Dictionary learning under global sparsity constraint

Deyu Meng, Yee Leung*, Qian Zhao, Zongben Xu

Abstract

A new method is proposed in this paper to learn overcomplete dictionary from training data samples. Differing from the current methods that enforce similar sparsity constraint on each of the input samples, the proposed method attempts to impose global sparsity constraint on the entire data set. This enables the proposed method to fittingly assign the atoms of the dictionary to represent various samples and optimally adapt to the complicated structures underlying the entire data set. By virtue of the sparse coding and sparse PCA techniques, a simple algorithm is designed for the implementation of the method. The efficiency and the convergence of the proposed algorithm are also theoretically analyzed. Based on the experimental results implemented on a series of signal and image data sets, it is apparent that our method performs better than the current dictionary learning methods in original dictionary recovering, input data reconstructing, and salient data structure revealing.

Index Terms

Dictionary learning, signal reconstruction, sparse principle component analysis, sparse representation, structure learning.

I. INTRODUCTION

In recent years, there has been a significant interest in using sparse representation over a redundant dictionary as a driving force for various data processing tasks [1]-[5]. All of these applications capitalize

Deyu Meng, Qian Zhao, and Zongben Xu are with the Institute for Information and System Sciences and Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi’an Jiaotong University, Xi’an, 710049, P.R. China. (dymeng@xjtu.edu.cn, zbxu@xjtu.edu.cn)

Yee Leung is with the Department of Geography & Resource Management; Institute of Environment, Energy and Sustainability; and Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, P.R. China (yeeleung@cuhk.edu.hk).

* is the corresponding author of this paper.
on the fact that salient features underlying a data sample (e.g., a signal or an image) can always be captured by its sparse representation over an appropriate dictionary. As such, the pre-specified dictionary is crucial to the success of the sparse representation model in practical applications. Most conventional studies use the “off-the-shelf” dictionaries, such as the wavelet [6] and DCT bases [7], [8], to build a sparsifying dictionary based on a mathematical model of the data. Current studies, however, have demonstrated the advantages of learning an often overcomplete dictionary matched to the data samples of interest [4], [9], [10], [11].

The dictionary learning task is generally described in the following mathematical way: For a collection of data samples \( X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d \times n} \), it is expected to find the dictionary \( D = [d_1, d_2, \cdots, d_m] \in \mathbb{R}^{d \times m} \) (the atom number \( m \) is set larger than \( d \), implying that the dictionary is redundant) through the optimization model [12][13]:

\[
\min_{D,A} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{2} \| x_i - D\alpha_i \|_2^2 + \lambda \mathcal{P}(\alpha_i) \right),
\]

where the vector \( \alpha_i \) contains the representation coefficients of \( x_i \), and we denote the coefficient matrix as \( A = [\alpha_1, \alpha_2, \cdots, \alpha_n] \in \mathbb{R}^{m \times n} \). The objective function of \((P_\lambda)\) involves two elements considered in dictionary learning task: the expression error term, \( \frac{1}{2} \| x_i - D\alpha_i \|_2^2 \), and the sparsity controlling term, \( \mathcal{P}(\alpha_i) \), with respect to each involved sample \( x_i \). The most widely utilized functions of \( \mathcal{P}(\alpha_i) \) include the \( l_0 \) penalty \( \| \alpha_i \|_0 \), the \( l_1 \) penalty \( \| \alpha_i \|_1 \), the MC penalty etc.. The task can also be accomplished by solving the following two alternative optimization problems:

\[
\min_{D,A} \| X - DA \|_F^2 \quad \text{s.t.} \quad \mathcal{P}(\alpha_i) \leq k, \quad \forall \ 1 \leq i \leq n, \quad (P_k)
\]

and

\[
\min_{D,A} \sum_{i=1}^{n} \mathcal{P}(\alpha_i) \quad \text{s.t.} \quad \| x_i - D\alpha_i \|_2 \leq \epsilon, \quad \forall \ 1 \leq i \leq n, \quad (P_\epsilon)
\]

where the notion \( \| A \|_F \) stands for the Frobenius norm, and \( \| A \|_0 \) counts the nonzero entries of the matrix or the vector \( A \). The tuning parameters \( \lambda, k, \) and \( \epsilon \) in the models \((P_\lambda), (P_k), \) and \((P_\epsilon)\) play a very important
role in the model performance. They intrinsically control the compromise between the expression error of the sparse representation and the sparsity of the representation coefficients of the final results.

It should be noted that a uniform parameter $\lambda$, $k$, or $\varepsilon$ formulated for the entire training data set $X$ is specified in the current model $(P_\lambda)$, $(P_k)$, or $(P_\varepsilon)$, respectively. Such formulation facilitates the parameter selection and algorithm construction against the model. The samples collected from applications, however, are always of varying interior structures. On one hand, some samples may be composed of complicated features and need to be very densely represented under the dictionary; while some might be of very simple structure and can be precisely represented with very sparse coefficient vectors. On the other hand, some collected samples may seriously deviate from the original (e.g., signals mixed with the analog-to-digital conversion errors or images transmitted through noisy channels) while some may be totally clean samples. This means that we should vary the parameter $\lambda$, $k$, or $\varepsilon$ with respect to different input samples to make the dictionary learning model adaptive to the intrinsic structures underlying the entire training samples. The uniform specification of the sparsity penalty $\lambda$, the maximal representation sparsity $k$, or the minimal representational error $\varepsilon$ of the conventional model $(P_\lambda)$, $(P_k)$, or $(P_\varepsilon)$, respectively, is thus not very reasonable, and always tends to conduct unstable performance of the model in applications.

The purpose of this paper is to construct a new dictionary learning model that will not impose similar penalty or constraint on each input sample like the conventional methods. Instead, the model will specify a global sparsity constraint on the entire sample set, so that it will adaptively tune the representation sparsity of diverse samples and properly reveal the intrinsic structures underlying the entire data set. An efficient algorithm is specifically designed for the proposed model. It is efficient, easy to be implemented, and is expected to converge to a local optimum of the problem. By a series of experiments, it is verified that the proposed algorithm, in comparison with the current dictionary learning methods, can not only deliver more faithful dictionary underlying the input samples, but can also more precisely recover the original data under the dictionary. Besides, the intrinsic structure underneath the entire training sample set can be very impressively depicted via the different representation sparsities of the samples under the
learned dictionary.

