

Laplacian sparse dictionary learning for image classification based on sparse representation^{*}

Fang LI^{1,2,3}, Jia SHENG¹, San-yuan ZHANG^{†‡1}

⁽¹⁾College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

⁽²⁾School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China)

⁽³⁾Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)

[†]E-mail: syzhang@zju.edu.cn

Received Jan. 25, 2016; Revision accepted May 12, 2016; Crosschecked Nov. 20, 2017

Abstract: Sparse representation is a mathematical model for data representation that has proved to be a powerful tool for solving problems in various fields such as pattern recognition, machine learning, and computer vision. As one of the building blocks of the sparse representation method, dictionary learning plays an important role in the minimization of the reconstruction error between the original signal and its sparse representation in the space of the learned dictionary. Although using training samples directly as dictionary bases can achieve good performance, the main drawback of this method is that it may result in a very large and inefficient dictionary due to noisy training instances. To obtain a smaller and more representative dictionary, in this paper, we propose an approach called Laplacian sparse dictionary (LSD) learning. Our method is based on manifold learning and double sparsity. We incorporate the Laplacian weighted graph in the sparse representation model and impose the l_1 -norm sparsity on the dictionary. An LSD is a sparse overcomplete dictionary that can preserve the intrinsic structure of the data and learn a smaller dictionary for each class. The learned LSD can be easily integrated into a classification framework based on sparse representation. We compare the proposed method with other methods using three benchmark-controlled face image databases, Extended Yale B, ORL, and AR, and one uncontrolled person image dataset, i-LIDS-MA. Results show the advantages of the proposed LSD algorithm over state-of-the-art sparse representation based classification methods.

Key words: Sparse representation; Laplacian regularizer; Dictionary learning; Double sparsity; Manifold
<https://doi.org/10.1631/FITEE.1600039>

CLC number: TP39


1 Introduction

Sparse representation has shown huge capabilities in handling problems such as computer vision, image processing, and visual tracking (Huang *et al.*, 2014; Lu and Li, 2014; Peleg and Elad, 2014; Zhu *et al.*, 2014; Zhang *et al.*, 2015). The core idea of sparse representation is to exploit a linear combination of some samples to represent the test

sample and then to calculate the representation solution that will be applied to reconstruct the desired results. In recent years in the image processing area, some models based on sparse representation have been proposed (Wright *et al.*, 2009; Yang *et al.*, 2010, 2012; Wang *et al.*, 2015). In these models, an input testing image is coded as a sparse linear combination of sample images or dictionary via l_1 -norm minimization. Because sample images or dictionaries are determinants in sparse representation, how to learn an optimal dictionary from training data becomes an important question worthy of further investigation (Shao *et al.*, 2014). Some sparse representation models that use the original image as the dictionary have shown promising results (Wright *et al.*, 2009; Qiao *et al.*, 2010), but there is a drawback with

[‡] Corresponding author

^{*} Project supported by the National Natural Science Foundation of China (Nos. 61272304 and 61363029) and the Guangxi Key Laboratory of Trusted Software (No. kx201313)

 ORCID: San-yuan ZHANG, <http://orcid.org/0000-0001-8604-874X>

© Zhejiang University and Springer-Verlag GmbH Germany 2017

these models. For example, original images have redundant, noisy and trivial information that can be negative to recognition. If the training samples are huge, the computation of sparse representation is time-consuming. Some attempts have been made to learn a compressed dictionary from training images and then use this learned dictionary for image analysis (Aharon *et al.*, 2006; Rubinstein *et al.*, 2010a; Yang *et al.*, 2010; 2014; Gao *et al.*, 2014).

In this paper, we focus on building a robust dictionary for sparse representation. We propose a new dictionary learning approach called Laplacian sparse dictionary (LSD), which is based on manifold embedding and double sparsity dictionary learning. After mapping a Laplacian weighted graph to the original sparse representations and imposing the l_1 -norm sparsity on the dictionary, a more compact and robust sparse dictionary is learned from the original images. This dictionary is then used to represent the input probe image. Since a Laplacian weighted graph can preserve the neighborhood structure of the data and a sparse dictionary shows much more stable performance, the learned LSD will be more representative for sparse representation and will achieve excellent reconstruction results.

2 Related work

Wright *et al.* (2009) proposed a typical sparse representation model that uses training samples as a dictionary in a technique called sparse representation-based classification (SRC). SRC assumes that the test sample can be represented by samples from the same class and will be classified as a member of the class, which leads to the minimum reconstruction error. Denote $\mathbf{X}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ as the training samples of the i th class and each column of \mathbf{X} as a sample vector. For a test sample $\mathbf{y} \in \mathbb{R}^m$ from the i th class, \mathbf{y} can be sufficiently represented by the linear combination of the samples from \mathbf{X}_i , i.e., $\mathbf{y} = \sum_{j=1}^{n_i} s_{i,j} \mathbf{x}_{i,j} = \mathbf{X}_i \mathbf{s}_i$, where $\mathbf{s}_i = [s_{i,1}, s_{i,2}, \dots, s_{i,n_i}]^T \in \mathbb{R}^{n_i}$ are the sparse coefficients.

