Sparsity-based Image Denoising via Dictionary Learning and Structural Clustering

Weisheng Dong
Xidian University

Xin Li
WVU

Lei Zhang
HK Polytech. Univ.

Guangming Shi
Xidian University

Abstract

Where does the sparsity in image signals come from? Local and nonlocal image models have supplied complementary views toward the regularity in natural images - the former attempts to construct or learn a dictionary of basis functions that promotes the sparsity; while the latter connects the sparsity with the self-similarity of the image source by clustering. In this paper, we present a variational framework for unifying the above two views and propose a new denoising algorithm built upon clustering-based sparse representation (CSR). Inspired by the success of l_1-optimization, we have formulated a double-header l_1-optimization problem where the regularization involves both dictionary learning and structural structuring. A surrogate-function based iterative shrinkage solution has been developed to solve the double-header l_1-optimization problem and a probabilistic interpretation of CSR model is also included. Our experimental results have shown convincing improvements over state-of-the-art denoising technique BM3D on the class of regular texture images. The PSNR performance of CSR denoising is at least comparable and often superior to other competing schemes including BM3D on a collection of 12 generic natural images.

1. Introduction

There have been two complementary views toward the regularization of image denoising problems: local vs. nonlocal. In the local view, a signal $\vec{x} \in R^n$ can be decomposed with respect to a collection of n-dimensional basis vectors in the Hilbert space (also-called dictionary) $\mathbf{\Phi} \in R^{n \times m}$, namely $\vec{x}_{n\times 1} = \mathbf{\Phi}_{n\times m}\vec{\alpha}_{m\times 1}$, where $\vec{\alpha}$ denotes the vector of weights. The sparsity of α can be characterized by its l_0-norm (nonconvex) or computationally more tractable l_1 norm [1]. This line of research has led to both construction of basis functions (e.g., ridgelet, contourlets) and adaptive learning of dictionary (e.g., K-SVD [2], stochastic approximation [3]). In the nonlcoal view, natural images contain self-repeating patterns. Exploiting the self-similarity of overlapping patches has led to a flurry of nonlocal image denoising algorithms - e.g., nonlocal mean [4], BM3D [5], locally learned dictionaries K-LLD [6], learned simultaneous sparse coding (LSSC) [7].

Among them, the PSNR performance of BM3D has remained the state-of-the-art since its publication. Despite the impressive performance of BM3D, there still lacks a solid understanding about *why* it performs so well. Moreover, the subtle relationship between sparsity (widely used for low-level vision tasks) and clustering (a common tool for the middle-level vision) remains elusive; we do acknowledge the most recent effort on joint/group sparsity [7] which attempts to shed some light on this issue. It seems desirable to connect the two class of most promising ideas, namely dictionary learning (e.g., K-SVD) and structural clustering (e.g., BM3D), under a *unified* theoretic framework.

In this paper, we achieve the above objective by proposing a new image model called clustering-based sparse representation (CSR). The basic idea behind our CSR model is to treat the local and nonlocal sparsity constraints (associated with dictionary learning and structural clustering respectively) as peers and incorporate them into a unified variational framework. The new regularization term can be viewed as a plausible formalization of joint/group sparsity discussed in [7]. Thanks to the unitary property of dictionary, we can show the equivalence between spatial-domain and transform-domain representation of this new term. Additionally, inspired by the success of compressed sensing, we propose to replace the l_2-norm of characterizing nonlocal sparsity with an l_1-norm, which forms a *double-header* l_1 optimization problem.

We have developed an iterative shrinkage solution to the above double-header l_1 optimization problem vis surrogate functions [8]. Our results further generalize those in [8] - from a single regularization parameter to a pair of regularization parameters. Such generalization allows us to simultaneously enforce local and nonlocal sparsity constraints by computationally efficient shrinkage operators. Additionally, we have borrowed ideas from reweighted l_1-optimization [9] to adaptively adjust the two regularization parameters and iterative regularization [10] to further improve the performance of CSR denoising. Extensive experimental results are reported that our CSR algorithm can achieve highly competitive (and often better) performance to other leading denoising techniques including the state-of-the-art BM3D.

2. Clustering-based Sparse Representation (CSR) Model

Following the notation used in [2], we first establish the connection between an image $\mathbf{X}$ and the set of sparse coefficients $\alpha = \{\vec{\alpha}_i\}$ (so-called sparseland model). Let $\mathbf{x}_i$ denote a patch extracted from $\mathbf{X}$ at the spatial location i; then we have

$$\mathbf{x}_i = \mathbf{R}_i \mathbf{X}, \tag{1}$$

where $\mathbf{R}_i$ denotes a rectangular windowing operator. Note that when overlapping is allowed, such patch-based representation is highly redundant and the recovery of $\mathbf{X}$ from $\{\mathbf{x}_i\}$ becomes an over-determined system. It is straightforward to obtain the following Least-Square solution

$$\mathbf{X} = (\sum_i \mathbf{R}_i^T \mathbf{R}_i)^{-1} (\sum_i \mathbf{R}_i^T \mathbf{x}_i), \tag{2}$$

which is nothing but an abstraction of the strategy of averaging overlapped patches. Meantime, for a given dictionary $\mathbf{\Phi}$, each patch is related to its sparse coefficients $\{\vec{\alpha}_i\}$ by

$$\mathbf{x}_i = \mathbf{\Phi} \vec{\alpha}_i. \tag{3}$$

Substituting Eq. (3) into Eq. (2), we obtain

$$\mathbf{X} = \mathbf{D}\vec{\alpha} \doteq (\sum_i \mathbf{R}_i^T \mathbf{R}_i)^{-1} (\sum_i \mathbf{R}_i^T \mathbf{\Phi} \vec{\alpha}_i), \tag{4}$$

where $\mathbf{D}$ is the operator dual to $\mathbf{R}$ (reconstructing image from sparse coefficients). Under the context of image denoising, one can formulate the following variational problem

$$\vec{\alpha} = \arg\min_{\vec{\alpha}} \frac{1}{2} ||\mathbf{Y} - \mathbf{D}\vec{\alpha}||_2^2 + \lambda ||\vec{\alpha}||_1, \tag{5}$$

where $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ is the noisy image and λ is the standard Lagrangian multiplier. Extensive studies have been done on the design/learning of dictionary [2] and computationally efficient/robust algorithms for solving the above convex optimization problem [11].