In what follows, related work in the literature is first reviewed in Section II. Details of our algorithm and its basic model are then presented in Section III. The convergence and the computational complexity of the proposed algorithm are also evaluated in this section. The experimental results are given in Section IV for substantiation and verification. The paper is then concluded with a summary and outlook for future research.

II. RELATED WORK

Using sparse representations of data under an appropriately specified redundant dictionary has advanced multiple data processing tasks, and has drawn a lot of attention in the past decade or so. In the conventional studies, the “off-the-shelf” dictionaries have always been employed in various applications. The typical ones include the Fourier [14], the wavelet [15]-[19], and the DCT bases [7], [8]. These bases have been applied to many practical problems, and sometimes tend to conduct nearly-optimal sparse representations of input samples.

Recent research, however, has demonstrated the significance of learning an overcomplete dictionary, instead of a fixed one, matched to the input data of interest. Various algorithms along this line have been formulated in recent years. For example, the algorithm proposed by Olshausen and Field [20] can find sparse linear codes for natural scenes. The dictionary composed by these codes complies with the intrinsic features of the localized, oriented, bandpass receptive fields of the neurons of the primary visual cortex. The method of optimal directions (MOD), proposed by Engan et al. [21], is also an appealing dictionary training algorithm. It improves the efficiency of the work by Olshausen and Field [20] both in the sparse coding and dictionary updating stages. Through generalizing the K-means clustering process to alternate between sparse coding and dictionary updating, Aharon et al. [4], [9] designed K-SVD method. There are two versions of the method: one achieves sparse data representations under strict sparsity constraint (corresponding to the model $(P_k)$, and is thus denoted as $K$-SVD$_{P_k}$) [4], and the other attains the dictionary by allowing a bounded representation error for each training sample (corresponding to the model $(P_\varepsilon)$,
and is thus denoted as $K$-SVD$\lambda$) [9]. The method has empirically shown the state-of-the-art performance in some image processing applications. Some methods have further been constructed to improve the efficiency of the dictionary learning problem: Mairal et al. [22] proposed an online model to efficiently solve the dictionary learning problem; Jenatton et al. [23] used a tree structured sparse representation to give a linear-time computation of the problem; and Lee et al. [12] designed an algorithm to speedup the sparse coding stage of the problem, allowing it to learn larger sparse codes than other algorithms. Recently, some algorithms have also been developed to extend the capability of dictionary learning based on some specific motivations. For instance, Shi et al. [13] developed an algorithm for dictionary learning with non-convex while continuous MC penalty; Mairal et al. [24] established a discriminative approach, instead of the purely reconstructive methods, to build a dictionary. All of the aforementioned methods are addressed to the models $(P_\lambda)$, $(P_k)$ and $(P_\epsilon)$ introduced in Section I.

For many real data processing applications, however, these models cannot fully tally with the practical data samples possessing intrinsic complicated structures, such as those with different content of features, or with spatial and/or spectral non-uniform noises. Along this line of research, Mairal et al. [25] addressed the case of removing nonhomogeneous white Gaussian noise of images by virtue of the a priori knowledge of the noise deviation at each pixel of the objective images. Such elaborate information about noise, however, generally cannot be attained in practice. Very recently, Zhou et al. [10], [11] designed a nonparametric Bayesian method, called the beta process factor analysis (BPFA) method, for dictionary learning. The method can learn a sparse dictionary in situ for input samples with spatially non-uniform noises, without having to know the a priori noise information. The method is thus used as one of the methods for comparison in our experiments.

In this paper, we propose a new dictionary learning method to improve the capability of the current dictionary learning methods by imposing a global sparsity constraint on all training samples to enable optimal atom assignment to individual training samples and adaptation to their intrinsic structures. In what follows, we give a detailed analysis of the proposed model and the associated algorithm.
III. DICTIONARY LEARNING UNDER GLOBAL SPARSITY CONSTRAINT: MODEL AND ALGORITHM

A. Model: From local to global constraint on sparse representation

The current dictionary learning models, namely $(P_\lambda)$, $(P_k)$, and $(P_\epsilon)$, enforce similar sparsity control parameter, including sparsity penalty $\lambda$, sparsity constraint $k$, or representation error bound $\epsilon$, for each involved sample. However, there are often counterexamples to such formulation in real applications, especially for data samples embedded with intrinsic heterogeneous sparsity structures. We take the image case as an instance, in which the training data set corresponds to the small local patches of the image in consideration. On one hand, the local parts of a real image may contain very different capacities of meaningful features, e.g., the region full of patterns with abundant textures, as compared to the area located at the background with small grayscale variations. In such case, smaller sparsity penalty $\lambda$ of $(P_\lambda)$, or larger representation sparsity $k$ of $(P_k)$, should be preset for the local patches located at the former region so that more atoms of the dictionary can be assigned to them. This phenomenon can easily be understood via the representations of patches $A$ and $B$ located at the house eave and the background parts of the “house” image, respectively, as shown in Fig. 1. On the other hand, the real noise mixed in the image is often of significant statistical heteroscedasticity. This means that the extents of noise corruption in various parts of the image, such as patches $C$ and $D$ in Fig. 1, may be significantly different. It is easy to see that patch $C$ is highly corrupted by noise while $D$ is almost clean and contains essentially no noise. Larger representation error bound $\epsilon$ of $(P_\epsilon)$ should then be set for the former patches so that the model can properly capture the variation of noise corruption across the image. It is thus foreseeable that the performance of the current dictionary learning models can be substantially enhanced by adaptively regulating the sparsity control parameter(s) with respect to the underlying structural characteristics of the entire training samples (image patches).