The SRC algorithm (Wright *et al.*, 2009) is summarized as follows:

1. Initialize \mathbf{D} with each column unit normalized.
2. Compute the sparse coefficients \mathbf{s} using the least absolute shrinkage and selection operator (lasso)

(Tibshirani, 1996) given by

$$\hat{\mathbf{s}}_i = \arg \min_{\mathbf{s}} \|\mathbf{s}\|_1 \quad \text{s.t. } \mathbf{y} = \mathbf{X}\mathbf{s}. \quad (1)$$

3. Compute the residual error

$$\mathbf{r}_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\delta}_i(\hat{\mathbf{s}}_i)\|_2, \quad (2)$$

where $\boldsymbol{\delta}_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the characteristic function that selects the coefficients associated with the i th class. For $\mathbf{s} \in \mathbb{R}^n$, $\boldsymbol{\delta}_i(\mathbf{s}) \in \mathbb{R}^n$ is a new vector whose only non-zero entries are those in \mathbf{s} that are associated with class i . The given test sample \mathbf{y} can be approximated as $\hat{\mathbf{y}}_i = \mathbf{X}\boldsymbol{\delta}_i(\hat{\mathbf{s}}_i)$ by using only the coefficients associated with the i th class.

4. Output label $(\mathbf{y}) = \arg \min_i r_i(\mathbf{y})$.

In this model, no actual dictionary training is performed because the entire training samples are used directly as the dictionary. If the training set is small, this approach is computationally very efficient because there is no overhead for the learning of the dictionary. Using the minimum residual error to classify an unseen test sample is easily interpretable because the class of the subdictionary leading to the minimum residual error can be inspected and assigned as the class label of the test sample. However, this method has a drawback. Due to noisy training instances, using the training samples as the dictionary may result in a very large and possibly inefficient dictionary (Gangeh *et al.*, 2013), especially in applications with large training sets.

One way to deal with this drawback is to use a more compact and robust set of bases as the dictionary to represent the input query image. Yang *et al.* (2010) proposed an approach which learns a smaller subdictionary from each class of training samples and then combines subdictionaries into one dictionary. In Yang's Metaface approach, each subdictionary \mathbf{D}_i is learned using the training sample \mathbf{X}_i in class i using the formulation given as $\min_{\mathbf{D}_i, \mathbf{S}_i} \|\mathbf{X}_i - \mathbf{D}_i \mathbf{S}_i\|_F^2 + \lambda \|\mathbf{S}_i\|_1$, where \mathbf{S}_i is the matrix of sparse coefficients representing \mathbf{X}_i . Computed subdictionaries are eventually composed into one dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c]$, where c is the number of classes. This algorithm has a smaller dictionary and higher accuracy than the original SRC.

3 Laplacian sparse dictionary learning

The goal of our method is to construct a compact and representative dictionary for sparse representation. We take Yang's model as a starting point to introduce a novel dictionary learning method, which is based on two theories: manifold learning and double sparsity. First, we deem that high-dimensional images can be sparsely represented or coded by representative samples on a lower dimensional subspace or sub-manifold. These representative samples constitute the dictionary that we want to build. A good dictionary has to discover the geometric structure of the image data manifold and preserve its intrinsic structure as faithfully as possible. Second, a sparse dictionary has a compact representation, low complexity and can reduce overfitting.

3.1 Manifold learning

In recent years, a number of studies have shown that high-dimensional images may reside on a lower dimensional subspace or sub-manifold (Roweis and Saul, 2000; Tenenbaum *et al.*, 2000). Many manifold learning and subspace learning methods have been proposed (Turk and Pentland, 1991; Belhumeur *et al.*, 1997; He and Niyogi, 2003; He *et al.*, 2005; Elhamifar and Vidal, 2013) and successfully used (Lu X *et al.*, 2013; Lu Y *et al.*, 2015; Wang *et al.*, 2015). An efficient subspace learning algorithm should be able to discover the manifold structure of the image space. In many real-world problems, the local manifold structure is more important than the global Euclidean structure. A Laplacian Eigenmap (Belkin and Niyogi, 2001) arises by solving a variational problem that optimally preserves the neighborhood structure of the dataset and has discriminating power although it is unsupervised. It is likely that a nearest neighbor search in the low-dimensional space will yield similar results to one in the high-dimensional space. Since real-world images distribute on low-dimensional manifolds embedded in the high-dimensional ambient space, it is natural to discretely approximate the manifold by using a graph. The vertices of the graph correspond to the data samples and the edge weight of the graph represents the affinity between the data points. Constructing a Laplacian weighted graph is a key process to keep the local manifold structure and includes two main steps:

1. Constructing the adjacency graph: let G denote a graph with n nodes. Put an edge between nodes i and j if \mathbf{x}_i and \mathbf{x}_j are 'close'. There are two variations: (1) ε -neighborhoods ($\varepsilon \in \mathbb{R}$). Nodes i and j are connected by an edge if $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \varepsilon$ where the norm is the usual Euclidean norm in \mathbb{R}^n . (2) k -nearest neighbors ($k \in \mathbb{N}$). Nodes i and j are connected by an edge if i is among k -nearest neighbors of j or j is among k -nearest neighbors of i .

2. Choosing the weights: there are two variations for weighting the edges. \mathbf{W} is a sparse symmetric $n \times n$ matrix with W_{ij} having the weight of the edge joining vertices i and j , and 0 if there is no such edge.

A possible way of defining \mathbf{W} is as follows:

$$W_{ij} = \begin{cases} \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t), & \|\mathbf{x}_i - \mathbf{x}_j\|^2 < \varepsilon, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

or

$$W_{ij} = \begin{cases} \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t), & \text{if } \mathbf{x}_j \in N(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N(\mathbf{x}_j), \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $N(\mathbf{x}_i)$ denotes the k -nearest neighbors of \mathbf{x}_i , $t \in \mathbb{R}$.