The key motivation lies in the observation that the sparse (nonzero) coefficients $\vec{\alpha}$ are NOT randomly distributed (please refer to Fig. 1 for a concrete example). Their location uncertainty is often related to the nonlocal self-similarity of image signals, which implies the possibility of achieving higher sparsity by exploiting such *location*-related constraint. From such perspective of resolving both intensity and location uncertainty, one might even make the connection with the idea called bilateral filtering originally proposed in [12]. Clustering represents a plausible approach of exploiting such nonlinear (since it is location-related) constraint; and indeed there are plenty of tools (e.g., kmeans, kNN, spectral, graph-cut) in the literature. However, it is often difficult to establish a synergic connection between data clustering and sparse representations partially because they are viewed as tools developed at different levels (middle vs. low). To gain deeper insight into how nonlocal self-similarity can promote the sparsity, we propose to study the following cost function

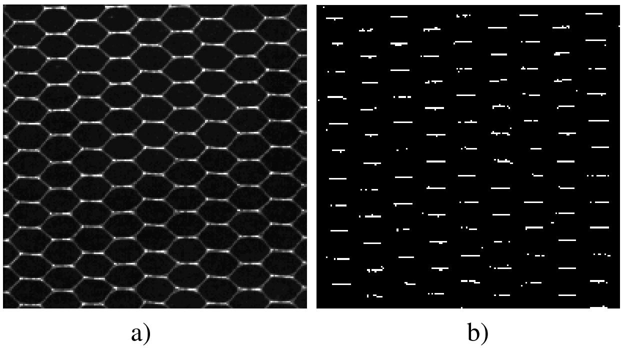

Figure 1. Limitation of K-SVD: a) an image of regular texture; b) spatial distribution of sparse coefficients corresponding to the 6th basis vector (note that their locations are NOT random).

$$\begin{aligned}(\vec{\alpha}, \vec{\mu}) &= \arg\min_{\vec{\alpha}, \vec{\mu}_k} \frac{1}{2} ||\mathbf{Y} - \mathbf{D}\vec{\alpha}||_2^2 + \lambda_1 ||\vec{\alpha}||_1 \\ &+ \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} ||\mathbf{\Phi}\vec{\alpha}_i - \vec{\mu}_k||_2^2. \end{aligned} \tag{6}$$

where $\vec{\mu}_k$ stands for the centroid of the k-th cluster C_k of coefficients $\vec{\alpha}$. An intuitive interpretation of the new clustering-based regularization term is that the weighting coefficients $\vec{\alpha}$ are re-encoded with respect to $\vec{\mu}_k$. With such further "compression", sparser representation can be obtained (the consequence of exploiting nonlocal self-similarity). Indeed previous works such as BM3D and LSSC are based on similar considerations about clustering and sparsity but their connection remains loose. To the best of our knowledge, this is the first rigorous mathematical formulation of combining clustering and sparsity under a unified variational framework.

To better understand the significance of the new regularization term, we can rewrite Eq. (6)

$$\begin{aligned}(\vec{\alpha}, \vec{\beta}) &= \arg\min_{\vec{\alpha}, \vec{\mu}_k} \frac{1}{2} ||\mathbf{Y} - \mathbf{D}\vec{\alpha}||_2^2 + \lambda_1 ||\vec{\alpha}||_1 \\ &+ \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} ||\mathbf{\Phi}\vec{\alpha}_i - \mathbf{\Phi}\vec{\beta}_k||_2^2. \end{aligned} \tag{7}$$

where $\vec{\mu}_k = \mathbf{\Phi}\vec{\beta}_k$ (i.e., all centroid vectors are represented with respect to the same dictionary $\mathbf{\Phi}$ as $\mathbf{x}_i$). Thanks to the unitary property of $\mathbf{\Phi}$, we have $||\mathbf{\Phi}\vec{\alpha}_i - \mathbf{\Phi}\vec{\beta}_k||_2^2 = ||\vec{\alpha}_i -$

$\vec{\beta}_k||_2^2$. Therefore, Eq. (6) boils down to the following joint optimization problem

$$
\begin{aligned}
(\vec{\alpha}, \vec{\beta}) &= \arg\min_{\vec{\alpha}, \vec{\mu}_k} \frac{1}{2}||\mathbf{Y} - \mathbf{D}\vec{\alpha}||_2^2 + \lambda_1||\vec{\alpha}||_1 \\
&+ \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} ||\vec{\alpha}_i - \vec{\beta}_k||_2^2. \qquad (8)
\end{aligned}
$$

Inspired by the success of compressed sensing (called l1magic by the authors of [1]), we propose to replace the L_2 norm in the new regularization term by L_1 norm.

