Recognition," *Pattern Recognitio*n, vol.43, no.1, pp.331-341, 2010.

[21] Z. Zhang, S. C. Yan, and M. B. Zhao, "Pairwise Sparsity Preserving Embedding for Unsupervised Subspace Learning and Classification," *IEEE Trans. on Image Processing*, accepted in July 2013.

[22] J. Wright, A. Yang, S. Sastry, Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.31, no.2, pp.210-227, 2009.

[23] Z. Zhang, M. B. Zhao, and T. W. S. Chow, "Binary- and Multi-Class Group Sparse Canonical Correlation Analysis for Feature Extraction and Classification," *IEEE Trans. on Knowledge and Data Engineering*, vol.25, no.10, pp. 2192-2205, October 2013.

[24] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. on Neural Networks*, vol.17, no.1, pp.157-165, 2006.

[25] Y. Q. Song, F. P. Nie, C. S. Zhang, and S. M. Xiang, "A unified framework for semi-supervised dimensionality reduction," *Pattern Recognition*, vol.41, no.9, pp.2789-2799, 2008.

[26] Z. Zhang, M. B. Zhao, and T. W. S. Chow, "Marginal Semi-Supervised Sub-Manifold Projections with Informative Constraints for Dimensionality Reduction and Recognition," *Neural Networks*, vol.36, pp.97- 111, 2012.

[27] E. Elhamifar, and R. Vidal, "Sparse subspace clustering," In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2790-2797, 2009.

[28] D. Q. Zhang, Z. H. Zhou, and S. C. Chen, "Semi-supervised dimensionality reduction," In: *Proceedings of the 7th SIAM International Conference on Data Mining (SDM)*, Minneapolis, MN, pp.629-634, 2007.

[29] M. S. Baghshah, and S. B. Shouraki, "Semi-Supervised Metric Learning Using Pairwise Constraints," In: *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 1217-1222, 2009.

[30] Z. Zhang, T. W. S. Chow, M. B. Zhao, "M-Isomap: Orthogonal Constrained Marginal Isomap for Nonlinear Dimensionality Reduction," *IEEE Trans. on Systems, Man and Cybernetics Part B: Cybernetics*, vol.43, iss.1, pp.180-192, Feb 2013.

[31] Z. Zhang, T. W. S. Chow, and M. B. Zhao, "Trace Ratio Optimization based Semi-Supervised Nonlinear Dimensionality Reduction for Marginal Manifold Visualization," *IEEE Trans. on Knowledge and Data Engineering,* vol.25, iss.5, pp.1148-1161, May 2013.

[32] M. Wang, X. Hua, T. Mei, R. Hong, G. Qi, Y. Song, and L. Dai, "Semi-supervised kernel diensity estimation for video annotation," *Computer Vision and Image Understanding*, vol.13, no.3, pp.384-396, 2009.

[33] J. Chen, J. Ye, Q. Li, "Integrating global and local structures: a least squares framework for dimensionality reduction, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[34] D. Cai, X. F. He, and J. W. Han, "Semi-supervised discriminant analysis," In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, pp.1-7, 2007.

[35] X. Zhu, "Semi-supervised learning literature survey," Technical Report 1530, Univ. Wisconsin-Madison. 2005.