One common assumption about the affinity between data points is the smoothness assumption, which says if two samples are close to each other in the input space, their corresponding outputs are close to each other (Chapelle *et al.*, 2006). According to the smoothness assumption, if data points \mathbf{x}_i and \mathbf{x}_j are close to each other, then their coefficients s_i and s_j should be close as well. Consistency with the geometry of the data which follows from the smoothness assumption motivates a regularizer term of the form:

$$\frac{1}{2} \sum_{i,j} (s_i - s_j) W_{ij} = \text{tr}(\mathbf{S}(\mathbf{D}\mathbf{D} - \mathbf{W})\mathbf{S}^T) = \text{tr}(\mathbf{S}\mathbf{L}\mathbf{S}^T), \quad (5)$$

where $\sum_j W_{i,j}$ is the degree of \mathbf{x}_i , $\mathbf{D}\mathbf{D} = \text{diag}(dd_1, dd_2, \dots, dd_n)$, and $\mathbf{L} = \mathbf{D}\mathbf{D} - \mathbf{W}$ is the Laplacian matrix.

To preserve locality information for the subdictionary, in our method, we incorporate the Laplacian regularizer (5) into the dictionary learning process. Since the Laplacian regularizer can preserve the local

structure, it serves as a data fidelity term in data representation. Our formulation is given as follows:

$$\min_{\mathbf{D}, \mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda \sum_i \|s_i\|_1 + \gamma \text{tr}(\mathbf{S}\mathbf{L}\mathbf{S}^T), \quad (6)$$

where γ is a positive scalar number.

3.2 Double sparsity

Building a dictionary for sparse signal representation involves balancing between complexity and adaptability. Rubinstein *et al.* (2010b) proposed a sparse dictionary model that satisfies these two considerations and has a simple and effective structure. This algorithm is based on the hypothesis that dictionary atoms themselves may have some underlying sparse structure over a more fundamental dictionary. Each atom of the proposed sparse dictionary has itself a sparse representation over some pre-specified base dictionary. Sparse dictionaries are efficient for large dictionaries and high-dimensional signals, show much more stable performance, and lead to higher compression rates. Motivated by Rubinstein's work, we introduced double sparsity into our work. Since natural images have high local redundancy, a sparse dictionary is useful because variations like expression, pose, session difference, and illumination usually lead to sparse changes in natural images (Yang M *et al.*, 2013). Some recent studies have shown that a learning dictionary, instead of off-the-shelf bases, leads to state-of-the-art performance in many applications (Yang *et al.*, 2010; Shao *et al.*, 2014). Employing sparse representations on an overcomplete dictionary with redundant information, the dictionary learning method has outperformed a pre-specified dictionary based on transformation functions. So, in our study, unlike in Rubinstein's model which learns a sparse dictionary from a fixed base dictionary, we learn sparse coefficients and a sparse dictionary simultaneously. We focus on a new overcomplete dictionary learning method in which the l_1 -norm sparsity is imposed not only on the coefficients but also on the dictionary atom. The sparse dictionary can be built by solving an alternative optimization problem.

3.3 Proposed method

In this section, we present a method for building an LSD which takes into account the local manifold

structure of the image data space and has a sparse structure.

3.3.1 Objective function

We want to build a dictionary that has a sparse structure and can better characterize the image's local information. To learn such a sparse dictionary from the original training database, we combine a Laplacian regularizer with Yang's model and impose the l_1 -norm sparse constraint on the dictionary. Let us denote $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ as the training samples of the i th class and each column of \mathbf{X} as a vector. The goal of our method is to learn an LSD $\mathbf{D}=[\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$ from \mathbf{X} , where $p \leq n$. It is required that each LSD vector \mathbf{d}_j ($j=1, 2, \dots, p$) should be a unit column vector. $\mathbf{S}=[\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n] \in \mathbb{R}^{p \times n}$ is the coefficient matrix where each column is a sparse representation for a data matrix. By incorporating the Laplacian regularizer into the original sparse coding and imposing the l_1 -norm sparsity on the dictionary, our objective function can be written as

$$J_{\mathbf{D}, \mathbf{S}} = \arg \min_{\mathbf{D}, \mathbf{S}} \left\{ \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda \sum_i \|s_i\|_1 + \beta \sum_j \|\mathbf{d}_j\|_1 + \gamma \text{tr}(\mathbf{S}\mathbf{L}\mathbf{S}^T) \right\} \quad \text{s.t. } \mathbf{d}_j^T \mathbf{d}_j = 1, \forall j, \quad (7)$$

where \mathbf{S} is the representation matrix of \mathbf{X} over the LSD \mathbf{D} . Parameters λ and β are positive scalar numbers that balance the sparsity and the reconstruction error. Parameter γ is the Laplacian coefficient and is also a positive scalar number. We let $\mathbf{d}_j^T \mathbf{d}_j = 1$ to avoid \mathbf{D} having an arbitrarily large l_2 -norm.

3.3.2 Laplacian sparse dictionary learning

Since the objective function (7) is not convex for \mathbf{D} and \mathbf{S} simultaneously, the minimization of Eq. (7) should be solved in an alternative manner over \mathbf{D} and \mathbf{S} . We split problem (7) into two sub-problems that are much easier to solve. The optimization processes can be iterated until converged. We iteratively optimize \mathbf{D} and \mathbf{S} using the following two-stage method:

Step 1: learn coefficient matrix \mathbf{S} with fixed dictionary \mathbf{D} .