$$
\begin{aligned}
(\vec{\alpha}, \vec{\beta}) &= \arg\min_{\vec{\alpha}, \vec{\mu}_k} \frac{1}{2}||\mathbf{Y} - \mathbf{D}\vec{\alpha}||_2^2 + \lambda_1||\vec{\alpha}||_1 \\
&+ \lambda_2 \sum_{k=1}^{K} \sum_{i \in C_k} ||\vec{\alpha}_i - \vec{\beta}_k||_{\mathbf{1}}. \qquad (9)
\end{aligned}
$$

To summarize the CSR model, we note that it offers a new way of understanding sparsity by unifying dictionary learning ($\vec{\alpha}_i$'s) and structural clustering $\vec{\beta}_k$'s into a variational framework. Higher sparsity is expected to be achieved by exploiting the structural redundancy in $\vec{\alpha}_i$'s. Another way of understanding $\vec{\beta}_k$'s is that they are exemplars learned through structural clustering to encode $\vec{\alpha}_i$'s at a higher level (conceptually similar to the idea of deconvolutional networks [13]).

3. Iterative Reweighted and Regularized l_1-Minimization

One of the major technical contributions of this paper is to solve the *double-header* l_1-optimization of Eq. (9) via an iterative algorithm alternatively updating $\vec{\alpha}$ and $\vec{\beta}$. Borrowing ideas from surrogate functions [8], we have derived an iterative shrinkage operator to update $\vec{\alpha}$ for fixed $\vec{\beta}$, i.e.,

$$
\alpha_j^{(i+1)} = \begin{cases} \mathbf{S}_{\tau_1,\tau_2}(v_j^{(i)}) & \beta_j \geq 0 \\ -\mathbf{S}_{\tau_1,\tau_2}(-v_j^{(i)}) & \beta_j < 0 \end{cases} \qquad (10)
$$

where

$$
\mathbf{v}^{(i)} = \frac{1}{c}\mathbf{D}^T(\mathbf{x} - \mathbf{D}\alpha^{(i)}) + \alpha^{(i)}. \qquad (11)
$$

and $\tau_1 = \frac{\lambda_1}{c}, \tau_2 = \frac{\lambda_2}{c}$ (c is an auxiliary parameter guaranteeing the convexity of surrogate function), superscript (i) denotes iteration number and subscript j denotes the j-th entry in a vector. Therefore, our result shows iterative shrinkage is also applicable to the case of two regularization parameters corresponding to local and nonlocal sparsity respectively, which further extends the result of [8] (**D** from unitary to non-unitary). Technical details of deriving the new bi-variate shrinkage operator $\mathbf{S}_{\tau_1,\tau_2}$ are referred to the Appendix. The update of $\vec{\beta}$ follows a similar procedure to nonlocal mean denoising [4] (iterative reweighted Least-Square [14] could offer a more systematic solution but has not been used in our current implementation).

Computational efficiency of iterative shrinkage allows us to refine the CSR model and its associated optimization algorithm. First, we have borrowed ideas from the literature of variational image restoration [15] and reweighted l_1-optimization [9] to adaptively adjust the two regularization parameters τ_1, τ_2. In [15], it was shown that the regularization parameter λ should be inversely proportional to signal-to-noise-ratio (SNR); the reweighting strategy proposed in [9] also suggests that the new weights are inversely proportional to signal magnitude $|\mathbf{x}|_1$ in the scenario of compressed sensing (since no noise is involved). Therefore, we have adopted the following strategy for updating τ_1, τ_2

$$
\tau_1 = c_1\frac{\sigma_w^2}{\sigma_\alpha}, \tau_2 = c_2\frac{\sigma_w^2}{\sigma_\gamma}. \qquad (12)
$$

where σ_w^2 is noise variance, $\vec{\gamma} = \vec{\alpha} - \vec{\beta}$ and c_1, c_2 are two predefined constants (we usually set $c_1 < c_2$ to emphasize the nonlocal term).

Second, inspired by the recent work [10], we propose to update the estimation of recovered image by

$$
\mathbf{X}^{(i+1)} = \tilde{\mathbf{S}}((1-\delta)\mathbf{X}^{(i)} + \delta\mathbf{Y}), \qquad (13)
$$

where $\tilde{\mathbf{S}} = \mathbf{D} \circ \mathbf{S} \circ \mathbf{R}$ denotes the projection onto the regularization constraint set and

$$
(1-\delta)\mathbf{X}^{(i)} + \delta\mathbf{Y} = \mathbf{X}^{(i)} + \delta(\mathbf{Y} - \mathbf{X}^{(i)}), \qquad (14)
$$

is the operator implementing the idea of iterative regularization. Note that the RHS of Eq. (14) can be viewed as a degenerated Landweber operator (when the blurring kernel reduces to an identity operator) and δ is a small positive number controlling the amount of noise fed back to the iteration. We have chosen to manually terminate the algorithm after three iterations. A complete description of the proposed CSR denoising algorithm is as follows.

Algorithm 1. Image Denoising via CSR

```
• Initialization: $\hat{\mathbf{X}} = \mathbf{Y}$;
• Outer loop (dictionary learning): for $i = 1, 2, ..., I$
    - update $\Phi$ via kmeans and PCA;
    • Inner loop (structural clustering): for $j = 1, 2, ..., J$
        - iterative regularization: $\tilde{\mathbf{X}} = \hat{\mathbf{X}} + \delta(\mathbf{Y} - \hat{\mathbf{X}})$;
        - regularization parameter update: obtain new estimate of $\tau_1, \tau_2$ via Eq. (12);
        - centroid estimate update: obtain new estimate of $\vec{\beta}_k$'s via kNN clustering;
        - image estimate update: obtain new estimate of $\mathbf{X}$ by $\hat{\mathbf{X}} = \mathbf{D} \circ \mathbf{S} \circ \mathbf{R}\tilde{\mathbf{X}}$;
```

4. Bayesian Interpretation and Extension of CSR Denoising

In this section, we provide a Bayesian interpretation of the above CSR denoising algorithm. In the literature of wavelet thresholding [16], the connection between sparse representation and Bayesian denoising has been well established. Such connection has been fruitful to the development of both theories in the past decade because it helps reconcile the differences between deterministic and probabilistic schools. The dual role played by regularization function and prior distribution in deterministic and probabilistic settings has coherently shown the equivalence between variational and Bayesian image restoration. Therefore, we deem it useful to extend the above connection from a local (dictionary-based) to nonlocal (clustering-based) framework.

