IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 28, 2020, accepted May 18, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997413

A Manifold Laplacian Regularized Semi-Supervised Sparse Image Classification Method With a Variant Trace Lasso Norm

WENJUAN ZHANG [1], XIANGCHU FENG [2], AND YUNMEI CHEN [3], (Member, IEEE)

[1] School of Science, Xi'an Technological University, Xi'an 710021, China
[2] School of Mathematics and Statistics, Xidian University, Xi'an 710071, China
[3] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

Corresponding author: Wenjuan Zhang (zhangwenjuan@xatu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772389 and Grant 61572392, in part by the Scientific Research Program funded by the Shaanxi Provincial Education Department under Grant 15JK1371 and Grant 18JK0373, in part by the Principal Fund Project of Xi'an Technological University under Grant XAGDXJJ17027, and in part by the Natural Science Basic Research Plan in Shaanxi Province, China, under Grant 2020 JM-569.

ABSTRACT Since the cost of labeling data is getting higher and higher, we hope to make full use of the large amount of unlabeled data and improve image classification effect through adding some unlabeled samples for training. In addition, we expect to uniformly realize two tasks, namely the clustering of the unlabeled data and the recognition of the query image. We achieve the goal by designing a novel sparse model based on manifold assumption, which has been proved to work well in many tasks. Based on the assumption that images of the same class lie on a sub-manifold and an image can be approximately represented as the linear combination of its neighboring data due to the local linear property of manifold, we proposed a sparse representation model on manifold. Specifically, there are two regularizations, i.e., a variant Trace lasso norm and the manifold Laplacian regularization. The first regularization term enables the representation coefficients satisfying sparsity between groups and density within a group. And the second term is manifold Laplacian regularization by which label can be accurately propagated from labeled data to unlabeled data. Augmented Lagrange Multiplier (ALM) scheme and Gauss Seidel Alternating Direction Method of Multiplier (GS-ADMM) are given to solve the problem numerically. We conduct some experiments on three human face databases and compare the proposed work with several state-of-the-art methods. For each subject, some labeled face images are randomly chosen for training for those supervised methods, and a small amount of unlabeled images are added to form the training set of the proposed approach. All experiments show our method can get better classification results due to the addition of unlabeled samples.

INDEX TERMS Image classification, manifold, semi-supervised, sparse, trace lasso.

I. INTRODUCTION

Image classification is one of the most active applications in image processing, computer vision and machine learning and has been extensively studied by numerous researchers. Meanwhile, numerous image classification and representation methods have been proposed. Wright *et al.* proposed a Sparse Representation Coding (SRC) method by applying the l_1-norm based sparse representation to Face Recognition (FR) [1]. SRC has shown interesting results in image classification and recognition and has been widely used and extended, as evidenced by its many followup papers. Some later works, on the other hand, began to investigate the role of sparsity in image representation [2]–[5]. Yang *et al.* [4] gave an insight into SRC and provided some theoretical support for its effectiveness. They argued that it is l_1 constraint rather than l_0 that makes SRC effective. Lei Zhang *et al.* indicated that most literatures emphasized too much on the role of l_1-norm sparsity in image classification. They demonstrated that it is actually the Collaborative Representation (CR), i.e., using the training samples from all classes to represent the query sample, but not the l_1-norm, that plays the essential role in SRC. Therefore, they proposed the CR based classification with regularized least square (CRC_RLS) [5], which has

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

significantly less complexity than SRC but leads to very competitive classification results. Zhang *et al.* [6] extended their CRC to the robust version, robust collaborative representation classification (RCRC), using the Laplacian estimator to deal with severe random pixel noise and illumination changes. Grave *et al.* proposed a Trace Lasso (TL) norm [7]. It is proved that TL interpolates between l_2-norm and l_1-norm. Its behaviour is adaptively related to the correlation of the training data. For a recognition task, if the labeling information is available, by integrating the labeling information with TL, Jian Lai *et al.* proposed a method named Supervised Trace Lasso (STL) [8]. This method can cluster the samples from the same subject but with different variation information together, which conforms to the goal of identification.

All the above works consider the query sample lies in the linear space spanned by the training samples from the same class. However, in practice, taking face image for instance, when reflectance is typically non-Lambertian and the pose of the subject varies, the data do not necessarily conform to linear subspace models. On the other hand, these methods are supervised methods, that is, they require the labeling information of all training samples. They are effective only in the small-sample-size case because when sample-size gets larger the computing burden for these methods will get heavy and particularly the cost of labeling data may be unaffordable. In recent years, a lot of methods based on deep learning [9]–[13] have been proposed. These methods obtained very competitive results on image classification and recognition task. Even though there are only a small amount of labeled samples, deep learning can also be used through semi-supervised network as long as the quantity of samples is enough. However, in some practical applications, especially for some cold research fields or some institutions with limited condition, public data sets may not be able to meet their requirement and data acquisition is also very difficult. As a result, they have to face the small-sample-size situation, which is the focus of our work here. We will show that when a small number of unlabeled samples are added to training set the classification result can be effectively improved.

Manifold regularized semi-supervised learning (MRSSL) is one of the most successful methods in computational imaging. A set of N by M images may be better modeled by a manifold embedded in an NM-dimensional Euclidean space, called an image manifold [14]. MRSSL exploits the local structure of data distribution including both labeled and unlabeled samples to leverage the generalization ability of a learning model. There are many representative works in MRSSL, in which the most prominent is Laplacian regularization which determines the underlying manifold by using the graph Laplacian [15], [16]. With the merits of simple calculation and promising performance, Laplacian regularization based semi-supervised learning has received extensive attention and many algorithms have been developed, including Laplacian regularized support vector machines, Laplacian regularized kernel least squares [15], [17], and Laplacian regularized nonnegative matrix factorization [18]. In addition, P-Laplace regularization is proposed in [19] to preserve the local geometry and is applied to support vector machines and kernel least squares. Reference [20] presented a hyper graph P-Laplacian regularization for remotely sensed image recognition. P-Laplacian is a natural generalization of the standard graph Laplacian. Besides Laplace regularization, [21] presented Hessian regularized multiset canonical correlations for multiview dimension reduction.