Based on the above rationale, we reformulate the model for dictionary learning into the following
The 'house' image

The 'house' image with nonhomogeneous gaussian noise

Fig. 1. The left figure is the 'house' image, and the right figure is the same image mixed with nonhomogeneous Gaussian noise. A,B,C,D are four patches cut from the two images, respectively. The upper two expressions at the bottom of the figure show the atoms and the coefficients utilized to sparsely represent the patches A and B over the dictionary learned from the algorithm proposed in Section III.B, respectively. The lower two ones demonstrate the ground truth of the noise separated from the patches C and D, respectively.

global-sparsity-constraint form:

$$\min_{D,A} \|X - DA\|^2_F \quad \text{s.t.} \quad \|A\|_0 \leq K, \quad (P_K)$$

where $K$ is the maximal size of the non-zero entries of the coefficient matrix $A$. Differing from the current models in which similar sparsity constraint is imposed on each input sample, our proposed model capitalizes on the global sparsity constraint superimposed upon the entire sample set. This formulation enables the model to adaptively assign different number of atoms, $k_i$, of the dictionary to represent each sample $x_i$ according to its intrinsic structure. This can easily be understood through the following equivalent reformulation of $(P_K)$:

$$\min_{D,A,\{k_i\}_{i=1}^n} \|X - DA\|^2_F \quad \text{s.t.} \quad \|\alpha_i\|_0 \leq k_i, \quad \forall \ 1 \leq i \leq n, \ \sum_{i=1}^n k_i = K. \quad (1)$$

In specific, for input data samples containing different capacities of features, more non-zero atoms will be assigned to represent more complex samples by the proposed model; and for samples corrupted by the heterogeneous noises, the distribution of the non-zero entries of the coefficient matrix tends to be
optimally balanced among the entire sample set, and the sparse representations of the samples over the
dictionary attained by the proposed model will possibly reveal the major information (the original data
samples) while eliminate the minor (the mixed noises) at the global scale. The dictionary learning model
\((P_K)\) is thus expected to outperform the current models.

We now construct an efficient algorithm for solving \((P_K)\).

B. Algorithm: Iterative updating of the column and row vectors of the coefficient matrix

The main idea of our algorithm is to iteratively update the column and row vectors of the coefficient
matrix \(A\) to attain a local optimum of \((P_K)\). Denote the column and row vectors of the coefficient
matrix \(A\) as \([\alpha_c^1, \alpha_c^2, \ldots, \alpha_c^n]\) and \([((\alpha_r^1)^T, (\alpha_r^2)^T, \ldots, (\alpha_r^m)^T)^T\), respectively. The column updating step is
to update each column vector \(\alpha_c^i\) \((i = 1, 2, \ldots, n)\) of \(A\), with the number of its non-zero entries, \(k_c^i\), fixed,
while the column positions of these non-zero elements are optimally relocated in an adaptive way. By
decomposing the model \((P_K)\), the corresponding task is to solve the following optimization problem for
each \(\alpha_c^i\) \((i = 1, \ldots, n)\):

\[
\min_{\alpha_c^i} \|x_i - D\alpha_c^i\|_2^2 \quad \text{s.t.} \quad \|\alpha_c^i\|_0 \leq k_c^i.
\]  \(2\)

The row updating step is to update each row vector \(\alpha_r^i\) \((i = 1, 2, \ldots, m)\) of \(A\), with the sparsity \(k_r^i\) of \(\alpha_r^i\) fixed, while its \(k_r^i\) non-zero elements are optimally adapted to the proper row positions. Since the atom
d\(d_i\) of the dictionary \(D\) corresponds completely to \(\alpha_r^i\) in the sense that \(DA = \sum_{i=1}^{m} d_i (\alpha_r^i)^T\) \([4]\), it is also
simultaneously updated in this step together with \(\alpha_r^i\). The corresponding optimization is of the following
form for each \(\alpha_r^i\) \((i = 1, \ldots, m)\):

\[
\min_{\alpha_r^i, d_i} \|E_i - d_i(\alpha_r^i)^T\|_F^2 \quad \text{s.t.} \quad \|\alpha_r^i\|_0 \leq k_r^i, \quad d_i^T d_i = 1,
\]  \(3\)

where \(E_i = X - \sum_{j \neq i} d_j (\alpha_j)^T\), stands for the representation error of all considered samples with the effect
of the \(i\)-th atom \(d_i\) removed.

The algorithm for solving the optimization model \((P_K)\) can then be summarized as Algorithm 1.
Algorithm 1 The Algorithm for dictionary learning under global sparsity constraint (GDL)

Given: $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d \times n}$, the global sparsity $K$

Execute:

1. Initialize the dictionary $D \in \mathbb{R}^{d \times m}$ and the coefficient matrix $A \in \mathbb{R}^{m \times n}$ with sparsity $K$, respectively.

2. Repeat

   2.1 (Column updating). Update the column vector $\alpha^c_i$ of $A$ by solving (2) for each $i = 1, \cdots, n$.

   2.2 (Row updating). Update the row vector $\alpha^r_i$ of $A$ and the atom $d_i$ of $D$ by solving (3) for each $i = 1, \cdots, m$.

Until the termination condition is satisfied

Return: the solution $D, A$ of $(P_K)$.

Now the question is how to efficiently solve the optimization problems (2) and (3). For (2), it is actually the well known $L_0$-norm model of sparse coding, and multiple effective algorithms have been investigated against this model. The typical ones, for example, include the thresholding methods, e.g., the hard algorithm [26], and the greedy methods, e.g., the OMP algorithm [27].

For (3), we give the following theorem:

**Theorem 1:** The optimum of the model (3) can be attained by solving the optimization problem:

$$\max_w w^T E_i^T E_i w \quad \text{s.t. } w^T w = 1, \|w\|_0 \leq \alpha^r_i,$$

in the sense of

$$\hat{d}_i = \frac{E_i \hat{w}}{\|E_i \hat{w}\|_2}, \hat{\alpha}^r_i = \|E_i \hat{w}\|_2 \hat{w},$$

where $\hat{d}_i, \hat{\alpha}^r_i$ are the optima of (3), and $\hat{w}$ is the optimum of (4), respectively.

The proof of Theorem 1 is given in the appendix.

It is very interesting that (4) is just the sparse principal component analysis or sparse PCA model [28], which has been thoroughly investigated in the last decade [28]-[32]. Many efficient algorithms have been constructed for solving the model, including SPCA [28], GPower [29], sPCA-rSVD [32] etc..

Thus, both models (2) and (3) can be efficiently solved, i.e., both of the column and row updating steps of the proposed algorithm can be effectively implemented, by employing the existing methods in sparse
coding and sparse PCA, respectively.

The remaining issues are then how to appropriately specify the initial dictionary $D$ and the coefficient matrix $A$ in step 1, and when to terminate the iterative process in step 2 of the proposed algorithm. In our experiments, $D$ and $A$ were simply initialized with data signals and sparse matrix with $K$ non-zero elements, whose positions are randomly generated in the matrix, respectively. As for the termination of the algorithm, since the entire representation error of signals, i.e., the objective function of $(P_K)$, decrease monotonically under the fixed global sparsity constraint throughout the iterative process (see details in the next section), the algorithm can reasonably be terminated when the decrease in value of the objective function is smaller than some preset small threshold, or the process has reached the pre-specified number of iterations.