The basic idea behind CSR is to assume that we can treat the centroids of K clusters $\vec{\beta}$ as the peer hidden variables to sparse coefficients $\vec{\alpha}$. Such idea is essentially to recognize the importance of resolving the organizational (location-related) uncertainty underlying image signals. Therefore, we might formulate the following maximum a posterior (MAP) estimation problem

$$(\vec{\alpha},\vec{\beta}) = \arg\max_{\vec{\alpha},\vec{\beta}} logP(\vec{\alpha},\vec{\beta}|\mathbf{Y}), \tag{15}$$

Using Bayesian formula, we can rewrite Eq. (15) into

$$(\vec{\alpha},\vec{\beta}) = \arg\max_{\vec{\alpha},\vec{\beta}} logP(\mathbf{Y}|\vec{\alpha},\vec{\beta}) + P(\vec{\alpha},\vec{\beta}), \tag{16}$$

where the two terms correspond to the likelihood and prior distributions respectively. The first term is easy to characterize by the degradation model $\mathbf{Y} = \mathbf{X} + \mathbf{W}$, namely

$$P(\mathbf{Y}|\vec{\alpha},\vec{\beta}) = \frac{1}{\sqrt{2\pi}\sigma_w} exp(-\frac{1}{2\sigma_w^2}||\mathbf{Y}-\mathbf{D}\vec{\alpha}||_2^2). \tag{17}$$

The art of statistical modeling often refers to the approximation of the second term - e.g., under the assumption of i.i.d., we can decompose $P(\vec{\alpha})$ into the product of the marginal distributions. One way of relaxing such assumption is to further exploit its structural constraint by data clustering - i.e.,

$$P(\vec{\alpha},\vec{\beta}) = P(\vec{\beta}|\vec{\alpha})P(\vec{\alpha}) = P(\vec{\gamma}|\vec{\alpha})P(\vec{\alpha}), \tag{18}$$

where $\vec{\gamma} = \vec{\alpha} - \vec{\beta}$ defines the deviation from each cluster. Such clustering-based differential prediction can be viewed as another level of sparse coding strategy so $\vec{\gamma}$ is approximately independent from $\vec{\alpha}$. If we choose to model both $\vec{\alpha}$ and $\vec{\gamma}$ by i.i.d. Laplacian distribution, the prior model is given by

$$\begin{aligned} P(\vec{\alpha},\vec{\beta}) &= \prod_i \frac{1}{\sqrt{2}\sigma_\alpha} exp(-\frac{||\vec{\alpha}_i||_1}{\sigma_\alpha}) \times \\ & \prod_k \prod_i \frac{1}{\sqrt{2}\sigma_\gamma} exp(-\frac{||\vec{\alpha}_i - \vec{\beta}_k||_1}{\sigma_\gamma}). \end{aligned} \tag{19}$$

Substituting Eqs. (17) and (19) into Eq. (16), we obtain

$$\begin{aligned} (\vec{\alpha},\vec{\beta}) &= \arg\min_{\vec{\alpha},\vec{\beta}} ||\mathbf{Y}-\mathbf{D}\vec{\alpha}||_2^2 + \frac{2\sqrt{2}\sigma_w^2}{\sigma_\alpha}\sum_i ||\vec{\alpha}_i||_1 \\ &+ \frac{2\sqrt{2}\sigma_w^2}{\sigma_\gamma}\sum_k\sum_i ||\vec{\alpha}_i - \vec{\beta}_k||_1. \end{aligned} \tag{20}$$

which is equivalent to Eq. (6) by setting $\lambda_1 = \frac{2\sqrt{2}\sigma_w^2}{\sigma_\alpha}$, $\lambda_2 = \frac{2\sqrt{2}\sigma_w^2}{\sigma_\gamma}$.

Probabilistic setting also allows us to reinspect some adhoc choice made in deterministic setting. For example, inspired by the kernel density estimation techniques (e.g., Parzen windows [17]) in nonparametric statistics, we can generalize Eq. (1) into

$$\mathbf{W}\mathbf{x}_i = \mathbf{W}\mathbf{R}_i\mathbf{X}, \tag{21}$$

where $\mathbf{W}$ denotes a nonuniform weighting operator in favor of samples closer to the center of the window. Accordingly, we can extend the formula of Eq. (3) into a weighted Least-Square solution

$$\mathbf{X} = (\sum_i \mathbf{R}_i^T\mathbf{W}\mathbf{R}_i)^{-1}(\sum_i \mathbf{R}_i^T\mathbf{W}\mathbf{x}_i). \tag{22}$$

In our current implementation, a Gaussian window is used for $\mathbf{W}$ (similar to the weighted window used in nonlocal mean [4]).