In this paper, a semi-supervised sparse image classification model on manifold is presented as follows

$$\begin{aligned}\arg\min_{\mathbf{x},\,\mathbf{Z}} & \left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_1+\lambda\left\|\mathbf{Z}\mathrm{Diag}\left(\mathbf{x}\right)\right\|_* \\ & +\frac{\upsilon}{2}\sum\nolimits_{i,j=1}^{n}\left\|\mathbf{z}_i-\mathbf{z}_j\right\|_2^2 S_{ij}+\sum\nolimits_{j=1}^{n}U_{jj}\left\|\mathbf{z}_j-\mathbf{g}_j\right\|_2^2 \qquad (1)\end{aligned}$$

The query image $\mathbf{y}\in R^m$ is first collaboratively represented by the whole training samples $\mathbf{A}=[\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_n]\in R^{m\times n}$, whether they are labeled or unlabeled. $\mathbf{x}$ is the representation coefficient. We assume the images of one class lie on a submanifold. Since manifold locally is a linear space, the query image can be approximately represented as the linear combination of its neighbour data, namely only the coefficients correlated with these data are not zero in the collaborative representation.

Assume among the whole n training samples, only samples whose index belongs to S have identity information. Let $\mathbf{z}_j\in R^c$ $(1\le j\le n)$ be the label vector of sample $\mathbf{a}_j$, here c is the number of classes. Since we don't know c, the dimension of $\mathbf{z}_j$ just needs to be set larger than c and at most equals to $\min\{m,\ n\}$. Here, for simplicity, we still use c to represent the dimension of $\mathbf{z}_j$. The label matrix is $\mathbf{Z}=[\mathbf{z}_1,\cdots,\mathbf{z}_n]\in R^{c\times n}$. Assume $\mathbf{g}_j=\left(g_{ij}\right)\in R^c$ $(j\in S)$ is the label vector of the labeled sample $\mathbf{a}_j$. The entries of $\mathbf{g}_j$ is very simple. $g_{ij}=1$ when $\mathbf{a}_j$ is in the ith class, while $g_{ij}=0$ otherwise. If $\mathbf{a}_j$ already has a label, namely $j\in S$, then $\mathbf{z}_j=\mathbf{g}_j$. This can be fulfilled by the last term in (1), in which $\mathbf{U}\in R^{n\times n}$ is a diagonal matrix, $\mathbf{U}_{jj}$ takes a large value when $\mathbf{a}_j$ is labeled, otherwise $\mathbf{U}_{jj}=0$. If $\mathbf{a}_j$ is not labeled, we set $\mathbf{g}_j$ to be a zero vector and $\mathbf{z}_j$ needs to be solved.

We present two regularization terms. The first regularization term, a variant Trace Lasso norm $\left\|\mathbf{Z}\mathrm{Diag}\left(\mathbf{x}\right)\right\|_*$, forces the group sparsity instead of the sample sparsity, which means the query image is presented by a small number of groups, and in each group, the training samples are fully used. This sparsity between groups and density within a group is preferred to the aim of image classification. The second one is the manifold Laplacian regularization $\sum\nolimits_{i,j=1}^{n}\left\|\mathbf{z}_i-\mathbf{z}_j\right\|_2^2 S_{ij}$, with which distance is calculated along the manifold of samples and for each class label can be appropriately propagated from labeled samples to unlabeled samples, then the accurate classification can be reached. λ and υ are two parameters used to balance the roles of two regularization terms.

The rest of the paper is organized as follows. The related works to our method are analysed in Section 2. In Section 3, we present the manifold Laplacian regularized semi-supervised sparse image classification model. And two regularization terms are presented. The first one is a variant Trace Lasso regularization by combining semi-supervised samples with Trace Lasso norm. The second one is the manifold Laplacian regularization. The numerical methods, i.e. the Augmented Lagrange Multiplier (ALM) scheme and Gauss Seidel Alternating Direction Method of Multiplier (GS-ADMM) are given in Section 4 to solve the problem numerically. In Section 5, through experiments with several commonly used databases, we compare the performance of the proposed method with several state-of-the-art methods to show the effect of our method. Concluding and discussing remarks are made in Section 7. As far as we know, it is the first time to use unlabeled samples in the sparse representation based image classification methods. Compared to the methods based on deep learning, our method belongs to the type of knowledge-based modeling approaches which has clear structure and better theoretically interpretability.

Some notations used in this work are defined as follows. For a vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ is the l_2 norm, $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$ is the l_1 norm, $\|\mathbf{x}\|_0$ is the l_0 norm which counts the number of nonzero elements of vector $\mathbf{x}$. $\mathrm{Diag}(\mathbf{x})$ is a diagonal matrix whose diagonal entries are $\mathbf{x}$. For a matrix $\mathbf{X} = \left(X_{ij}\right)$, $\mathbf{x}_j$ and $\mathbf{x}^j$ represent the jth column and jth row of $\mathbf{X}$ separately, $\|\mathbf{X}\|_{\mathrm{F}} = \sqrt{\sum \left|X_{ij}\right|^2}$ denotes the Frobenius norm, $\|\mathbf{X}\|_* = \sum \sigma_i(\mathbf{X})$ is the nuclear norm, here $\sigma_i(\mathbf{X})$ is the ith singular value of $\mathbf{X}$. $\mathbf{X}^{\mathrm{T}}$ refers to the transpose of $\mathbf{X}$. $\mathrm{Diag}(\mathbf{X})$ describes the diagonal matrix with diagonal components being X_{ii}, $\mathrm{diag}(\mathbf{X})$ is a vector with entries X_{ii}. $\mathrm{tr}(\mathbf{X})$ is the trace function of the square matrix $\mathbf{X}$.

II. RELATED WORKS

A. SRC

Wright *et al.* considered the following model

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_0 \tag{2}$$

where the query sample is represented as a linear combination of all labeled training samples. Suppose the training samples from the same subject to be in a single subspace, therefore the l_0-norm forces the query sample to be sparsely represented and we hope that the training samples significantly contribute to the representation are from the same subspace with the query.

However, (2) is an NP-hard problem. It has been proved that, if $\mathbf{x}$ is sparse enough, the solution of l_0 minimization problem (2) is equivalent to the solution of the following l_1 minimization problem called SRC

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{3}$$

Many efficient methods have been proposed to solve this problem [22]–[24]. Further analysis showed that if training data are highly correlated, to achieve the sparse goal, SRC may randomly select one sample. This randomness could cause the SRC unstable and lead to misclassification by selecting the sample from the wrong subject.

B. CRC_RLS

Lei Zhang *et al.* indicated that it is the CR, but not the l_1-norm sparsity, that plays an essential role for classification in SRC. Therefore they proposed to use the l_2-norm, which can have similar classification results but with significantly lower complexity

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 \tag{4}$$

However, when columns of $\mathbf{A}$ are orthogonal to each other, we need many samples to faithfully represent $\mathbf{y}$, then the discrimination ability of (4) becomes weaker.