When \mathbf{D} is fixed, the optimization function (7) can be rewritten as the following objective function:

$$\min_s \left\{ \|X - DS\|_F^2 + \lambda \sum_i \|s_i\|_1 + \gamma \text{tr}(SLS^T) \right\}. \quad (8)$$

The standard unconstrained optimization methods cannot be employed to address Eq. (8) because this problem with l_1 -norm is non-differentiable when s_i contains values of 0. Some sparse representation methods with l_1 -norm minimization have been proposed (Lee et al., 2006; Yang and Zhang, 2011; Yang J et al., 2012; Yang AY et al., 2013). Adopting an optimization method based on coordinate descent, it is obvious that Eq. (8) is convex and can obtain a global minimum. Instead of optimizing the whole sparse coefficient matrix S , we update each vector s_i one by one, while keeping all the other vectors s_j ($i \neq j$) fixed. To optimize the problem over each s_i , we rewrite Eq. (8) in a vector form (Zheng et al., 2011).

The reconstruction error $\|X - DS\|_F^2$ can be rewritten as

$$\sum_i \|x_i - Ds_i\|_2^2. \quad (9)$$

The Laplacian regularizer $\text{tr}(SLS^T)$ can be rewritten as

$$\text{tr}(SLS^T) = \text{tr} \left(\sum_{i,j} L_{ij} s_i s_j^T \right) = \sum_{i,j} L_{ij} s_j^T s_i = \sum_{i,j} L_{ij} s_i^T s_j, \quad (10)$$

where L_{ij} is the entry of the Laplacian matrix.

Combining Eqs. (9) and (10), Eq. (8) can be rewritten as

$$\min \sum_i \|x_i - Ds_i\|_2^2 + \lambda \sum_i \|s_i\|_1 + \gamma \sum_{i,j} L_{ij} s_i^T s_j. \quad (11)$$

When updating S_i , the other vectors $\{S_j\}_{j \neq i}$ are fixed. Thus, we obtain the following optimization problem:

$$\min_{s_i} f(s_i) = \|x_i - Ds_i\|_2^2 + \lambda \sum_j s_i^{(j)} + \gamma L_{ii} s_i^T s_i + s_i^T h_i, \quad (12)$$

where $h_i = 2\gamma \sum_{j \neq i} L_{ij} s_j$ and $s_i^{(j)}$ is the j th coefficient of s_i .

We solve problem (8) by following the feature-sign search algorithm proposed by Lee et al. (2006).

For details, please refer to Zheng et al. (2011).

Step 2: learn dictionary D with fixed coefficient matrix S .

When S is fixed, the optimization function (7) can be rewritten as the following objective function:

$$\min_D \|X - DS\|_F^2 + \beta \sum_j \|d_j\|_1 \quad \text{s.t. } d_j^T d_j = 1, \forall j. \quad (13)$$

To update the sparse dictionary atom by atom (Yang M et al., 2013), we rewrite S as $S = [s_1, \dots, s_j, \dots, s_p]$, where s_j is the j th row of S and p is the number of dictionary atoms. Eq. (13) can be written as follows:

$$\min_D \left\| X - \left(\sum_{j \neq k} d_j s_j + d_k s_k \right) \right\|_F^2 + \beta \sum_j \|d_j\|_1 \quad (14)$$

$$\text{s.t. } d_j^T d_j = 1, \forall j.$$

Let $Z = X - \sum_{j \neq k} d_j s_j$. By fixing all the other atoms d_j ($j \neq k$), the updating of d_k can be rewritten as

$$\min_{d_k} \|Z - d_k s_k\|_F^2 + \beta \|d_k\|_1 \quad \text{s.t. } \|d_k\|_2 = 1. \quad (15)$$

By using Lemma 1 of Rubinstein et al. (2010b) and letting $l = \|d_k\|_2$, Eq. (15) can be rewritten as

$$\min_{d_k} \|Zs_k^T / l^2 - d_k\|_F^2 + \frac{\beta}{\sqrt{l}} \|d_k\|_1 \quad \text{s.t. } \|d_k\|_2 = 1. \quad (16)$$

Let us take the derivative of the object function (16) and then set the derivative function as zero. Then d_k can be updated as

$$d_k = T_{\frac{\beta}{2\sqrt{l}}} (Zs_k^T / l^2) \left/ \left\| T_{\frac{\beta}{2\sqrt{l}}} (Zs_k^T / l^2) \right\|_2 \right., \quad (17)$$

where T_τ is a soft-threshold operator defined as

$$T_\tau(x) = \begin{cases} 0, & |x_\eta| \leq \tau, \\ x_\eta - \text{sign}(x_\eta)\tau, & \text{otherwise.} \end{cases} \quad (18)$$

The dictionary D is updated once all atoms d_k are updated.

3.3.3 Algorithm

The optimization procedures of the proposed LSD are described in Algorithm 1.

Algorithm 1 Laplacian sparse dictionary learning

Input: $\mathbf{x}=[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ where \mathbf{x}_i ($i=1, 2, \dots, n$) is the training sample of the i th object class.

Output: Laplacian sparse subdictionary \mathbf{D} .

1. Initialize each column of \mathbf{D} to have unit l_2 -norm.
2. Fix \mathbf{D} and solve \mathbf{S} using convex optimization techniques to solve the following function:

$$\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \lambda \sum_i \|\mathbf{s}_i\|_1 + \gamma \text{tr}(\mathbf{S}\mathbf{L}\mathbf{S}^T).$$

3. Fix \mathbf{S} and update \mathbf{D} . Update all the \mathbf{d}_j 's one by one by solving the function

$$\min_{\mathbf{d}_j} \|\mathbf{X} - \mathbf{D}\mathbf{S}\|_F^2 + \beta \sum_j \|\mathbf{d}_j\|_1 \quad \text{s.t. } \mathbf{d}_j^T \mathbf{d}_j = 1, \forall j.$$

4. Go back to 2. The iterative minimization process is continued until the stopping criterion is met.
 5. Output \mathbf{D} .
-

The proposed LSD learning algorithm can generate a Laplacian sparse subdictionary \mathbf{D} for each i th object class. We concatenate all \mathbf{D}_i 's into one dictionary $\mathbf{D}_t=[\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c]$, where c is the number of classes. After the dictionary \mathbf{D}_t has been built, the class label of a test sample is computed in the same way as in SRC, where \mathbf{X} in Eq. (1) is replaced by \mathbf{D}_t .