Now we investigate the convergence and the computational complexity of the proposed algorithm in the next section.

C. Convergence and computational complexity

We first discuss the convergence of the proposed algorithm, i.e., the iterations between the column and row updating steps of the algorithm. Under the assumption that models (3) and (4) can be precisely solved, each of the updating iterations in step 2 monotonically decreases the total representation error $\|DA - X\|_F^2$ under the guarantee that the constraint $\|A\|_0 \leq K$ is consistently held. Therefore, a local minimum of (2) is expected to be attained by the proposed algorithm. Although the above claim depends on the success of the sparse coding and sparse PCA techniques employed in approximating the solutions of (3) and (4), respectively, the algorithms employed on the tasks performed well in our experiments and could empirically generate a rational solution of (5) after multiple iterations, as demonstrated in Section IV.

The computational complexity of the proposed algorithm is essentially determined by the iterative process between the column and row updating steps. By employing the recent sparse coding and sparse PCA technologies, e.g., OMP [27] and sPCA-rSVD algorithms [32], respectively, in our experiments, both
steps can be efficiently performed, requiring around $O(dnm\hat{k}) \times Iter$ computational cost, where $\hat{k}$ is the maximal value of $k_c^i$s, and $Iter$ is the iteration number of the algorithm\footnote{In each iteration of the proposed algorithm, the optimization problem needs to be solved for $i = 1, \cdots, n$, each requiring $O(dmk_c^i)$ cost from utilizing the OMP algorithm [9], [27], and problem needs to be solved for $i = 1, \cdots, m$, each costing $O(dn)$ computation from employing the sPCA-rSVD algorithm [32]. Thus, the total computational complexity of the proposed algorithm is around $O(dnm\hat{k}) \times Iter$.}. That is, the computational time of the proposed algorithm increases linearly with the dimensionality and the size of the input samples, as well as the number of atoms in the dictionary. The computational complexity of the proposed algorithm is comparable to that of the current dictionary learning algorithms [9], [10], [20].

IV. EXPERIMENTAL RESULTS

To test the effectiveness of the proposed algorithm on dictionary learning, it was applied to a series of synthetic signals and real images for substantiation. The results are summarized in the following discussion. All programs were implemented on the Matlab 7.0 platform.

A. Synthetic signal experiments

We first apply the proposed algorithm to synthetic signal data to test whether the algorithm can recover the generating dictionary and reconstruct the original signals.

Two series of signal experiments were implemented, each involving 11 sets of signals. Each signal set contained 1500 20-dimensional signals, denoted as $X = [x_1, x_2, \cdots, x_{1500}] \in \mathbb{R}^{20 \times 1500}$, which were created by a linear combination of a pre-specified dictionary $D = [d_1, d_2, \cdots, d_{50}] \in \mathbb{R}^{20 \times 50}$ and representation coefficients $A = [\alpha_1, \alpha_2, \cdots, \alpha_{1500}] \in \mathbb{R}^{50 \times 1500}$, and mixed with different extents of homogeneous Gaussian white noise. The entries of each dictionary $D$ were first generated by random sampling, and each column (atom) $d_i$ of $D$ was then divided by its $l_2$-norm for normalization. For the first series of experiments, each column $\alpha_i$ of $A$ contained 3 non-zero elements with randomly chosen values and locations. For the second series of experiments, each coefficient matrix $A$ consisted of 4500 randomly valued and located non-zero entries. Added to the 11 signal sets in both series of experiments were
homogeneous Gaussian noises with standard deviations varying from 0 to 0.1 with interval 0.01. The SNR values of these signal sets ranged from infinity to around 10, correspondingly. It should be indicated that for the first experimental series, the signals in each experiment were of similar representation sparsity over the preset dictionary, which complies with the preassumption of the model \( P_k \); and for both series of experiments, the signals were corrupted with homogeneous noises, which tallies with the preassumption underlying the model \( P_e \).

Four of the current dictionary learning methods, including K-SVD \( P_k \) [4], K-SVD \( P_e \) [9], MOD [21], and BPFA [10], were applied to these signal sets for comparison. The dictionary of the first three methods, as well as the proposed method, were initialized as the randomly selected signals from the input set, and the initialization of the last method was based on the singular value decomposition technique [10]. Since both MOD and K-SVD \( P_e \) need the a priori deviation of the noise mixed in the signals to preset the representation error parameter, we directly used the ground truth information to optimally specify the parameter value [9]. For K-SVD \( P_k \), the sparsity constraint parameter \( k \) was set as the real sparsity in the first experimental series. For the second series, it was specified by running the method 5 times on each signal set under different \( k \)s, and selecting the best one as the final output. All parameters involved in the BPFA were automatically inferred by using a full posterior on the model [10]. For the proposed GDL method, the global sparsity \( K \) was set as 4500 for all experiments. The results of all involved methods were attained after 100 iterations.

Two criteria are utilized to assess the performance of the employed methods for dictionary learning. The first is computed by sweeping through each atom of the generating dictionary and checking whether it is recovered by the dictionary attained by a utilized method via the following formula [4]:

\[
1 - |d_i^T \hat{d}_i|,
\]

where \( d_i \) is the atom of the original dictionary and \( \hat{d}_i \) is its corresponding closest element in the recovered

\(^2\)The signal set mixed with Gaussian noise with deviation 0 means that the set is clean and contains no noise. The corresponding value of SNR is infinity.
dictionary. If the value of (6) was less than 0.01, then it was considered as a success. The rate of the successfully recovered atoms in the generating dictionary (called the dictionary recovery rate, or DR in brief) is then taken as the first criterion, which evaluates the capability of the method in delivering the original dictionary beneath the input signals. The second criterion is the mean of the standard deviations of the reconstructed signals from the original signals (called the representation error, or RE briefly). This value assesses the performance of the method in recovering the input signals.

In all of the implemented experiments, the DR and RE values in the iterative processes of the four current methods and the proposed method were recorded. The upper and lower panels of Fig. 2 depict the RE and DR curves of the five methods in the iterative processes of three of the first series of experiments, respectively. Fig. 3 shows the corresponding results in three cases of the second series of experiments. Panels (a) and (c) of Fig. 4 show the final RE and DR values of the five methods in 11 experiments of the first series, respectively. For easy comparison, panels (b) and (d) of the figure display the mean values of RE and DR of the five methods as vertical bars. Fig. 5 depicts the cases of the second series of experiments.