C. TL

Grave *et al.* proposed a new norm named Trace Lasso (TL) norm $\|\mathbf{A}\mathrm{Diag}(\mathbf{x})\|_*$. It is shown that TL interpolates between l_2-norm and l_1-norm. Its behaviour adaptively depends on the correlation of training data. When all columns of $\mathbf{A}$ are the same, the result of TL is the same as that of l_2-norm of $\mathbf{x}$. When all columns of $\mathbf{A}$ are orthogonal to each other, the result of TL is the same as that of l_1-norm of $\mathbf{x}$. So TL norm shares the advantages of l_1-norm and l_2-norm. Using it as the regularization term, the following problem is obtained

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{A}\mathrm{Diag}(\mathbf{x})\|_* \tag{5}$$

TL naturally clusters the highly correlated initial sampling data $\mathbf{A}$. However, it is well known that face images include much information, such as identity and variations (e.g. illumination and expression). In the uncontrolled environment, variation information can be more significant than identity. In this case, the correlation will depend more on variations than identity. Therefore, face images from different subjects with similar variations could have a higher correlation than those from the same subject but with different variations. As a result, TL naturally clusters the samples with similar variations together. The outcome of TL is contradictory to the goal of identification, which is to cluster the samples according to their identities.

D. STL

For a recognition task, if the labeling information is available, by integrating the labeling information with TL, Jian Lai *et al.* proposed a method named STL. Their method is as following

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 + \lambda \|\mathbf{G}\mathrm{Diag}(\mathbf{x})\|_* \tag{6}$$

in which a class dependent matrix $\mathbf{G} \in R^{m\times n}$ is introduced in the trace lasso term, where $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \cdots, \mathbf{G}_c]$ and $\mathbf{G}_i \in R^{m\times n_i}$, here n_i is the number of training samples in the ith class and c is the number of training classes. In $\mathbf{G}_i$, all elements in the ith row are one and those in the other rows are zero. With $\mathbf{G}$, the correlation of column vectors within the class is one and that between the classes is zero. This method can cluster the samples from the same subject but with different variation information together, which conforms to the goal of image recognition.

In all the above methods, CR is used, that is, the query image is considered as a linear combination of all training samples. The query image lies in the linear subspace spanned by the training data from the same subject. When this subspace has sufficient samples and can be expanded by these samples, namely it is complete, the query sample can be faithfully represented and representation error approaches zero. Unfortunately, sometimes image classification may be a typical small-sample-size problem, even the amount of samples may not meet the completeness requirement, not to mention having to label all the training samples. When only a small number of labeled images are available, they will lead to wrong classification results.

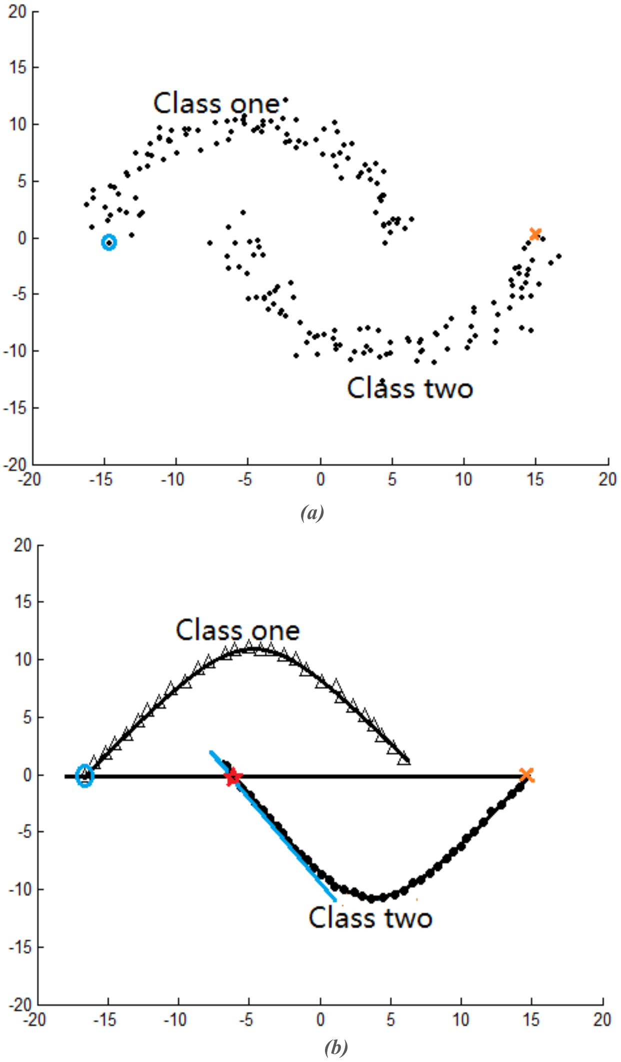
III. PROPOSED APPROACH

A. THE PROPOSED MODEL

Nowadays, it is easy to collect unlabeled samples because of the convenience supplied by Internet. MRSSL successfully exploits the local structure of data distribution including both labeled and unlabeled samples. With the unlabeled samples, which are from various different subjects, the number of samples from the same class with the query is firstly increased and therefore the representation ability is improved. Besides, the unlabeled images can be automatically labeled using the proposed model rather than manual participation.

Manifold usually means the graph locally having the property of Euclidean space. We assume all samples lie on a low dimensional manifold which is embedded in a high dimensional Euclidean space. Images of the same class have the same label and lie on a sub-manifold. Since manifold locally can be approximated as a linear space, any point on it can be approximated by the linear combination of the neighboring points. Consider the query sample $\mathbf{y} \in R^m$ as a collaborative representation of all training samples $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in R^{m\times n}$, then only the linear representation coefficients, which are correlated with the data on the same sub-manifold with $\mathbf{y}$ and at the neighborhood of $\mathbf{y}$, are nonzero and the other coefficients are all zero. This is equivalent to find a kind of sparse representation of $\mathbf{y}$ about all training samples.

Fig.1 (a) is a practical example where there are two classes of data and only two samples (one for each class) are labeled. These two labeled data are marked with blue circle and orange cross respectively. The other points are all unlabeled so we need to find the labels of all these data. This is a very difficult clustering task due to the lack of labeled data. For the convenience of illustration, Fig.1 (a) is simply shown as Fig.1 (b). The data in Class one are shown as small triangles and data in Class two are shown as black dots. The two curves represent two sub-manifolds associated with the two classes. In full-supervised case, that is, only labeled samples can be used. Therefore the point marked with red star can only be represented with the blue circle point and orange cross point. Since these three points are in the same linear space (the black straight line through the three points), the red star point can be represented as the linear combination of the blue circle point and orange cross point. And both representation coefficients may not be zero, which implies the red star point can be classified to Class one or Class two. This may lead to a wrong classification result. While as the unlabeled points are added and under the assumption of image manifold, the red star point can be approximately represented as the linear combination of the points on the tangent plane of the sub-manifold (the blue straight line), and all these points are from Class two. This is the classification result we expect.