5. Image Denoising Experiments

We have implemented the proposed CSR denoising algorithm under MATLAB (source codes accompanying this work can be accessed at `http://www.csee.wvu.edu/~xinl/CSR.html`). The following parameters have been used in our experiments: block-size $B = 7$, $\lambda = 0.03$, dictionary-size $K = 64$ and $I = J = 3$. In a nutshell, our CSR algorithm can be viewed as the hybrid of dictionary learning (similar to K-SVD but with 64 dictionaries) and structural clustering (similar to BM3D but in the transform domain). In CSR, the updating of dictionary is implemented by kmeans and PCA which attempts to better handle the spatially varying characteristics than K-SVD; the clustering is performed with respect to transform coefficients as a second-stage sparse coding (while clustering and filtering are disconnected in BM3D).

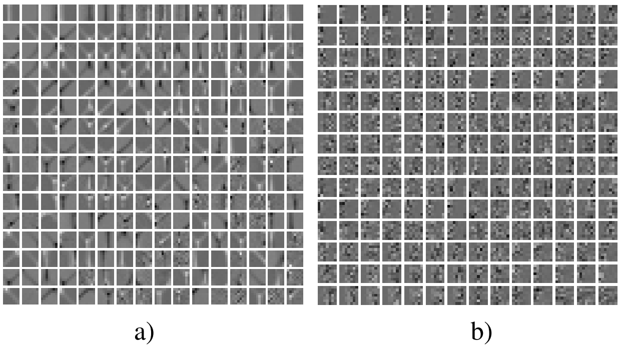
Figure 2. Comparison of learned dictionaries from the test image $D34$ between a) K-SVD ($K = 256$) and b) CSR (only four out of 64 sets of dictionaries is displayed).

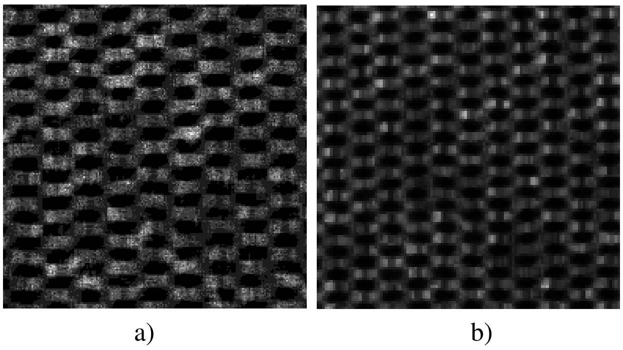
Figure 3. Comparison of sparsity distribution between K-SVD and CSR: a) spatial distribution of $\vec{\alpha}$ plotted on a block-level ($B = 8$); b) spatial distribution of $\vec{\gamma} = \vec{\alpha} - \vec{\beta}$ plotted on a block-level ($B = 7$); note that how the introduction of $\vec{\beta}$ (cluster centroids) makes the CSR sparser.

To understand how the idea of structural clustering improves sparsity, we have compared the outputs of CSR and K-SVD on the same noisy image as shown in Fig. 1. Figs. 2 and 3 include the comparison between learned dictionaries and sparsity distributions. For this specific image, we can observe that the basis images learned by K-SVD and CSR are visually similar. However, the actual sparsity (measured on a block-by-block basis) varies as the consequence of structural clustering. It can be seen from Fig. 3 that CSR appears to be sparser (i.e., reencode $\vec{\alpha}$ into $\vec{\gamma} = \vec{\alpha} - \vec{\beta}$) due to the exploitation of nonlocal similarity.

It is not surprising to see that CSR significantly outperforms both K-SVD and BM3D on such image of regular texture pattern. The PSNR gain over K-SVD and BM3D is over $0.77dB$ and $1.97dB$ respectively as shown in Fig. 4). Apparently when image is highly self-repeating, dictionary learning plays a more important role than structural clustering (as verified by the gain of K-SVD over BM3D); but combining them together leads to further impressive improvement. Dramatic gain has also been observed for other images of regular texture patterns in the Brodatz database (not included here due to space limit).

We have also compared the CSR algorithm and other leading denoising techniques in the literature at different noise levels for a collection of 12 photographic images. The denoising results of three benchmark schemes (K-SVD [2], SA-DCT [18] and BM3D [5]) are all based on the source codes or executables released by the original authors. Table I includes the PSNR comparison on the set of 12 images (the highest PSNR values among fout are highlighted in each cell). We conclude that the proposed CSR algorithm has achieved highly competitive PSNR performance to BM3D; on the average CSR has outperformed BM3D by a small margin. To the best of our knowledge, this is the first time that under fair comparison situations[1], denoising results comparable to BM3D are reported in the open literature. Subjective quality comparisons for two typical test images (one abundant with textures and the other with edges) are shown in Figs. 5 and 6. The PSNR gain of CSR over BM3D on these two images is less impressive than for $D34$ but still in the range of $0.3 - 0.4dB$.

6. Discussions: Think Globally, Fit Locally?

What have we learned from the new theory of CSR and the above denoising experiments? It is enlightening to understand the relationship between dictionary learning and structural clustering from a manifold perspective. Globally the collection of patches in natural images would form a nonlinear manifold consisting of many constellations; how to discover the local geometry of such nonlinear manifold is a problem that has attracted lots of attention in recent years[2]. Image denoising can also be cast under the framework of manifold learning/reconstruction except that unsupervised learning works with noisy data. Dictionary learning such as K-SVD separates image signals from additive noise by thinking globally (i.e., change-of-coordinates); while structural clustering such as BM3D achieves the same objective by locally fitting the hypersurface in the patch space (i.e., iterative shrinkage). What CSR has shown is the benefit of combining global thinking with local fitting.

Acknowledgements

This work was partially supported by grant NSF-CCF-0914353, HK RGC General Research Fund (PolyU 5375/09E), NSFC (No. 60736043,61072104,

[1] The authors of LSSC [7] have reported slightly higher PSNR results than BM3D but their experiments have used a large number of additional training data for dictionary learning.