It can easily be observed from the upper panels of Figs. 2 and 3 that the RE values obtained by the proposed method tend to decrease monotonically throughout the iterative process. Besides, after around 10 iterations, the GDL algorithm attains the smallest (5 out of 6 depicted cases) or the second smallest RE values among the five methods. In the final output, the GDL algorithm also achieves the smallest (7 out of 11 experiments in the first series and 8 out of 11 experiments in the second series) or the second smallest RE values in all of the experiments, as clearly shown in Fig. 4(a) and Fig. 5(a). On the average, the proposed algorithm outperforms the other four methods in both series of experiments, as shown in Fig. 4(b) and Fig. 5(b). This demonstrates the excellent capability of the proposed algorithm in reconstructing the input signals.

Furthermore, from the lower panels of Figs. 2 and 3, it can be observed that the DR curves of the proposed method tend to increase monotonically in all experiments, and the method obtains the largest (4
Fig. 2. The upper panels: The RE curves of the K-SVD$_{P_k}$, K-SVD$_{P_e}$, MOD, BPFA, and the proposed GDL methods in the iterative processes in three of the first series of experiments. The lower panels: the corresponding DR curves. It should be noted that the BPFA fails to take effect on the signals in the sense that the method always obtains RE values larger than 0.2 and zero DR values.

Fig. 3. The upper panels: The RE curves of the five methods in the iterative processes in three of the second series of experiments. The lower panels: the corresponding DR curves. It should be noted that BPFA fails to take effect on the signals in the sense that the method always obtains RE values larger than 0.3 and zero DR values.
out of 6 depicted cases) or the second largest values among the five methods after 10 iterations. Moreover, by observing Fig. 4(c) and Fig. 5(c), the proposed algorithm apparently yields the most stable DR values among the five methods, and successfully detects more than 90% atoms of the original dictionary in each of the experiments. In specific, the algorithm attains the largest (6 out of 11 experiments in the first series and 9 out of 11 experiments in the second series) or second largest DR values in all experiments, and achieves the largest average DR values in both experimental series, as shown in Fig. 4(d) and Fig. 5(d). This substantiates the good capability of the GDL algorithm in recovering the original dictionary.

In the next section, we further verify the effectiveness of the proposed method on image reconstruction
from data with more complex intrinsic structures and more complicated nonhomogeneous noises.

B. Real image experiments

A series of test images of $512 \times 512$ or $256 \times 256$ pixels were utilized for the image reconstruction problems. These images were generated by combining 6 gray-scale images, all of which are widely used in the image processing literature [33], with different levels of nonhomogeneous noise. In our experiments, four types of nonhomogeneous noise were employed: (1) nonhomogeneous Gaussian noise with extent $\delta$: the standard deviation of the Gaussian noise increasing uniformly from 0 for lower-right pixels to $\delta$ for the upper-left pixels across the image; (2) salt-pepper noise with extent $p$: corrupting the image with $p$ percentage of dead pixels with either maximum or minimum intensity values; (3) mixture of homogeneous Gaussian and salt-pepper noise, with extents $\sigma$ and $p$, respectively; (4) mixture of nonhomogeneous Gaussian and salt-pepper noise, with extents $\delta$ and $p$, respectively.

The dictionary was trained on the overlapping patches, of $8 \times 8$ pixels (i.e., the input samples are with dimension $d = 64$), from a noisy image in each experiment, and thus each experiment included $n = (256 - 7)^2 = 62,001$ patches (all available patches from the $256 \times 256$ images, and every second patch from every second row in the $512 \times 512$ images). In each of the experiments, the dictionary $D$ contained $m = 256$ atoms. For each dictionary learning method, the images were rebuilt by averaging the overlapping reconstructed patches over the dictionary attained by the method.

Three of the current methods, DCT [7], K-SVD$_{P_{\epsilon}}$ [9], and BPFA [11]$^3$, were also applied to these images for comparison. The dictionary utilized by the DCT method is the overcomplete DCT bases, while dictionaries of K-SVD$_{P_{\epsilon}}$, BPFA and GDL, were trained from the images. The randomly selected image patches were used as the initialization of the K-SVD$_{P_{\epsilon}}$ and GDL methods, and the singular-value-decomposition-based initialization was used for BPFA. The K-SVD$_{P_{\epsilon}}$, BPFA, GDL results were all obtained by 10 iterations. Since both DCT and K-SVD$_{P_{\epsilon}}$ need to preset the parameter that evaluates

$^3$The results of the K-SVD$_{P_{\epsilon}}$ and MOD methods are omitted since their performance are comparatively less than the other adopted methods in the images.
the mean noise deviation of the entire image pixels, we implemented both methods 10 times on each test image under different initializations of this parameter and only recorded the best one as the final result. In comparison, the proposed GDL algorithm was implemented only once under $K = 15000$ for all experiments. The results are quantitatively measured by the PSNR value (with unit dB) defined by

$$\text{PSNR}(f^r) = 10 \log_{10} \frac{255^2}{\|f - f^r\|_2^2},$$

where $f$ and $f^r$ denote the original and the reconstructed images, respectively. Besides, by looking through the numbers of atoms required for representing the image patches (i.e., the numbers of the non-zero elements in the corresponding representation coefficients) and averaging the results over the entire image, an atom-using-frequency figure, of the same resolution as the original image, can be achieved. For the atom-using-frequency figures so constructed, the brighter is a pixel, the more atoms are assigned to represent the image patches containing the pixel, and thus the more emphasis is placed on the region around the pixel by the corresponding method, and vice versa. Therefore, such an atom-using-frequency figure can qualitatively reflect the intrinsic image structure explored by the utilized method.

Table I lists details of the PSNR results of the reconstructed images. Figs. 6-9 depict the recovered images, along with their PSNR values obtained by applying the DCT, K-SVD, BPFA, and GDL methods to four typical test images mixed with different types of noises, respectively. The corresponding dictionaries and atom-using-frequency figures attained by these methods are also displayed in these figures for easy evaluation.

