MULTISCALE DICTIONARY LEARNING FOR HIERARCHICAL SPARSE REPRESENTATION

Yangmei Shen, Hongkai Xiong

Department of Electronic Engineering
Shanghai Jiao Tong University, China
{shenyangmei0214, xionghongkai}@sjtu.edu.cn

Wenrui Dai

Department of Biomedical Informatics
University of California, USA
wed004@ucsd.edu

ABSTRACT

In this paper, we propose a multiscale dictionary learning framework for hierarchical sparse representation of natural images. The proposed framework leverages an adaptive quadtree decomposition to represent structured sparsity in different scales. In dictionary learning, a tree-structured regularized optimization is formulated to distinguish and represent high-frequency details based on varying local statistics and group low-frequency components for local smoothness and structural consistency. In comparison to traditional proximal gradient method, block-coordinate descent is adopted to improve the efficiency of dictionary learning with a guarantee of recovery performance. The proposed framework enables hierarchical sparse representation by naturally organizing the trained dictionary atoms in a prespecified arborescent structure with descending scales from root to leaves. Consequently, the approximation of high-frequency details can be improved with progressive refinement from coarser to finer scales. Employed into image denoising, the proposed framework is demonstrated to be competitive with the state-of-the-art methods in terms of objective and visual restoration quality.

Index Terms— dictionary learning, multiscale representation , structured sparsity, hierarchical structure, image denoising

1. INTRODUCTION

Sparse representation over redundant dictionary is a powerful model to adapt real-world signals, which is validated by well-established theoretical frameworks and state-of-the-art empirical results [1]. Its basic assumption suggests that a natural signal $\mathbf{x} \in \mathbb{R}^m$ is approximately represented by a *sparse* linear combination of atoms selected from an *overcomplete dictionary* $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$ ($m < p$), with the corresponding sparse representation vector $\alpha \in \mathbb{R}^p$. In general, a sparse coding problem is formulated to derive the sparse representation model.

$$\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1, \tag{1}$$

where λ is a regularization parameter balancing fidelity and sparsity, and $\|\alpha\|_1$ is a sparsity-inducing norm leading to the well-known Lasso or basis pursuit problems.

In comparison to pre-defined analytical dictionaries e.g., wavelets, the trained dictionaries can significantly improve the approximation performance for natural images by capturing the varying structures [2, 3]. ℓ_1-regularized optimization was formulated to adaptively derive the dictionaries from the sampled signals with batch gradient descent. These batch procedures would fail for large-scale high-dimensional signals, due to high computational complexity and low convergence speed. Thus, *online* dictionary learning methods have been widely concerned to achieve faster convergence with guaranteed accuracy [4, 5]. However, ℓ_1-regularized optimization is still restricted for sparse representation of multiscale high-dimensional signals like images and videos, as it independently generates the atoms by ignoring their structural relationship [6, 7].

To sufficiently exploit prior knowledge, *structured sparsity* methods were developed to adopt sparsity-inducing regularization capable for the higher-order information about the patterns of nonzero coefficients. One such possibility is the search for group dictionaries, where group structures of bags of visual descriptors at image level are considered for image classification [8]. Another alternative has been the pursuit of hierarchical dictionaries, which involve a tree-structured group-Lasso penalty addressed efficiently by dedicated proximal methods [9]. Inspired by independent component analysis, [10] goes beyond one-dimensional patterns and puts a 2-D grid structure on decomposition coefficients to infer topographic dictionaries by network flow optimization. Since all the above algorithms work off-line, [11] develops an online structured learning scheme using variational methods, making it possible to efficiently process large and partially observable training data. However, all of these structured dictio-

The work was supported in part by the NSFC grants 61501294, 61622112, 61529101, 61472234, 61425011, China Postdoctoral Science Foundation 2015M581617 and Program of Shanghai Academic Research Leader 17XD1401900.

naries have been traditionally restricted to fixed atom scales, which is insufficient to characterize the diverse and complex natural phenomena.