3.3.4 Convergence discussion

The algorithm of LSD is summarized in Algorithm 1. LSD is an alternating optimization problem. At the update step for \mathbf{S} , we solve Eq. (8) by following the feature-sign search algorithm proposed by Lee *et al.* (2006). They proved that the feature-sign search algorithm converges to a global optimum in a finite number of steps by proceeding in a series of feature-sign steps, in which each step reduces the objective function. At the update step for \mathbf{D} , we update the sparse dictionary atom by atom while not violating the sparsity constraint. Executing a series of such steps ensures monotonic reduction. In our case, in each round of alternative minimization, the objective function of LSD decreases. LSD converges within three to six iterations in all the experiments to be presented. Fig. 1 plots the empirical convergence curve of LSD applied to the i-LIDS-MA database,

from which we see that the proposed LSD algorithm converges after about five iterations.

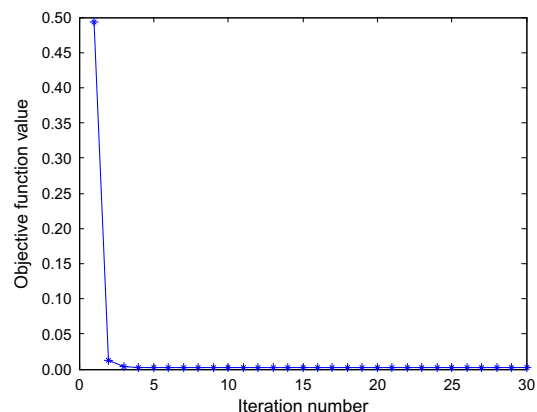


Fig. 1 Example of the convergence of Laplacian sparse dictionary (LSD)

4 Experimental results

The LSD we propose is a general dictionary and can be used in any application. We tested the performance of LSD by embedding it in SRC. After learning LSD, we find the coefficients for a test sample using LSD instead of the whole training set in Eq. (1).

The class label of a test sample is obtained in the same way as in the SRC approach. The residuals given in Eq. (2) are calculated, and the test sample is assigned to the class that yields the minimum error. Three widely used controlled face image databases, Extended Yale B (Georghiades *et al.*, 2001; Lee *et al.*, 2005), ORL, and AR (Martinez and Benavente, 1998), and one uncontrolled person image dataset, i-LIDS-MA (Bağ *et al.*, 2012), were used in the experiments. In our implementation, we applied Eigenfaces (Turk and Pentland, 1991) to reduce the dimensionality of face images. We compared SRC incorporating the proposed LSD with the original SRC (Wright *et al.*, 2009), SRC with MFL (Yang *et al.*, 2010), and SRC with GSC (Zheng *et al.*, 2011). As in the study by Wright *et al.* (2009), the k NN classifier was used in the experiments as a reference.

4.1 Image database and experimental setting

1. Extended Yale B database: The Extended Yale B database consists of images captured from 38 individuals under various laboratory-controlled lighting

conditions. This database has a total of 2414 frontal-face images. Fig. 2 shows some examples. For each subject, 32 images were randomly selected for training and the rest were used for testing. In the sparse representation model, parameter λ (Eq. (1)) was adjusted to achieve the best performance. To ensure a fair comparison, we used the same experimental setting as used by Yang *et al.* (2010). The associated dimension of features was set to 504. In our proposed LSD, we set $\lambda=\beta=0.001$, $\gamma=0.01$, and $k=5$.



Fig. 2 Some examples from the Extended Yale B database

2. ORL database: The ORL database contains a total of 400 images captured from 40 individuals. The images show variation in facial expression and facial details (glasses/no glasses). Fig. 3 shows some examples. In our experiment, the first six images of each individual were used for training and the remaining four for testing. The parameters λ and p were chosen to achieve the best performance for each method and were set the same as in the study of Yang *et al.* (2010). The associated dimension of features was set to 140. In our proposed LSD, we set $\lambda=\beta=0.001$, $\gamma=0.01$, and $k=5$.



Fig. 3 Some examples of the ORL database

3. AR database: The AR database contains frontal images captured from 126 individuals (Martinez and Benavente, 1998). For each individual, 26 pictures were taken with different illumination and expressions in two separate sessions. Fig. 4 shows some examples from the AR database. To ensure a fair comparison, we used the same experimental settings as used by Yang *et al.* (2010). In our experiment, a subset of the database consisting of 50 male and 50 female subjects was used. For each subject, the training set contained seven images from session 1 and the test set contained seven images from session 2. Following Yang *et al.* (2010), parameters λ and p were

selected to achieve the best performance. The associated dimension of features was set to 300. In our proposed LSD, we set $\lambda=\beta=0.001$, $\gamma=0.01$, and $k=5$.