The advantages of the proposed method can easily be observed in these results. First, our algorithm best rebuilds the original images among all employed methods. Specifically, as compared with the DCT, K-SVD and BPFA methods, our method achieves the largest PSNR values in almost all of the experiments, and on average, our obtained PSNR value are the largest among those of the employed methods for each image experiment. This advantage can also be visualized in the first rows of Figs. 6-9. The noise is most prominently removed from the noisy images, e.g. the bookshelf of the Barbara image in Fig. 6 by our method, and the details of the original image are mostly recovered, e.g. the windows of the House image...
| σ | Lena | Barbara | Boat | Fingerprint | House | Peppers |
|---|------|---------|------|-------------|-------|---------|
| 51.10 | 27.44 | 27.72 | 24.57 | 25.15 | 25.72 | 26.06 | 23.15 | 23.56 | 27.49 | 28.03 | 25.88 | 26.67 |
|      | 24.34 | 29.93 | 25.39 | 25.67 | 25.08 | 27.44 | 25.17 | 25.86 | 25.19 | 31.67 | 25.10 | 27.66 |
| 61.32 | 26.45 | 26.62 | 23.48 | 23.97 | 24.68 | 24.85 | 22.02 | 22.66 | 26.15 | 26.47 | 24.62 | 25.37 |
|      | 22.69 | 28.88 | 23.81 | 25.41 | 23.29 | 26.83 | 23.49 | 25.17 | 23.44 | 30.61 | 23.54 | 27.32 |
| 71.54 | 25.71 | 25.75 | 22.67 | 22.71 | 23.90 | 24.04 | 20.77 | 21.41 | 25.20 | 25.46 | 23.97 | 24.54 |
|      | 21.12 | 28.04 | 22.28 | 24.82 | 21.74 | 26.11 | 22.13 | 24.44 | 22.23 | 29.58 | 22.01 | 26.75 |
| 81.76 | 25.44 | 25.41 | 22.41 | 22.34 | 23.46 | 23.39 | 19.99 | 20.55 | 24.70 | 24.85 | 23.11 | 23.60 |
|      | 20.07 | 27.08 | 21.23 | 24.22 | 20.55 | 25.48 | 20.84 | 23.55 | 21.16 | 28.59 | 20.88 | 26.01 |
| 91.98 | 25.20 | 25.19 | 22.31 | 22.25 | 23.21 | 23.15 | 19.15 | 19.68 | 24.07 | 24.03 | 22.43 | 22.74 |
|      | 19.18 | 26.30 | 20.19 | 23.49 | 19.44 | 24.82 | 19.77 | 22.62 | 19.93 | 26.91 | 19.82 | 25.35 |
| 102.20 | 25.06 | 25.06 | 22.17 | 22.16 | 23.07 | 23.05 | 18.39 | 18.83 | 23.97 | 24.00 | 21.80 | 22.05 |
|      | 18.12 | 25.47 | 19.39 | 22.90 | 18.71 | 24.15 | 18.83 | 21.72 | 19.02 | 26.60 | 19.06 | 24.75 |

TABLE I

Summary of the PSNR results of the image experiments. In each cell, the upper-left, upper-right, lower-left, and lower-right values refer to the DCT, K-SVD, BPFA, and GDL methods, respectively. The best result in each cell is highlighted. The last row is the average results of the four methods over all noise cases for each image.
Fig. 6. Results on the Barbara image mixed with nonhomogeneous Gaussian noise with extent 61.32. The panels from top to bottom: the reconstructed images, the dictionaries, and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA, and GDL methods, respectively.

Fig. 7. Results on the Peppers image mixed with salt-pepper noise with extent $p = 10$. The panels from top to bottom: the reconstructed images, the dictionaries, and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA, and GDL methods, respectively.
Fig. 8. Results on the Boat image corrupted by mixture of homogeneous Gaussian noise and salt-pepper noise with extents $\sigma = 20$ and $p = 5$. The panels from top to bottom: the reconstructed images, the dictionaries, and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA, and GDL methods, respectively.

Fig. 9. Results on the House image corrupted by mixture of nonhomogeneous Gaussian noise and salt-pepper noise with extents $\delta = 51.10$ and $p = 4$. The panels from top to bottom: the reconstructed images, the dictionaries, and the atom-using-frequency figures obtained by the DCT, K-SVD, BPFA, and GDL methods, respectively.
in Fig. 9 (the details can better be seen by zooming in onto the images in the computer). These results show the excellent capability of the proposed algorithm in reconstructing the original images.

Second, the proposed method more robustly attains the proper dictionary underlying the images as compared with the other dictionary learning methods. This can easily be observed in the second rows of Figs. 6-9. It is evident that the dictionaries attained by our method capture more meaningful features underlying the images and are least affected by the noise in all cases. These results validate the capability of the proposed algorithm in properly generating the dictionary for images corrupted by nonhomogeneous noises.

Third, the atom-using-frequency figures obtained by our method faithfully reflect the intrinsic structures underlying the images. Particularly, from the third rows of Figs. 6-9, it can be observed that the atom-using-frequency figures obtained by the proposed method clearly depict the basic edge information underlying the images. This is due to the fact that the related patches are of relatively complicated structures, and our method can adaptively assign more atoms to represent these image patches. In comparison, such meaningful structures are not so noticeably detected by the atom-using-frequency figures of the other methods in the experiments. These results demonstrate the capability of the GDL method in detecting major/meaningful structure information while at the same time eliminating the minor/useless noise information underlying the images at the global scale.

V. Conclusion

In this paper we have proposed a novel dictionary learning method. Instead of enforcing similar sparsity constraint on each input sample like the current methods, the new method imposes global sparsity constraint on all training data samples, which makes the new method capable of optimally assigning atoms for representing the various samples and adapting to the intrinsic data structures at the global scale. An efficient algorithm has also been correspondingly developed, which is easy to be implemented based on the sparse coding and sparse PCA techniques, and is expected to converge to a locally optimal solution of the problem. Based on the experimental results on a series of signal and image data sets, it has been
substantiated that as compared with the current dictionary learning methods, the proposed method can more faithfully deliver the ordinary dictionary and properly reconstruct the input samples. Besides, it has been theoretically analyzed and empirically verified that by utilizing the proposed method, the atoms of the dictionary can be optimally adapted to represent samples with various intrinsic complexities, and the frequency of atom-using can facilitate revealing the intrinsic structure underlying the input samples.

Multiple problems still need to be further investigated in our future research. For example, the effectiveness of the proposed algorithm requires to be further testified in real signals and images with complicated noise types, e.g., the poisson noise. Besides, qualitatively speaking, the more complex is the entire structure or the less noise is contained in the input data, the larger the global sparsity parameter $K$ should be properly preset. Further investigation, however, still needs to be made to design an automatic quantitative parameter selection strategy to further improve the quality of the proposed method. Furthermore, research is needed to further improve the efficiency of the proposed algorithm by virtue of the online [22] or convexification [12] techniques.