(a)

(b)

FIGURE 1. An example of data classification. (a) Two classes of data with only two labeled data, (b) The sketch of (a).

We measure the reconstruction error with l_1-norm, which is much more robust than l_2-norm to handle real-world contamination.

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 \tag{7}$$

There are two regularization terms in our model.

The first term is the manifold Laplacian regularization

$$\frac{\upsilon}{2}\sum\nolimits_{i,j=1}^{n}\left\|\mathbf{z}_i-\mathbf{z}_j\right\|_2^2 S_{ij} \tag{8}$$

υ is the regularization parameter used to adjust the smoothness of manifold. Assume $\mathbf{S} = \left(S_{ij}\right)_{n\times n}$ is a matrix with element S_{ij} being the similarity between two samples $\mathbf{a}_i$ and $\mathbf{a}_j$. The similarity matrix is used to obtain the labels of the unlabeled training samples. Let

$$S_{ij}=\begin{cases}\exp\left(\dfrac{-\left\|\mathbf{a}_i-\mathbf{a}_j\right\|^2}{\sigma^2}\right), & \mathbf{a}_i,\mathbf{a}_j \text{ is } k\text{-nearest neighbor}\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

For alleviating the number of parameters, here the similarity S_{ij} only takes value 0 or 1. $\mathbf{S}$ can be simplified as

$$S_{ij}=\begin{cases}1, & \mathbf{a}_i,\mathbf{a}_j \text{ is } k\text{-nearest neighbor}\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

(10) is the exceptional case of (9) as $\sigma \to \infty$. Equation (8) means when similarity degree between $\mathbf{a}_i$ and $\mathbf{a}_j$ is 1, their labels should be as same as possible.

Using all the training samples $\mathbf{a}_j\,(j = 1, 2, \cdots, n)$ as nodes, $\mathbf{a}_i$ and $\mathbf{a}_j$ have a connection between them as $S_{ij} = 1$ and no connection as $S_{ij} = 0$. Then we can obtain a graph of all the samples. Since for each sample the most similar sample must come from the same class with it, under an appropriate threshold k (note that a small k will be fine in formula (10)), each node must connect with at least one node on the same sub-manifold. Then through the function of manifold Laplacian regularization (8), the label can be propagated from the labeled nodes to unlabeled nodes along the connections. The connected nodes (samples cluster) therefore can share the same label. This can be simply illustrated by Fig.2, where there are two classes of data and only two are labeled (one for each class and tagged with blue circle and orange cross separately). The other points are all unlabeled. For each class the label can be properly propagated to unlabeled data along the connections because of the function of the manifold Laplacian regularization.

The second regularization, a variant Trace Lasso norm is proposed as follows

$$\left\|\mathbf{Z}\mathrm{Diag}(\mathbf{x})\right\|_* \tag{11}$$

The TL term $\left\|\mathbf{Z}\mathrm{Diag}(\mathbf{x})\right\|_*$ can be considered as an approximation to the rank of $\mathbf{Z}\mathrm{Diag}(\mathbf{x})$. We set $\mathbf{Z}_i = \left[\mathbf{z}_1^i, \mathbf{z}_2^i, \cdots, \mathbf{z}_{n_i}^i\right] \in R^{c\times n_i}$ is composed of the label vectors of all samples from ith class. Since the label can be accurately propagated from labeled data to unlabeled data among the samples on the same sub-manifold, $\mathbf{Z}_i$ will have the structure that all the elements in ith row are one and those in other rows are zero. Therefore, the formula (11) can automatically seek a sparsity of the number of classes, which means the query image is represented by a small number of groups. Once one class is selected, it is in favor of using more samples from the same class, just as that illustrated in [8]. Therefore, the second regularization forces the group sparsity, and in each group, the training samples are fully used. This sparsity between groups and density within a group are preferred to the aim of image classification.

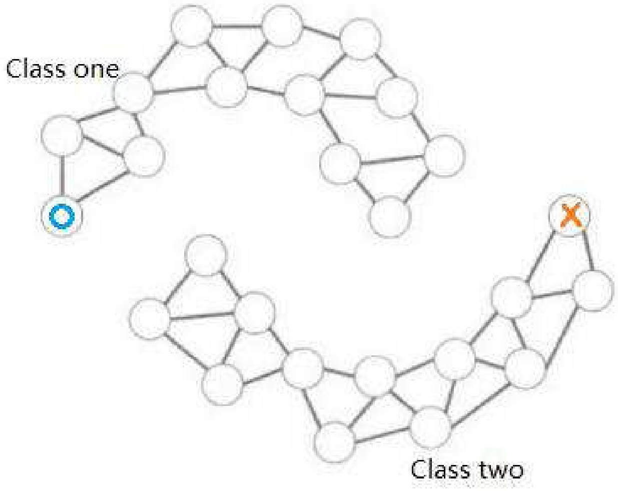


FIGURE 2. Semi-supervised classification on manifold.

As a sum of above, the complete model we propose is

$$\begin{aligned}\underset{\mathbf{x},\,\mathbf{Z}}{\arg\min}\ &\left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_1+\lambda\left\|\mathbf{Z}\mathrm{Diag}\left(\mathbf{x}\right)\right\|_*\\ &+\frac{\upsilon}{2}\sum\nolimits_{i,j=1}^{n}\left\|\mathbf{z}_i-\mathbf{z}_j\right\|_2^2 S_{ij}+\sum\nolimits_{j=1}^{n}U_{jj}\left\|\mathbf{z}_j-\mathbf{g}_j\right\|_2^2\end{aligned} \tag{12}$$

This model is a generalization of STL. If all the training samples are labeled, namely $\mathbf{Z}$ is known, the third and fourth terms will automatically disappear, then the formula (12) is the same with that of STL. If all training samples or a part of them are unlabeled, we can obtain the unknown labels at the same time of identifying the query image by (12).