Fig. 4 Some examples from the AR database

4. i-LIDS-MA dataset: The i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset with multiple camera views was originally released by the Home Office in the UK. Images were captured from non-overlapping multiple camera views subject to significant occlusions and large variation in view angle and illumination. i-LIDS-MA images were collected from i-LIDS video surveillance data captured at an airport (Bağ *et al.*, 2012). This database contains multiple images of 40 individuals extracted from two non-overlapping cameras (cameras 1 and 3 in their original setting) and there are large viewpoint changes. For each individual, 46 frames are annotated manually from both cameras. There are $40 \times 2 \times 46 = 3680$ annotated images. Fig. 5 shows some examples from the i-LIDS-MA database. The dataset is very challenging since it was built from data captured in real scenarios without predefined environmental settings. It is an uncontrolled recognition problem unlike the previously described controlled recognition problems.



Fig. 5 Sample images from the i-LIDS-MA dataset
Top and bottom lines correspond to images from different cameras. Columns illustrate the same person

In our experiment, for each person, the training set contained 46 images from camera 1 and the test set contained 46 images from camera 3. Parameters λ and p were selected to achieve the best performance. The associated dimension of features was set to 400. In our proposed LSD, $\lambda=\beta=0.001$, $\gamma=0.01$, and $k=5$.

4.2 Results and analysis

4.2.1 Results

The results from five methods applied to the three face image databases and one person image dataset are shown in Table 1, which lists the maximum recognition rate of each method. Our extensive experimental results show that SRC with the proposed LSD achieves the highest recognition rate and outperforms the original SRC method, SRC with MFL, and SRC with GSC. The classical k NN method performs the worst. The improvement from using LSD was about 6.96% over NN, 2.33% over SRC, 1.8% over MFL, and 0.91% over GSC on the Extended Yale B database. The improvement was about 4.38% over NN, 3.75% over SRC, 2.5% over MFL, and 1.25% over GSC on the ORL database. The improvement was about 6.43% over NN, 3.43% over SRC, 2.71% over MFL, and 0.86% over GSC on the AR database. The improvement was about 5.61% over NN, 3.16% over SRC, 1.86% over MFL, and 0.92% over GSC on the i-LIDS-MA database.

Table 1 The top recognition rates of different methods on test databases

Method	Top recognition rate (%)			
	E-Yale B	ORL	AR	i-LIDS-MA
k NN	92.46%	94.37%	88.14%	53.42%
SRC	97.09%	95.00%	91.14%	55.87%
MFL	97.62%	96.25%	91.86%	57.17%
GSC	98.51%	97.50%	93.71%	58.11%
LSD	99.42%	98.75%	94.57%	59.03%

To prove that the enhancement of recognition accuracy from using LSD was statistically significant, we carried out some statistical analyses. For example, we tested the null hypothesis that GSC and LSD methods yield the same recognition rate on the ORL

database. We compared 10 times top recognition rates of LSD with 10 times top recognition rates of GSC. Results from a Wilcoxon signed rank test are shown in Fig. 6. Since the test rejected the null hypothesis, we can conclude that the proposed LSD method yields higher recognition rates than GSC.

Hypothesis test summary

	Null hypothesis	Test	Sig.	Decision
1	The median of differences between GSC and LSD equals 0	Related-samples Wilcoxon signed rank test	0.038	Reject the null hypothesis

Asymptotic significances are displayed. The significance level is 0.05.

Fig. 6 Results from a Wilcoxon signed rank test

The experimental results and statistical analysis suggest that the proposed LSD algorithm can build a more robust and representative dictionary enabling an improvement in classification accuracy.

4.2.2 Parameter sensitivity analysis

In this section, we present the recognition accuracies with different parameter values. In our LSD, there are four parameters: λ is the tradeoff parameter used to balance the sparsity and the reconstruction error, β is a constant to balance the sparse dictionary basis term and reconstruction error, γ is the regularization parameter, and k is the number of nearest neighbors. λ is the same parameter as that used in SRC and MFL. We selected λ from $\{0.01, 0.001, 0.0001\}$, β from $\{0.001, 0.0005, 0.0001\}$, γ from $\{0.01, 0.005, 0.001, 0.0005, 0.0001\}$, and k from $\{2, 3, 4, 5, 6\}$.

Based on the ORL database, we show the top recognition rates when β varies from 0.001 to 0.0001, γ varies from 0.01 to 0.0001 and $\lambda=0.01$ (Fig. 7), 0.001 (Fig. 8), or 0.0001 (Fig. 9). Fig. 10 shows the top recognition rate when k varies from 2 to 6, λ varies from 0.01 to 0.0001, and $\beta=\gamma=0.001$.

The performance of LSD varied from 97.5% to 99.3% with different λ , β , γ , and k values. The recognition accuracy of LSD was better than that of the other algorithms and stable over a large range of parameter values.

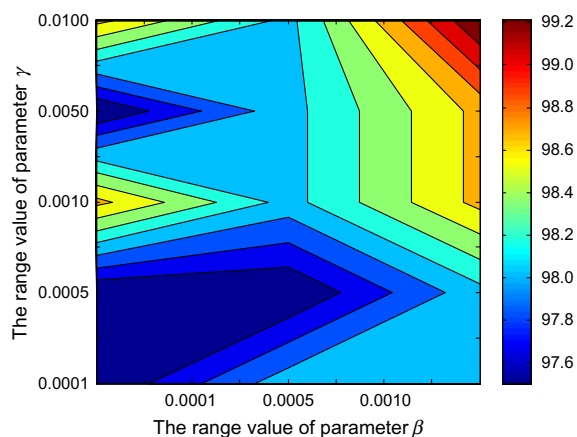


Fig. 7 Recognition accuracy of LSD on the ORL database with different values of parameters β and γ where $\lambda=0.01$