APPENDIX: PROOF OF THEOREM 1

Theorem 1: The optimum of the model

$$\begin{align*}
\min_{\alpha_i^r, d_i} \| E_i - d_i (\alpha_i^r)^T \|_F^2 \\
\text{s.t. } \| \alpha_i^r \|_0 \leq k_i^r, \quad d_i^T d_i = 1,
\end{align*}$$

(3)
can be attained by solving the optimization problem

$$\begin{align*}
\max_{w} w^T E_i^T E_i w \\
\text{s.t. } w^T w = 1, \quad \| w \|_0 \leq k_i^r,
\end{align*}$$

(4)
in the sense of

$$\tilde{d}_i = \frac{E_i \hat{w}}{\| E_i \hat{w} \|_2}, \quad \tilde{\alpha}_i^r = \| E_i \hat{w} \|_2 \hat{w},$$

(5)
where $\tilde{d}_i$, $\tilde{\alpha}_i^r$ are the optima of (3), and $\hat{w}$ is the optimum of (4), respectively.
Proof:

(i) For the optimal solutions $\hat{d}_i$, $\hat{\alpha}_i^r$ of (3), it holds that

$$\hat{d}_i = \frac{E_i \hat{\alpha}_i^r}{\|E_i \hat{\alpha}_i^r\|_2}.$$  

Proof of (i): The objective of (3) can be equivalently transformed as:

$$\|E_i - d_i(\alpha_i^r)^T\|^2_F = \text{tr}((E_i - d_i(\alpha_i^r)^T)(E_i - d_i(\alpha_i^r)^T)^T)$$

$$= \text{tr}(E_iE_i^T - E_i\alpha_i^r(d_i)^T - d_i(\alpha_i^r)^TE_i^T + d_i(\alpha_i^r)^T\alpha_i^r(d_i)^T)$$

$$= \text{tr}(E_iE_i^T) + (\alpha_i^r)^T\alpha_i^r - 2(E_i\alpha_i^r)^Td_i.$$  

By fixing $\alpha_i^r$ as the optimal solution $\hat{\alpha}_i^r$ of (3), it is easy to deduce that the optimal $d_i$ of (3) can be obtained by solving the reformulated optimization problem:

$$\max_{d_i} (E_i \hat{\alpha}_i^r)^T d_i \quad \text{s.t.} \quad d_i^T d_i = 1. \quad (7)$$

That is, the optimal $\hat{d}_i$ of (7) should be the unit vector with minimal inner product with the vector $E_i \hat{\alpha}_i^r$.

It is then easy to attain that $\hat{d}_i$ should be parallel to the direction of $E_i \hat{\alpha}_i^r$, i.e.,

$$\hat{d}_i = \frac{E_i \hat{\alpha}_i^r}{\|E_i \hat{\alpha}_i^r\|_2}.$$  

The proof of (i) is then completed.

(ii) It holds that

$$\|\hat{\alpha}_i^r\|^2_2 = \|E_i \hat{\alpha}_i^r\|_2.$$  

Proof of (ii): Denote $\hat{w} = \frac{\hat{\alpha}_i^r}{\|\hat{\alpha}_i^r\|_2}$ and $\hat{\rho} = \|\hat{\alpha}_i^r\|_2$. It is then easy to deduce that $\hat{\rho}$ is the optimum of the following optimization problem:

$$\min_{\rho} \|E_i - \rho \left(\hat{d}_i \hat{w}^T\right)\|^2_F.$$
since for any value of $\rho$,

$$
\left\| E_i - \rho \left( \hat{d}_i \hat{w}^T \right) \right\|_F^2 = \left\| E_i - \hat{d}_i (\hat{\alpha}_i^r)^T \right\|_F^2 \leq \left\| E_i - \left( \hat{d}_i \hat{\beta}^T \right) \right\|_F^2 = \left\| E_i - \left( \hat{d}_i (\rho \hat{w})^T \right) \right\|_F^2
$$

where $\beta = \rho \hat{w}$ is also a vector with $k_i^\nu$ non-zero entries. Since

$$
F(\rho) = \left\| E_i - \rho \left( \hat{d}_i \hat{w}^T \right) \right\|_F^2 = \text{tr} \left( \left( E_i - \rho \left( \hat{d}_i \hat{w}^T \right) \right) \left( E_i - \rho \left( \hat{d}_i \hat{w}^T \right) \right)^T \right)
$$

$$
= \text{tr} \left( E_i E_i^T - 2 \rho E_i \hat{w} \hat{d}_i^T + \rho^2 \hat{d}_i \hat{w}^T \hat{d}_i \hat{w}^T \right)
$$

$$
= \text{tr} \left( E_i E_i^T - 2 \rho \hat{d}_i^T E_i \hat{w} + \rho^2 \right),
$$

based on (i), it is easy to deduce that

$$
F'(\rho) = 0 \Rightarrow \rho = \hat{d}_i^T E_i \hat{w} \\
\Rightarrow \rho = \left( \frac{E_i \hat{\alpha}_i^r}{\|E_i \hat{\alpha}_i^r\|_2} \right)^T \frac{E_i \hat{\alpha}_i^r}{\|\hat{\alpha}_i^r\|_2} \\
\Rightarrow \rho = \frac{\|E_i \hat{\alpha}_i^r\|_2}{\|\hat{\alpha}_i^r\|_2} \\
\Rightarrow \|\hat{\alpha}_i^r\|_2 = \frac{\|E_i \hat{\alpha}_i^r\|_2}{\|\hat{\alpha}_i^r\|_2}.
$$

The proof of (ii) is completed.

(iii) The model (3) can be solved through the optimization problem (4) in the sense of the transformations (5).