Assume $\mathbf{G} = \left(\mathbf{g}_j\right)_{j=1}^n \in R^{c\times n}$ and $D_{jj} = \sum_{i=1}^n S_{ij}$, $\mathbf{D}$ is a diagonal matrix, $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the graph Laplacian matrix. Then (12) can be reformulated as

$$\begin{aligned}\underset{\mathbf{x},\,\mathbf{Z}}{\arg\min}\ &\left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_1+\lambda\left\|\mathbf{Z}\mathrm{Diag}\left(\mathbf{x}\right)\right\|_*\\ &+\upsilon\mathrm{tr}\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right)+\mathrm{tr}((\mathbf{Z}-\mathbf{G})\,\mathbf{U}\,(\mathbf{Z}-\mathbf{G})^{\mathrm{T}})\end{aligned} \tag{13}$$

B. OPTIMIZATION

Since the first two terms of formula (13) are not differentiable, this makes it impossible to achieve the solution directly through optimization methods such as gradient descent. The original problem is converted to the following equivalent constrained problem

$$\begin{aligned}\underset{\mathbf{e},\,\mathbf{J},\,\mathbf{Z},\,\mathbf{x}}{\arg\min}\ &\left\|\mathbf{e}\right\|_1+\lambda\left\|\mathbf{J}\right\|_*+\upsilon\mathbf{tr}\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right)\\ &+\mathrm{tr}((\mathbf{Z}-\mathbf{G})\,\mathbf{U}\,(\mathbf{Z}-\mathbf{G})^{\mathrm{T}})\\ \text{s.t.}\ \ &\mathbf{e}=\mathbf{y}-\mathbf{A}\mathbf{x},\ \ \mathbf{J}=\mathbf{Z}\mathrm{Diag}(\mathbf{x})\end{aligned} \tag{14}$$

We use the ALM scheme to derive the following unconstrained optimization problem

$$\begin{aligned}\arg\min_{\mathbf{e},\,\mathbf{J},\,\mathbf{Z},\,\mathbf{x}} \mathscr{L}(\mathbf{e},\ \mathbf{J},\ \mathbf{Z},\ \mathbf{x}) = \|\mathbf{e}\|_1 + \lambda\,\|\mathbf{J}\|_* + \upsilon\mathbf{tr}\left(\mathbf{ZLZ}^{\mathrm{T}}\right)\\ +\mathrm{tr}((\mathbf{Z}-\mathbf{G})\,\mathbf{U}\,(\mathbf{Z}-\mathbf{G})^{\mathbf{T}}) + \boldsymbol{\theta}^{\mathrm{T}}\,(\mathbf{y}-\mathbf{Ax}-\mathbf{e})\\ +\mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\,(\mathbf{Z}\mathrm{Diag}\,(\mathbf{x})-\mathbf{J})\right)\\ +\frac{\mu}{2}\left(\|\mathbf{y}-\mathbf{Ax}-\mathbf{e}\|_2^2 + \|\mathbf{Z}\mathrm{Diag}\,(\mathbf{x})-\mathbf{J}\|_{\mathrm{F}}^2\right) \qquad (15)\end{aligned}$$

where $\mathbf{Y} \in R^{c\times n}$ and $\boldsymbol{\theta} \in R^m$ are the Lagrangian multipliers, $\mu > 0$ is the penalty parameter. Instead of optimizing all arguments simultaneously, we solve them individually and iteratively using GS-ADMM.

By fixing $\mathbf{J}$, $\mathbf{Z}$, $\mathbf{x}$, we optimize $\mathbf{e}$ by the following subproblem

$$\begin{aligned}\arg\min_{\mathbf{e}} \|\mathbf{e}\|_1 + \boldsymbol{\theta}^{\mathrm{T}}\,(\mathbf{y}-\mathbf{Ax}-\mathbf{e}) + \frac{\mu}{2}\,\|\mathbf{y}-\mathbf{Ax}-\mathbf{e}\|_2^2\\ = \arg\min_{\mathbf{e}} \|\mathbf{e}\|_1 + \frac{\mu}{2}\left\|\mathbf{e}-\left(\mathbf{y}-\mathbf{Ax}+\frac{\boldsymbol{\theta}}{\mu}\right)\right\|_2^2 \qquad (16)\end{aligned}$$

The solution of (16) can be achieved via soft-thresholding.

To update $\mathbf{J}$, the following sub-problem is solved

$$\begin{aligned}\arg\min_{\mathbf{J}} \lambda\,\|\mathbf{J}\|_* + \mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\,(\mathbf{Z}\mathrm{Diag}(\mathbf{x})-\mathbf{J})\right)\\ +\frac{\mu}{2}\,\|\mathbf{Z}\mathrm{Diag}(\mathbf{x})-\mathbf{J}\|_{\mathrm{F}}^2\\ = \arg\min_{\mathbf{J}} \lambda\,\|\mathbf{J}\|_* + \frac{\mu}{2}\left\|\mathbf{J}-\left(\mathbf{Z}\mathrm{Diag}(\mathbf{x})+\frac{\mathbf{Y}}{\mu}\right)\right\|_{\mathrm{F}}^2 \qquad (17)\end{aligned}$$

Problem (17) can be solved by singular value thresholding operator.

The optimized $\mathbf{x}$ can be obtained as

$$\begin{aligned}\arg\min_{\mathbf{x}} \boldsymbol{\theta}^{\mathrm{T}}\,(\mathbf{y}-\mathbf{Ax}-\mathbf{e}) + \mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\,(\mathbf{Z}\mathrm{Diag}(\mathbf{x})-\mathbf{J})\right)\\ +\frac{\mu}{2}\left(\|\mathbf{y}-\mathbf{Ax}-\mathbf{e}\|_2^2 + \|\mathbf{Z}\mathrm{Diag}(\mathbf{x})-\mathbf{J}\|_{\mathrm{F}}^2\right) \qquad (18)\end{aligned}$$

This problem can be solved by solving the following linear system

$$\begin{aligned}\mu\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathrm{Diag}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\right)\mathbf{x} = \mathrm{diag}\left(\mu\mathbf{J}^{\mathrm{T}}\mathbf{Z} - \mathbf{Y}^{\mathrm{T}}\mathbf{Z}\right)\\ +\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta} + \mu\mathbf{A}^{\mathrm{T}}\,(\mathbf{y}-\mathbf{e}) \qquad (19)\end{aligned}$$

As the left multiplied matrix $(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathrm{Diag}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}))$ is inversible, $\mathbf{x}$ can be solved directly.