[2] Note that the term local or global refers to the view toward image manifold in the patch space (i.e., nonlocal image processing in fact corresponds to fitting the manifold locally - an unfortunate inconsistency of term used by different communities).

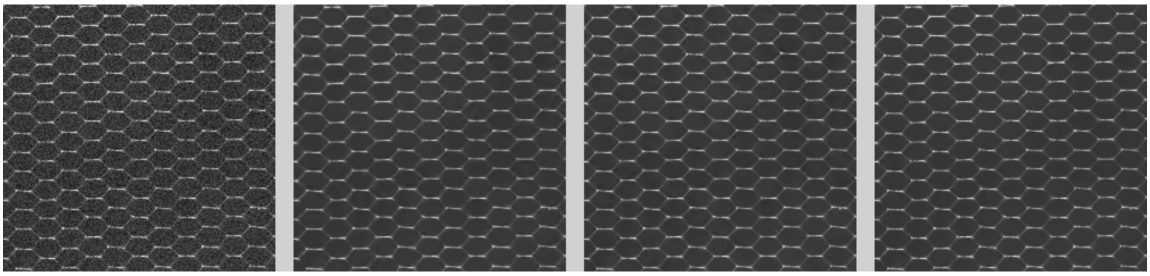

Figure 4. Denoising performance comparison for $D34$ image: a) noisy ($\sigma_w = 20$); b) BM3D ($PSNR = 29.33dB, SSIM = 0.9178$); c) K-SVD ($PSNR = 30.53dB, SSIM = 0.9327$); d) CSR ($PSNR = 31.30dB, SSIM = 0.9426$).

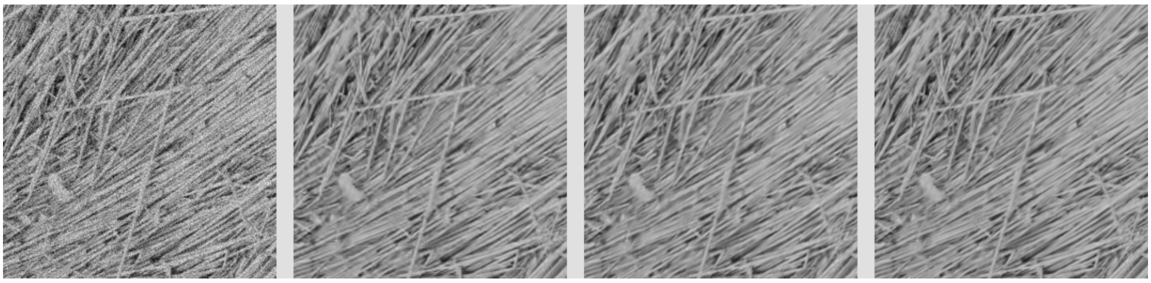

Figure 5. Denoising performance comparison for $straw$ image: a) noisy ($\sigma_w = 20$); b) BM3D ($PSNR = 27.09dB, SSIM = 0.8963$); c) K-SVD ($PSNR = 26.95dB, SSIM = 0.8899$); d) CSR ($PSNR = 27.50dB, SSIM = 0.9061$).

61070138,and 61071170), and the Fundamental Research Funds of the Central Universities of China (No. K50510020003).

Appendix: Iterative Shrinkage via Surrogate Functions

To simplify the notation, we will write α, β directly instead of their vectorial forms. The classical l_1-optimization problem is written as

$$\alpha = \arg\min_{\alpha} \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda||\alpha||_1, \qquad (23)$$

The simplest case to solve Eq. (23) is when $\mathbf{D}$ is unitary. With the assumption of $\mathbf{D}\mathbf{D}^T = \mathbf{I}$, the objective function becomes

$$\begin{aligned} f(\alpha) &= \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda||\alpha||_1 \\ &= \frac{1}{2}||\mathbf{D}(\mathbf{D}^T\mathbf{x} - \alpha)||_2^2 + \lambda||\alpha||_1 \\ &= \frac{1}{2}||\mathbf{D}(\alpha_0 - \alpha)||_2^2 + \lambda||\alpha||_1 \\ &= \frac{1}{2}||\alpha_0 - \alpha||_2^2 + \lambda||\alpha||_1. \qquad (24) \end{aligned}$$

where $\alpha_0 = \mathbf{D}^T\mathbf{x}$ and we have used $||\mathbf{x}||_2^2 = ||\mathbf{D}\mathbf{x}||_2^2$. Note that the consequence of the above procedure is "diagonalization" of the objective function - i.e.,

$$f(\alpha) = \sum_i [\frac{1}{2}(\alpha_0(i) - \alpha(i))^2 + \lambda|\alpha(i)|], \qquad (25)$$

which simplifies Eq. (24) into a scalar minimization problem

$$g(t) = \frac{1}{2}(t - t_0)^2 + \lambda|t|, \qquad (26)$$

whose solution is given by a soft shrinkage operator

$$S_\lambda(t) = \{ \begin{array}{ll} 0 & |t_0| \le \lambda \\ t_0 - sgn(t_0)\lambda & |t_0| > \lambda \end{array} . \qquad (27)$$