Proof of (iii): Denote the vector $\hat{w}$ as the normalized $\hat{\alpha}_i^r$, i.e.,

$$
\hat{w} = \frac{\hat{\alpha}_i^r}{\|\hat{\alpha}_i^r\|_2}.
$$

Based on (i) and (ii), we have

$$
\hat{d}_i = \frac{E_i \hat{\alpha}_i^r}{\|E_i \hat{\alpha}_i^r\|_2} = \frac{E_i \hat{w}}{\|E_i \hat{w}\|_2},
$$

and

$$
\hat{\alpha}_i^r = \|\hat{\alpha}_i^r\|_2 \hat{w} = \frac{\|E_i \hat{\alpha}_i^r\|_2}{\|\hat{\alpha}_i^r\|_2} \hat{w} = \|E_i \hat{w}\|_2 \hat{w}.
$$

The minimum of the objective function of (3) can then be equivalently transformed as

$$
\left\| E_i - \hat{d}_i (\hat{\alpha}_i^r)^T \right\|_F^2 = \left\| E_i - E_i \hat{w} \hat{w}^T \right\|_F^2.
$$
We further transform (10) as follows:

\[ \| E_i - E_i \hat{w} \hat{w}^T \|^2_F = \text{tr}((E_i - E_i \hat{w} \hat{w}^T)(E_i - E_i \hat{w} \hat{w}^T)^T) \]

\[ = \text{tr}(E_i E_i^T) - \text{tr}(E_i \hat{w} \hat{w}^T E_i^T) \]

\[ = \| E_i \|^2_F - \hat{w}^T E_i^T E_i \hat{w}. \]

Thus, the optimum of (3) can be attained through the optimal solution of the following optimization problem

\[ \min_w \| E_i \|^2_F - w^T E_i^T E_i w \quad \text{s.t. } w^T w = 1, \| w \|_0 \leq k_i^r \]

in the sense of the transformations (8) and (9). Since \( \| E_i \| \) is a fixed constant for optimization, we can equivalently reformulate the above optimization problem as (4).

This thus completes the proof of the theorem.

\[ \square \]

REFERENCES

[1] X. Mei and H.B. Ling. “Robust visual tracking and vehicle classification via sparse representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 11, pp. 2259-2272, Nov. 2011.

[2] T.V. Pham and A.W.M. Smeulders, “Sparse representation for coarse and fine object recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 4, pp. 555-567, Apr. 2006.

[3] J.K. Pillai, V.M. Patel, R. Chellappa, and N. K. Ratha, “Secure and robust iris recognition using random projections and sparse representations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 9, pp. 1877-1893, Sep. 2011.

[4] M. Aharon, M. Elad, and A.M. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[5] 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.

[6] S. Mallat, “A wavelet tour of signal processing,” Academic Press, 2009.

[7] O.G. Guleryuz, “Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part I-Theory,” IEEE Trans. Image Process., vol. 15, no. 3, pp. 539-554, Mar. 2005.

[8] O.G. Guleryuz, “Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part II-Adaptive algorithms,” IEEE Trans. Image Process., vol. 15, no. 3, pp. 555-571, Mar. 2005.
[9] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[10] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, “Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images,” to be published in *IEEE Trans. Image Process.*, 2011.

[11] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, “Non-parametric bayesian dictionary learning for sparse image representations,” In in *Adv. Neural Inf. Process. Syst.*, vol. 22, pp. 2295-2303, 2009.

[12] H. Lee, A. Battle, R. Raina, and A.Y. Ng, “Efficient sparse coding algorithms,” in *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 801-808, 2007.

[13] J.P. Shi, X. Ren, G. Dai, J.D. Wang, and Z.H. Zhang, “A non-convex relaxation approach to sparse dictionary learning,” In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2010.

[14] G.P. Tolstov, “Fourier series,” Courier-Dover, 1976.

[15] E.J. Candes and D.L. Donoho, “Recovering edges in ill-posed inverse problems: Optimality of curvelet frames,” *Ann. Statist.*, vol. 30, no. 3, pp. 784-842, Jun. 2002.

[16] M.N. Do and M. Vetterli, “Contourlets, beyond wavelets,” New York: Academic, 2003.

[17] D.L. Donoho, “Wedgelets: Nearly minimax estimation of edges,” *Ann. Statist.*, vol. 27, no. 3, pp. 859-897, Jun. 1998.

[18] S. Mallat and E. LePennec, “Bandelet image approximation and compression,” to be published in *SIAM J. Multiscale Model. Simul.*, 2005.

[19] W.T. Freeman and E.H. Adelson, “The design and use of steerable filters,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891-906, Sep. 1991.

[20] B.A. Olshausen and D.J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?” *Vision Res.*, vol. 37, no. 23, pp. 3311-3325, Dec. 1998.

[21] K. Engan, S.O. Aase, and J.H. Husøy, “Multi-frame compression: Theory and design,” *EURASIP Signal Process.*, vol. 80, no. 10, pp. 2121-2140, Oct. 2000.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” In *Proc. IEEE Int. Conf. Mach. Learn.*, vol. 26, pp. 689-696, 2009.

[23] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, “Proximal methods for sparse hierarchical dictionary learning,” In *Proc. IEEE Int. Conf. Mach. Learn.*, 2010.

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. “Supervised dictionary learning,” in *Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 1033-1040, 2009.

[25] J. Mairal, M. Elad, and G. Sapiro, “Sparse representation for color image restoration,” *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53-69, Jan. 2008.

[26] K. Bredies and D.A. Lorenz, “Iterated hard shrinkage for minimization problems with sparsity constraints,” *SIAM J. Sci. Comput.*, vol. 30, np. 2, pp. 657-683, 2008.
[27] Y.C. Pati, R. Rezaifar, and P.S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” In Proc. Ann. Asilomar Conf. Signals Systems and Computers, pp. 40-44, 1993.

[28] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” J. Comput. Graph. Stat., vol. 15, no. 2, pp. 265-286, Jun. 2006.

[29] M. Journéee, Y. Nesterov, P. Richtárik, and R. Sepulchre, “Generalized power method for sparse principal component analysis,” J. Mach. Learn. Res., vol. 11, pp. 517-553, Feb. 2010.

[30] D.Y. Meng, Q. Zhao, and Z.B. Xu, “Improve robustness of sparse PCA by L1-norm maximization,” Pattern Recognit., vol. 45, no. 1, pp. 487-497, Jan. 2012.

[31] C.D. Sigg and J.M. Buhmann, “Expectation maximization for sparse and non-negative PCA,” in Proc. IEEE Int. Conf. Mach. Learn., vol. 25, pp. 960-967, 2008.

[32] H.P. Shen and J.Z. Huang. “Sparse principal component analysis via regularized low rank matrix approximation,” J. Multivar. Anal., vol. 99, no. 6, pp. 1015-1034, Jul. 2008.

[33] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli, “Image denoising using scale mixtures of gaussians in the wavelet domain,” IEEE Trans. Image Process., vol. 12, no. 11, pp. 1338-1351, Nov. 2003.