By fixing $\mathbf{e}$, $\mathbf{J}$, $\mathbf{x}$, we optimize $\mathbf{Z}$ by the following subproblem

$$\begin{aligned}\arg\min_{\mathbf{Z}} \upsilon\mathbf{tr}\left(\mathbf{ZLZ}^{\mathrm{T}}\right) + \mathrm{tr}((\mathbf{Z}-\mathbf{G})\,\mathbf{U}\,(\mathbf{Z}-\mathbf{G})^{\mathrm{T}})\\ +\mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Z}\mathrm{Diag}(\mathbf{x})\right) + \frac{\mu}{2}\,\|\mathbf{Z}\mathrm{Diag}(\mathbf{x})-\mathbf{J}\|_{\mathrm{F}}^2 \qquad (20)\end{aligned}$$

It can be solved using the following equation

$$\begin{aligned}\mathbf{Z}\left(\mu\mathrm{Diag}(\mathbf{x})^2 + 2\upsilon\mathbf{L} + 2\mathbf{U}\right) = \mu\mathbf{J}\mathrm{Diag}(\mathbf{x})\\ -\mathbf{Y}\mathrm{Diag}(\mathbf{x}) + 2\mathbf{GU} \qquad (21)\end{aligned}$$

TABLE 1. Algorithm 1.

```
Input: Training set $\mathbf{A}$ ; query image $\mathbf{y}$ ; known labels $\mathbf{g}_j\ (j \in S)$ ; parameters $\lambda$ , $\upsilon$ , $k$ ; $\mu_{\max} = 10^6$ , $\rho = 1.1$ , $\varepsilon = 10^{-7}$ .
Initialization: $\mu^0 = 10^{-6}$ , and all the other variables are initialized to be zero vector or zero matrix. For $i = 0,1,2,\cdots$
    while $\left\|\mathbf{J}^{i+1} - \mathbf{Z}^{i+1}\mathrm{Diag}\left(\mathbf{x}^{i+1}\right)\right\|_\infty \ge \varepsilon$ or $\left\|\mathbf{y} - \mathbf{A}\mathbf{x}^{i+1} - \mathbf{e}^{i+1}\right\|_\infty \ge \varepsilon$ do
        1. Update $\mathbf{e}^{i+1} = \arg\min_{\mathbf{e}} \mathscr{L}\left(\mathbf{e}, \mathbf{J}^i, \mathbf{Z}^i, \mathbf{x}^i; \boldsymbol{\theta}^i, \mathbf{Y}^i\right)$ as problem (16).
        2. Update $\mathbf{J}^{i+1} = \arg\min_{\mathbf{J}} \mathscr{L}\left(\mathbf{e}^{i+1}, \mathbf{J}, \mathbf{Z}^i, \mathbf{x}^i; \boldsymbol{\theta}^i, \mathbf{Y}^i\right)$ as problem (17).
        3. Solve $\mathbf{x}^{i+1} = \arg\min_{\mathbf{x}} \mathscr{L}\left(\mathbf{e}^{i+1}, \mathbf{J}^{i+1}, \mathbf{Z}^i, \mathbf{x}; \boldsymbol{\theta}^i, \mathbf{Y}^i\right)$ using equation (19).
        4. Solve $\mathbf{Z}^{i+1} = \arg\min_{\mathbf{Z}} \mathscr{L}\left(\mathbf{e}^{i+1}, \mathbf{J}^{i+1}, \mathbf{Z}, \mathbf{x}^{i+1}; \boldsymbol{\theta}^i, \mathbf{Y}^i\right)$ by equation (21).
        5. Update the multipliers
            $\mathbf{Y}^{i+1} == \arg\min_{\mathbf{Y}} \mathscr{L}\left(\mathbf{e}^{i+1}, \mathbf{J}^{i+1}, \mathbf{Z}^{i+1}, \mathbf{x}^{i+1}; \boldsymbol{\theta}^i, \mathbf{Y}\right)$
        and
            $\boldsymbol{\theta}^{i+1} = \arg\min_{\boldsymbol{\theta}} \mathscr{L}\left(\mathbf{e}^{i+1}, \mathbf{J}^{i+1}, \mathbf{Z}^{i+1}, \mathbf{x}^{i+1}; \boldsymbol{\theta}, \mathbf{Y}^{i+1}\right)$
        by (22), (23).
        6. Update the parameter $\mu$ by $\mu^{i+1} = \min\left(\rho\mu^i, \mu_{\max}\right)$ .
    end
Output: Coefficient vector $\mathbf{x}$ , the unknown label vectors $\mathbf{z}_j\ (j \notin S)$ .
```

The Lagrangian multipliers are updated as

$$\mathbf{Y} = \mathbf{Y} + \mu\,(\mathbf{Z}\mathrm{Diag}(\mathbf{x}) - \mathbf{J}) \qquad (22)$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \mu\,(\mathbf{y} - \mathbf{Ax} - \mathbf{e}) \qquad (23)$$

The steps (16), (17), (19), (21), (22), (23) are repeated until the convergence conditions are attained. Algorithm 1 summarizes the procedures to solve the optimization problem. The numerical experiments in the next section can confirm the convergence of this algorithm.

C. CLASSIFICATION

Once the matrix $\mathbf{Z}$ is obtained, for the unlabeled data, the element $|Z_{ij}|$ describes the probability of jth data belonging to ith class. If Z_{ij} is the element with the largest absolute value in vector $\mathbf{z}_j$, we set $Z_{ij} = 1$ and all other elements are set to zero.

We classify the query sample according to the representative coefficients vector $\mathbf{x}$. The l_1-norm is still used to measure the reconstruction error to be consistent with the first term of (13). The reconstruction error of each class is

$$r\,(i) = \left\|\mathbf{y} - \mathbf{A} \times \left(\mathbf{Z}^i \times \mathrm{diag}\,(\mathbf{x})\right)^{\mathrm{T}}\right\|_1 \qquad (i = 1, 2, \cdots, c) \qquad (24)$$

Here, product $\mathbf{Z}^i \times \operatorname{diag}(\mathbf{x})$ is to extract the representation coefficients correlated to ith class. $r(i)$ is the representative error of using all the training samples from ith class to represent $\mathbf{y}$. Finally, the query sample is labeled to the class with the minimum residual as following

$$i^* = \arg\min_i r(i) \tag{25}$$

IV. EXPERIMENTAL RESULTS

As an important application of image classification, face recognition is mainly considered in this section. Of course, our method can be extensively applied to other data classification task as long as the data distribution conforms to the manifold assumption. The proposed method is compared with the state-of-the-art approaches including SRC, RCRC, TL, STL, nuclear norm based matrix regression (NMR) classification [25], weighted group sparse classifier (WGSC) [26], iterative re-constrained group sparse classification (IRGSC) [27]. We use three popular face databases: Extended Yale B database [28], AR Face Database [29] and ORL [30]. For the first two databases and the methods SRC, TL and STL, we use the similar setting as that used in [8] and directly cite some results reported in [8]. We compute the recognition accuracy (RA) as

$$\text{Recognition Accuracy} = \frac{\begin{array}{c}\text{Number of correctly}\\ \text{recognized testing data}\end{array}}{\begin{array}{c}\text{Total number of}\\ \text{testing data}\end{array}} \tag{26}$$

The average RA (ARA) are the results of over 10 runs across various methods for each testing image of every subject. We directly utilize the grey level as the feature in all experimental scenarios for all approaches. The best results are shown with bold font in all the tables below.