The basic idea behind surrogate functions is to show that the simple procedure of iterative shrinkage in the scalar case is also applicable to more general case (i.e., $\mathbf{D}$ is not unitary) [8]. In [8], authors have introduced the following surrogate function

$$\mathbf{\Psi}(\alpha, \alpha_0) = \frac{c}{2}||\alpha - \alpha_0||_2^2 - \frac{1}{2}||\mathbf{D}\alpha - \mathbf{D}\alpha_0||_2^2, \qquad (28)$$

where c is chosen to make $\mathbf{\Psi}(\alpha, \alpha_0)$ convex. Then the surrogate objective function for Eq. (23) becomes

$$\begin{aligned} \tilde{f}(\alpha, \alpha_0) &= \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda||\alpha||_1 \\ &+ \frac{c}{2}||\alpha - \alpha_0||_2^2 - \frac{1}{2}||\mathbf{D}\alpha - \mathbf{D}\alpha_0||_2^2. (29) \end{aligned}$$

After some manipulation, the above function can be simplified into

$$\tilde{f}(\alpha, \alpha_0) = const + \lambda||\alpha||_1 + \frac{c}{2}||\alpha - \mathbf{v}_0||_2^2, \qquad (30)$$

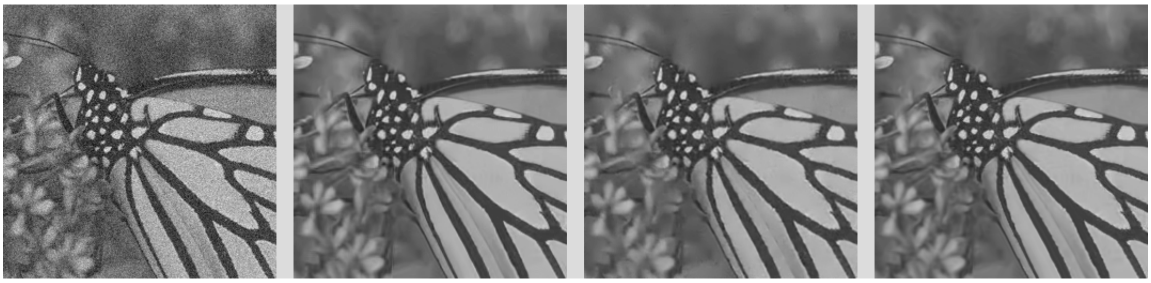

Figure 6. Denoising performance comparison for $monarch$ image: a) noisy ($\sigma_w = 20$); b) BM3D ($PSNR = 30.37dB, SSIM = 0.9209$); c) K-SVD ($PSNR = 29.89dB, SSIM = 0.9075$); d) CSR ($PSNR = 30.70dB, SSIM = 0.9197$).

where $\mathbf{v}_0 = \frac{1}{c}\mathbf{D}^T(\mathbf{x} - \mathbf{D}\alpha_0) + \alpha_0$. This form is similar to Eq. (25), which admits the following iterative shrinkage solution

$$\alpha_{i+1} = S_{\lambda/c}[\frac{1}{c}\mathbf{D}^T(\mathbf{x} - \mathbf{D}\alpha_i) + \alpha_i]. \qquad (31)$$

Next, we show how to solve the double-header l_1-optimization problem in Eq. (9) via surrogate functions. Without loss of generality, we describe our result for a single patch $\mathbf{x}$ and a chosen cluster (so the subscript k can be dropped). The simplified objective function is given by

$$f(\alpha, \beta) = \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda_1||\alpha||_1 + \lambda_2||\alpha - \beta||_{\mathbf{1}}, \qquad (32)$$

Similarly, we introduce the following surrogate objective function

$$\begin{aligned} \tilde{f}(\alpha, \beta, \alpha_0) &= \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda_1||\alpha||_1 + \lambda_2||\alpha - \beta||_1 \\ &+ \frac{c}{2}||\alpha - \alpha_0||_2^2 - \frac{1}{2}||\mathbf{D}\alpha - \mathbf{D}\alpha_0||_2^2. \qquad (33) \end{aligned}$$

After some similar manipulation to Eq. (29), we can simplify the above function into

$$\tilde{f}(\alpha, \alpha_0, \beta) = \mathit{const} + \lambda_1||\alpha||_1 + \lambda_2||\alpha - \beta||_1 + \frac{c}{2}||\alpha - \mathbf{v}_0||_2^2, \qquad (34)$$

where the definition of $\mathbf{v}_0$ is the same as above.

After translating the above minimization problem into its scalar version, we obtain

$$g(t) = \frac{1}{2}(t - t_0)^2 + \tau_1|t| + \tau_2|t - b|, \qquad (35)$$

where $\tau_1 = \frac{\lambda_1}{c}, \tau_2 = \frac{\lambda_2}{c}$ are scaled relaxation parameters and b is the scalar component of β. It follows that the solution to Eq. (32) is given by

$$\alpha_j^{(i+1)} = \begin{cases} \mathbf{S}_{\tau_1, \tau_2, \beta_j}(v_j^{(i)}) & \beta_j \geq 0 \\ -\mathbf{S}_{\tau_1, \tau_2, -\beta_j}(-v_j^{(i)}) & \beta_j < 0 \end{cases} \qquad (36)$$

where

$$\mathbf{v}^{(i)} = \frac{1}{c}\mathbf{D}^T(\mathbf{x} - \mathbf{D}\alpha^{(i)}) + \alpha^{(i)}. \qquad (37)$$

the generalized shrinkage operator $\mathbf{S}_{\tau_1, \tau_2, b}(t)$ is defined by

$$\mathbf{S}_{\tau_1, \tau_2, b}(t) = \begin{cases} t + \tau_1 + \tau_2 & t < -\tau_1 - \tau_2 \\ 0 & -\tau_1 - \tau_2 \leq t \leq \tau_1 - \tau_2 \\ t - \tau_1 + \tau_2 & \tau_1 - \tau_2 < t < \tau_1 - \tau_2 + b \\ b & \tau_1 - \tau_2 + b \leq t \leq \tau_1 + \tau_2 + b \\ t - \tau_1 - \tau_2 & t > \tau_1 + \tau_2 + b \end{cases} \qquad (38)$$

The approach based on surrogate functions can be interpreted as a proximal-point algorithm in convex optimization or a nonexpansive mapping in fixed-point theory. It is straightforward to justify the nonexpansive property for operator $\mathbf{S}_{\tau_1, \tau_2, b}(t_0)$ (i.e., $|\mathbf{S}_{\tau_1, \tau_2, b}(t_0)| \leq |t_0|$).