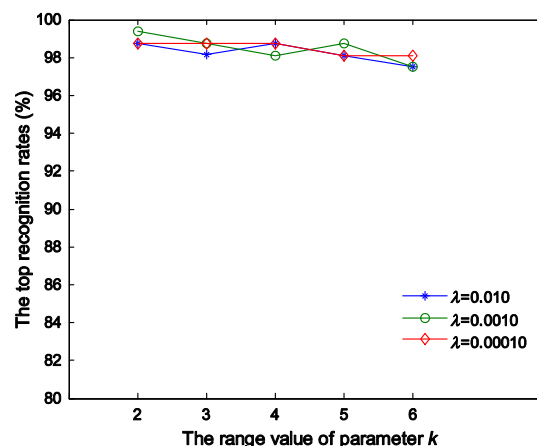


Fig. 10 Recognition accuracy of LSD on the ORL database with different numbers of nearest neighbors k

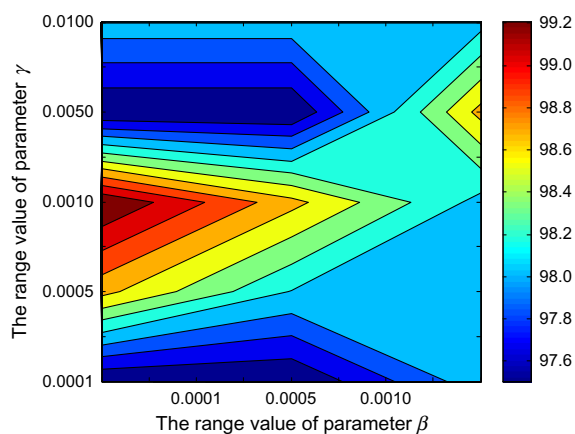


Fig. 8 Recognition accuracy of LSD on the ORL database with different values of parameters β and γ where $\lambda=0.001$

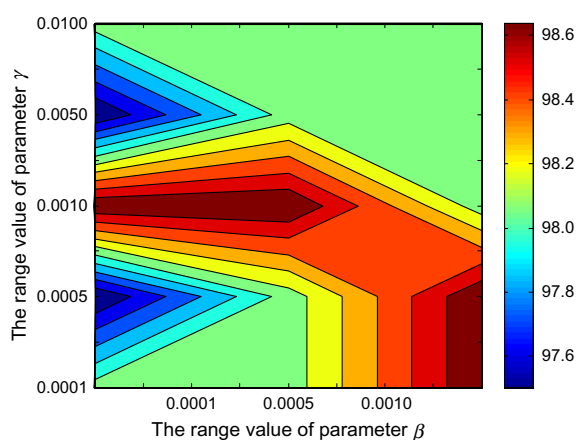


Fig. 9 Recognition accuracy of LSD on the ORL database with different values of parameters β and γ where $\lambda=0.0001$

5 Conclusions

In this paper, we proposed a Laplacian sparse dictionary (LSD) learning method based on manifold learning and double sparsity. The proposed LSD algorithm is a general dictionary learning method in that a Laplacian sparse subdictionary is learned for each class from the samples within that class. The Laplacian sparse dictionary has a sparse structure and can preserve the local structure of the data space. Embedding LSD into sparse representation-based classification (SRC) can improve the performance of SRC-based image classification. Our experiments on the Extended Yale B, ORL, and AR face image databases and the i-LIDS-MA person image dataset demonstrated that the proposed LSD algorithm has high accuracy and stable performance. Comparative experiments using the four benchmark image databases showed that the proposed LSD was superior to the state-of-the-art dictionary methods.

References

- Aharon, M., Elad, M., Bruckstein, A., 2006. *K*-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, **54**(11): 4311-4322. <https://doi.org/10.1109/TSP.2006.881199>
- Bak, S., Corvee, E., Bremond, F., et al., 2012. Boosted human re-identification using Riemannian manifolds. *Image Vis. Comput.*, **30**(6):443-452. <https://doi.org/10.1016/j.imavis.2011.08.008>
- Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Patt. Anal. Mach.*