There are three parameters needed to be tuned in our method: λ, υ and k, where λ and υ are used to balance the roles of two regularization terms. k is the parameter used to choose the most relevant samples for each data in formula (10). Because the label can be accurately propagated from labeled data to unlabeled data by manifold Laplacian regularization, k can take a small value. In the following experimental scenarios, $k = 2$ achieves good results. Fig.3 shows variations of ARA with parameters λ and υ. Here for each subject from Extended Yale B database we randomly choose 13 images with 8 labeled images and 5 unlabeled images as training set. Other 32 images are used for testing. Then we run our algorithm 10 times and calculate ARA. We can see when the value of λ is taken from interval $[1, 10]$, the highest ARA can be obtained, here paramenter υ is fixed as 1. As $\lambda < 1$, ARA rapidly decreases. In the same way, we can get the best choice for υ is $[1, 22]$ as fixing $\lambda = 1$. The ARA is more sensitive to small λ and υ. Experiments show these choices also achieve the best results in all the following experimental setting. The parameters for each other method are also finely tuned to achieve its best result.

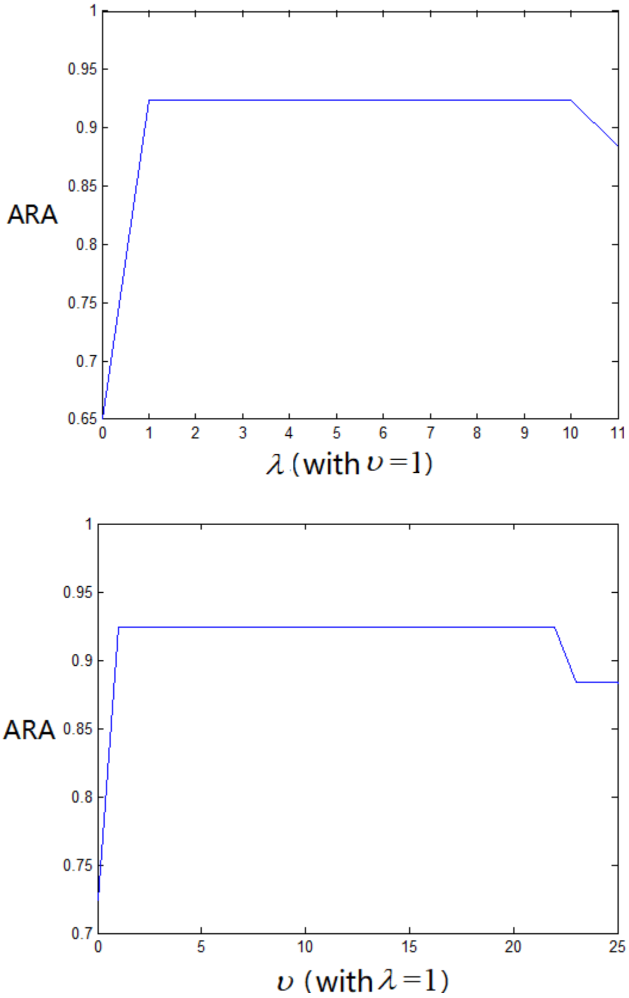


FIGURE 3. Variations of ARA with each parameter.

A. EXTENDED YALE B DATABASE

There are 38 subjects in Extended Yale B database. Each subject includes about 64 face images captured under different illuminations. All the images are down-sampled to 48*42. For per subject, we randomly select $t = 8$ images for training for the methods SRC, RCRC, TL, IRGSC, NMR, WGSC and STL. These are all fully supervised methods, which means they require all the training samples to be labeled. Based on the t labeled images, unlabeled samples are added then we check the recognition effect of our method. The number of unlabeled training images is denoted as s, therefore the training images for our method is $t + s$ in total. 32 images are used for testing for all the eight methods. The ARA are reported in Table 2. With the choice of $s = 24$, our method can achieve best result than the other seven methods.

Table 3 shows the influence of the value taken for s on ARA in our method. When s equals to 0, namely our model degenerates to that of STL, the ARA of our method is the same with that of STL method. While when the unlabeled samples are added, our method achieves higher and higher ARA with increase of the number of unlabeled images.

TABLE 2. Comparison of ARA on extended Yale B database.

METHODS		ARA(%)
SRC	$t=8$	84.94
RCRC		85.96
TL		86.13
IRGSC		89.07
NMR		86.67
WGSC		88.76
STL		90.83.
OURS($s=24$)		**96.68**

TABLE 3. Influence of *s* on ARA.

s	ARA(%)
0	90.83
5	92.39
10	94.28
15	95.94
24	**96.68**

TABLE 4. Comparison of ARA on AR database.

METHODS		ARA(%)
SRC	$t=4$	87.5
RCRC		88.53
TL		89.56
IRGSC		89.56
NMR		89.08
WGSC		88.75
STL		91.56
OURS($s=3$)		**95.36**

B. AR DATABASE

AR database includes 126 subjects. For each subject, 26 face images are taken in two separate sessions. Each session is with the expression, illumination and disguise variation. In this paper, a subset of 100 subjects is used with each subject getting 14 images selected and only with expression or illumination changing. All images are down-sampled to 50*40. For each subject, $t = 4$ face images from Session 1 are used for training for the methods SRC, RCRC, TL, IRGSC, NMR, WGSC and STL, and all these images are labeled. $s = 3$ unlabeled images of Session 1 are added to form the training set of our method. All the samples from Session 2 are used for testing. Table 4 shows the results of all involved methods. With the addition of the unlabeled samples, our method can achieve better results than all the other seven methods.

C. ORL DATABASE

The ORL data set consists of face images of 40 distinct subjects, each subject having 10 face images under varying lighting conditions, with different facial expressions and facial details. In our experiment each image is down-sampled from 112×92 to 32×32. For each subject, $t = 3$ labeled face images are used for training for the methods SRC, RCRC, TL, IRGSC, NMR, WGSC and STL, and $s = 2$ unlabeled images are added for training in our method. 5 images are used for testing. Table 5 gives the ARA of different methods. We can observe that our method can get better classification results than other methods due to the addition of the unlabeled samples, which further confirms the role of the unlabeled data.