References

[1] E. J. Candès, J. K. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information." *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. on Image Proc.*, vol. 15, no. 12, pp. 3736–3745, December 2006.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696.

[4] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," *CVPR*, vol. 2, pp. 60–65, 2005.

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. on Image Processing*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[6] P. Chatterjee and P. Milanfar, "Clustering-based denoising with locally learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1438–1451, 2009.

[7] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2272–2279.

Table 1. The PSNR (dB) results for different denoising methods. In each cell, the results of four denoising algorithms are reported. Top left: SA-DCT; Top right: K-SVD; Bottom left: BM3D; Bottom right: CSR (this work).

σ	5		10		15		20		25		30	
Lena	38.54	38.62	35.58	35.51	33.87	33.72	32.63	32.38	31.66	31.35	30.86	30.46
	38.71	**38.74**	**35.94**	35.90	**34.29**	34.20	**33.07**	32.96	**32.08**	31.98	**31.28**	31.16
Monarch	38.02	37.77	33.87	33.69	31.61	31.43	30.11	29.90	29.02	28.74	28.09	27.80
	38.23	**38.43**	34.14	**34.49**	31.88	**32.25**	30.38	**30.71**	29.26	**29.52**	28.36	**28.56**
Barbara	37.50	38.11	33.52	34.42	31.39	32.38	29.98	30.81	28.93	29.58	28.07	28.57
	38.33	**38.43**	34.97	**35.10**	33.09	**33.17**	31.74	**31.78**	**30.66**	**30.66**	**29.77**	29.72
Boat	37.14	37.23	33.63	33.64	31.78	31.73	30.48	30.36	29.45	29.28	28.62	28.43
	37.22	**37.31**	**33.92**	33.88	**32.15**	32.05	**30.89**	30.78	**29.92**	29.78	**29.11**	28.94
C. Man	38.13	37.85	33.92	33.70	31.67	31.44	30.21	29.96	29.14	28.93	28.29	28.07
	38.29	38.18	**34.13**	34.06	**31.91**	31.89	**30.51**	30.49	**29.51**	29.48	**28.70**	28.64
Couple	37.32	37.30	33.72	33.48	31.73	31.44	30.34	29.98	29.27	28.85	28.41	27.91
	37.49	37.41	**34.02**	33.95	**32.10**	32.00	**30.75**	30.60	**29.70**	29.52	**28.84**	28.62
F. Print	35.90	36.66	31.71	32.42	29.58	30.06	28.15	28.45	27.06	27.25	26.17	26.28
	36.59	**36.85**	32.53	**32.70**	30.35	**30.47**	28.87	**28.97**	27.76	**27.84**	26.88	**26.95**
Hill	37.04	37.03	33.45	33.38	31.61	31.46	30.39	30.15	29.49	29.19	28.77	28.37
	37.12	**37.12**	33.62	**33.66**	**31.88**	31.87	**30.73**	30.65	**29.85**	29.75	**29.14**	28.97
House	39.44	39.42	36.03	36.03	34.14	34.35	32.89	33.16	31.93	32.19	31.12	31.24
	39.91	**39.98**	36.82	**36.88**	35.07	**35.11**	**33.92**	33.86	**32.99**	32.98	**32.21**	32.11
Man	37.59	37.50	33.71	33.55	31.65	31.44	30.27	30.05	29.26	29.02	28.49	28.23
	37.80	37.78	33.94	**33.96**	31.88	**31.91**	30.54	**30.56**	**29.56**	**29.56**	**28.81**	28.75
Peppers	37.96	37.78	34.51	34.25	32.54	32.22	31.11	30.77	30.01	29.69	29.09	28.82
	38.09	38.03	**34.72**	34.64	**32.75**	32.69	**31.31**	31.25	**30.23**	30.14	**29.31**	29.22
Straw	34.99	35.47	30.21	31.00	27.82	28.58	26.26	26.95	25.09	25.70	24.11	24.69
	35.44	**35.89**	30.99	**31.51**	28.67	**29.14**	27.10	**27.50**	25.93	**26.21**	24.99	**25.16**
Average	37.47	37.56	33.66	33.76	31.62	31.69	30.24	30.24	29.19	29.15	28.34	28.24
	37.77	**37.85**	34.15	**34.23**	32.17	**32.23**	30.82	**30.84**	**29.79**	**29.79**	**28.95**	28.90

[8] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[9] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[10] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Modeling and Simulation*, vol. 4, no. 2, pp. 460–489, 2005.

[11] M. Zibulevsky and M. Elad, "L1-l2 optimization in signal and image processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 76 –88, may. 2010.

[12] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998, pp. 839–846.

[13] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, "Deconvolutional networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2528–2535.

[14] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[15] N. Galatsanos and A. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Transactions on Image Processing*, vol. 1, no. 3, pp. 322 – 336, Mar 1992.

[16] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2744–2756, Nov 2002.

[17] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[18] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. on Image Processing*, vol. 16, no. 5, pp. 1395–1411, May 2007.