- Intell.*, **19**(7):711-720.
<https://doi.org/10.1109/34.598228>
- Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *In: Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, p.585-591.
- Chapelle, O., Schölkopf, B., Zien, A., 2006. *Semi-supervised Learning*. MIT Press, Cambridge, MA.
- Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Patt. Anal. Mach. Intell.*, **35**(11):2765-2781.
<https://doi.org/10.1109/TPAMI.2013.57>
- Gangeh, M.J., Ghodsi, A., Kamel, M.S., 2013. Kernelized supervised dictionary learning. *IEEE Trans. Signal Process.*, **61**(19):4753-4767.
<https://doi.org/10.1109/TSP.2013.2274276>
- Gao, S., Tsang, I.W.H., Ma, Y., 2014. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. *IEEE Trans. Image Process.*, **23**(2): 623-634. <https://doi.org/10.1109/TIP.2013.2290593>
- Georghiadis, A.S., Belhumeur, P.N., Kriegman, D.J., 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Patt. Anal. Mach. Intell.*, **23**(6):643-660.
<https://doi.org/10.1109/34.927464>
- He, X., Niyogi, P., 2003. Locality preserving projections. 17th Annual Conf. on Neural Information Processing Systems, p.186-197.
- He, X., Yan, S., Hu, Y., et al., 2005. Face recognition using Laplacian faces. *IEEE Trans. Patt. Anal. Mach. Intell.*, **27**(3):328-340.
<https://doi.org/10.1109/TPAMI.2005.55>
- Huang, M., Yang, W., Jiang, J., et al., 2014. Brain extraction based on locally linear representation-based classification. *NeuroImage*, **92**:322-339.
<https://doi.org/10.1016/j.neuroimage.2014.01.059>
- Lee, H., Battle, A., Raina, R., et al., 2006. Efficient sparse coding algorithms. *In: Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, p.801-808.
- Lee, K.C., Ho, J., Kriegman, D.J., 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Patt. Anal. Mach. Intell.*, **27**(5):684-698.
<https://doi.org/10.1109/TPAMI.2005.92>
- Lu, X., Li, X., 2014. Group sparse reconstruction for image segmentation. *Neurocomputing*, **136**:41-48.
<https://doi.org/10.1016/j.neucom.2014.01.034>
- Lu, X., Wu, H., Yuan, Y., et al., 2013. Manifold regularized sparse NMF for hyperspectral unmixing. *IEEE Trans. Geosci. Remote Sens.*, **51**(5):2815-2826.
<https://doi.org/10.1109/TGRS.2012.2213825>
- Lu, Y., Lai, Z., Fan, Z., et al., 2015. Manifold discriminant regression learning for image classification. *Neurocomputing*, **166**:475-486.
<https://doi.org/10.1016/j.neucom.2015.03.031>
- Martinez, A.M., Benavente, R., 1998. The AR Face Database. CVC Technical Report, No. 24. Centre de Visió per Computador, Universitat Autònoma de Barcelona, Edifici O, Bellaterra, Barcelona.
- Peleg, T., Elad, M., 2014. A statistical prediction model based on sparse representations for single image super-resolution. *IEEE Trans. Image Process.*, **23**(6):2569-2582. <https://doi.org/10.1109/TIP.2014.2305844>
- Qiao, L., Chen, S., Tan, X., 2010. Sparsity preserving projections with applications to face recognition. *Patt. Recogn.*, **43**(1):331-341.
<https://doi.org/10.1016/j.patcog.2009.05.005>
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500): 2323-2326.
<https://doi.org/10.1126/science.290.5500.2323>
- Rubinstein, R., Bruckstein, A.M., Elad, M., 2010a. Dictionaries for sparse representation modeling. *Proc. IEEE*, **98**(6):1045-1057.
<https://doi.org/10.1109/JPROC.2010.2040551>
- Rubinstein, R., Zibulevsky, M., Elad, M., 2010b. Double sparsity: learning sparse dictionaries for sparse signal approximation. *IEEE Trans. Signal Process.*, **58**(3): 1553-1564.
<https://doi.org/10.1109/TSP.2009.2036477>
- Shao, L., Yan, R., Li, X., et al., 2014. From heuristic optimization to dictionary learning: a review and comprehensive comparison of image denoising algorithms. *IEEE Trans. Cybern.*, **44**(7):1001-1013.
<https://doi.org/10.1109/TCYB.2013.2278548>
- Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500):2319-2323.
<https://doi.org/10.1126/science.290.5500.2319>
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)*, **58**(2):267-288.
- Turk, M., Pentland, A., 1991. Eigenfaces for recognition. *J. Cogn. Neurosci.*, **3**(1):71-86.
- Wang, W., Wang, R., Huang, Z., et al., 2015. Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, p.2048-2057.
<https://doi.org/10.1109/TCSVT.2014.2367357>
- Wright, J., Yang, A.Y., Ganesh, A., et al., 2009. Robust face recognition via sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(2):210-227.
<https://doi.org/10.1109/TPAMI.2008.79>
- Yang, A.Y., Zhou, Z., Balasubramanian, A.G., et al., 2013. Fast- l_1 minimization algorithms for robust face recognition. *IEEE Trans. Image Process.*, **22**(8):3234-3246.
<https://doi.org/10.1109/TIP.2013.2262292>
- Yang, J., Zhang, L., Xu, Y., et al., 2012. Beyond sparsity: the role of l_1 -optimizer in pattern classification. *Patt. Recogn.*, **45**(3):1104-1118.
<https://doi.org/10.1016/j.patcog.2011.08.022>

- Yang, J.F., Zhang, Y., 2011. Alternating direction algorithms for ℓ_1 -problems in compressive sensing. *SIAM J. Sci. Comput.*, **33**(1):250-278.
<https://doi.org/10.1137/090777761>
- Yang, M., Zhang, L., Yang, J., et al., 2010. Metaface learning for sparse representation-based face recognition. 17th IEEE Int. Conf. on Image Processing, p.1601-1604.
<https://doi.org/10.1109/ICIP.2010.5652363>
- Yang, M., van Gool, L., Zhang, L., 2013. Sparse variation dictionary learning for face recognition with a single training sample per person. IEEE Int. Conf. on Computer Vision, p.689-696.
<https://doi.org/10.1109/ICCV.2013.91>
- Yang, M., Dai, D., Shen, L., et al., 2014. Latent dictionary learning for sparse representation-based classification. IEEE Conf. on Computer Vision and Pattern Recognition, p.4138-4145.
<https://doi.org/10.1109/CVPR.2014.527>
- Zhang, Z., Xu, Y., Yang, J., et al., 2015. A survey of sparse representation: algorithms and applications. *IEEE Access*, **3**:490-530.
<https://doi.org/10.1109/ACCESS.2015.2430359>
- Zheng, M., Bu, J., Chen, C., et al., 2011. Graph regularized sparse coding for image representation. *IEEE Trans. Image Process.*, **20**(5):1327-1336.
<https://doi.org/10.1109/TIP.2010.2090535>
- Zhu, P., Zuo, W., Zhang, L., et al., 2014. Image set-based collaborative representation for face recognition. *IEEE Trans. Inform. Forens. Secur.*, **9**(7):1120-1132.
<https://doi.org/10.1109/TIFS.2014.2324277>