TABLE 5. Comparison of ARA on ORL database.

METHODS		ARA(%)
SRC	$t=3$	71.13
RCRC		72.33
TL		73.79
IRGSC		72.76
NMR		73.47
WGSC		72.24
STL		75.76
OURS($s=2$)		**76.36**

V. CONCLUSION AND DISCUSSION

For the small-sample-size case especially when only a small number of labeled images are available and for the use of the unlabeled samples, a semi-supervised sparse image classification technique is proposed. The query image is collaboratively represented by the whole training data, whether they are labeled or unlabeled. Based on the assumption that images of the same class lie on a sub-manifold and the local linear property of manifold, an image can be approximately represented as the linear combination of its neighbouring data. There are two regularization terms. A generalized trace lasso regularization term is proposed by combing semi-supervised samples with a variant trace lasso norm. This term seeks the sparsity of the number of classes instead of the number of training samples, which directly coincides with the objective of data classification. By using manifold Laplacian regularization, the label of labeled images can be propagated to unlabeled images within a class along the distance of samples manifold. Both aims of image recognition and finding out the unknown identities of samples are achieved simultaneously. ALM Method and GS-ADMM are applied to solve the whole model.

Nowadays a discussion hot point in computational imaging is if it is the time to discard the classic methods and fully replace them by deep learning based methods. On the one hand, a prerequisite for deep learning based methods is a huge amount of samples. However, there are indeed some situations where there are only a small number of samples, at this time the knowledge based modeling methods are more suitable. On the other hand, classical methods have clear structure and theoretical guarantee. They are based on the knowledge of the problem we are trying to solve rather than seeking for best performance by intuitively choosing architectures or trial an error. In the future work, it is possibly better to integrate the classical knowledge based approaches into the deep learning architecture, making the algorithm enjoy both the flexibility of the deep learning based methods and the clear structure of the classical approaches. For example, the result of our algorithm is dependent on the selection of the similarity matrix $\mathbf{S}$, if $\mathbf{S}$ is not properly selected the label can't be accurately propagated. We will try to solve this problem and all the parameters that need to be determined by designing a deep network.

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] R. Rigamonti, M. A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" in *Proc. CVPR*, Jun. 2011, pp. 1545–1552.

[3] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. CVPR*, Jun. 2011, pp. 553–560.

[4] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L1-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, Mar. 2012.

[5] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.

[6] L. Zhang, M. Yang, X. C. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," *Comput. Sci.*, vol. 321, pp. 276–283, 2012.

[7] E. Grave, G. Obozinski, and F. Bach, "Trace lasso: A trace norm regularization for correlated designs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2187–2195.

[8] J. Lai and X. Jiang, "Supervised trace lasso for robust face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[9] Y. Li, G. Wang, L. Nie, Q. Wang, and W. Tan, "Distance metric optimization driven convolutional neural network for age invariant face recognition," *Pattern Recognit.*, vol. 75, pp. 51–62, Mar. 2018.

[10] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 964–975, Feb. 2018.

[11] Y. Zhao, Z. Zhen, X. Liu, Y. Song, and J. Liu, "The neural network for face recognition: Insights from an fMRI study on developmental prosopagnosia," *NeuroImage*, vol. 169, pp. 151–161, Apr. 2018.

[12] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "Towards pose invariant face recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2207–2216.

[13] Y. Jun, Y. Xiaokang, G. Fei, and T. Dacheng, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4014–4024, Dec. 2017.

[14] H.-M. Lu, Y. Fainman, and R. Hecht-Nielsen, "Image manifolds," in *Proc. SPIE, Appl. Artif. Neural Netw. Image Process. III*, vol. 3307, Apr. 1998.

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

[16] J. Tang, L. Shao, X. Li, and K. Lu, "A local structural descriptor for image matching via normalized graph Laplacian embedding," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 410–420, Feb. 2016.

[17] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1227–1233, Jun. 2012.

[18] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, Jul. 2011.

[19] W. Liu, X. Ma, Y. Zhou, D. Tao, and J. Cheng, "*p*-Laplacian regularization for scene recognition," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2927–2940, Aug. 2019.

[20] X. Ma, W. Liu, S. Li, D. Tao, and Y. Zhou, "Hypergraph *p*-Laplacian regularization for remotely sensed image recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1585–1595, Mar. 2019.

[21] W. Liu, X. Yang, D. Tao, J. Cheng, and Y. Tang, "Multiview dimension reduction via hessian multiset canonical correlations," *Inf. Fusion*, vol. 41, pp. 119–128, May 2018.

[22] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, Jul. 2007.

[23] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[24] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast l1-minimization algorithms for robust face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234–3246, May 2013.

[25] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[26] X. Tang, G. Feng, and J. Cai, "Weighted group sparse representation for undersampled face recognition," *Neurocomputing*, vol. 145, pp. 402–415, Dec. 2014.

[27] J. Zheng, P. Yang, S. Chen, G. Shen, and W. Wang, "Iterative reconstrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, May 2017.

[28] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[29] A. Martinez and R. Benavente, "The AR face database," CVC, Bellatera, Spain, Tech. Rep. 24, Jun. 1998.

[30] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

WENJUAN ZHANG received the M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2005 and 2013, respectively. She was a Visiting Scholar with the Department of Mathematics, University of Florida. She is currently an Associate Professor with the School of Science, Xi'an Technological University. Her research interests include applications of variation and regularization methods, partial differential equations in image segmentation, and low-rank and sparse approximation for image processing.

XIANGCHU FENG received the B.S. degree in computational mathematics from Xian Jiaotong University, Xi'an, China, in 1984, and the M.S. and Ph.D. degrees in applied mathematics from Xidian University, Xi'an, in 1989 and 1999, respectively. He is currently a Professor of applied mathematical with the School of Mathematics and Statistics, Xidian University. His research interests include numerical analysis, wavelets, low-rank matrix approximation, and partial differential equations for image processing.

YUNMEI CHEN (Member, IEEE) received the B.S. degree in mathematics from Tongji University, Shanghai, China, in 1967, and the M.S. and Ph.D. degrees in mathematics from Fudan University, Shanghai, in 1981 and 1985, respectively. She is currently a Distinguished Professor of mathematics with the Department of Mathematics, University of Florida. Her research interests include partial differential equations (PDE)/variational methods and methods of machine learning for image processing, optimization techniques and applications in imaging and machine learning, medical image analysis, PDE, and nonlinear analysis.