On the other hand, multiscale dictionaries have been considered to take advantage of multiscale property and data matching capability. Mairal *et al*. [12] fully decomposed images through a quadtree structure and learn multiple sub-dictionaries from patches with different scales using optimization methods like K-SVD. Modeling dictionary as a multiplication of Discrete Cosine Transform by a learned sparse matrix, the double-sparsity formulation made the first successful attempt towards the harmonic analysis [13]. In the context of Wavelet, learning process was applied into the analysis domain of Wavelet decomposition, where separate sub-dictionaries at different bands are trained by K-SVD [14]. Recently, Sulam *et al*. [15] extended the double-sparsity model by replacing the DCT dictionary with a new cropped Wavelet decomposition, which enabled dictionary learning to be up-scaled to a relative higher dimension. However, multiscale dictionaries would degrade the performance of sparse representation without considering the underlying dependencies or hidden structures between dictionary atoms.

In this paper, we develop a multiscale dictionary learning framework to enable hierarchical sparse representation, which naturally organizes atoms in a hierarchy with a descending order for node-sizes and increasing frequencies from root to leaves. Adaptive quadtree decomposition is proposed to recursively partition the images based on local statistics, which groups low-frequency components into large patches and distinguishes high-frequency details in small ones. A hierarchical regularized optimization is formulated to enforce sparsity patterns with rooted and connected subtrees. Thus, effective decomposition of image content is achieved using atoms from different scales. To improve learning efficiency, a joint hierarchical sparse coding step by proximal gradient method and a separate multiscale atom update procedure via block-coordinate descent are alternately performed. The learned dictionary is able to make sparse representation based on patches across multiple scales under a constraint of the tree-structured prior for nonzero patterns. For signal approximation, large atoms near the root provide low-frequency components, whereas fine details are hierarchically refined by small atoms in finer scales. In a nut-shell, the proposed framework can effectively represent multiscale signals with hierarchical trained dictionaries from sampled signals with multiple scales constrained by tree-structured sparsity. To validate the efficacy, the proposed framework was employed into image denoising task. Experimental results show that it is competitive with the state-of-the-art methods and allows practical applications to take a more global outlook over the diversity of real world signals.

The rest of this paper is organized as follows. Section 2 presents the proposed framework of multiscale dictionary learning for hierarchical sparse representation. Experimental results are shown in Section 3 for validation. Finally, we conclude the contributions in Section 4.

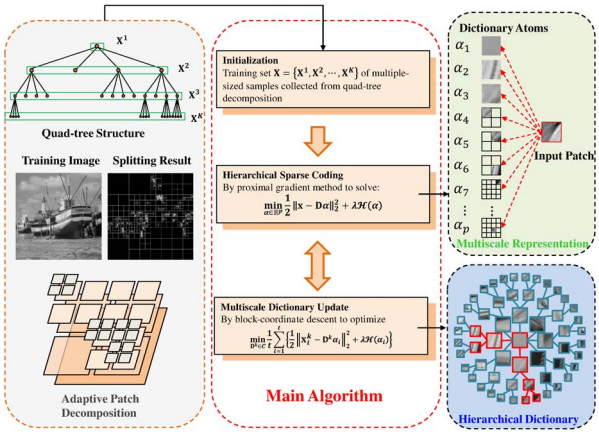


Fig. 1. The proposed framework for dictionary learning.

2. LEARNING MULTISCALE DICTIONARY FOR HIERARCHICAL SPARSE REPRESENTATION

It is widely known that natural image information spreads across multiple scales. Depending on specific structures, different images prefer different patch sizes for optimal representations. As depicts in Fig. 1, we present an attempt to explicitly exploit multiple scales simultaneously: using an efficient quadtree (QT) decomposition, an input image is recursively split into quadrants up to the selected sizes based on local features; by alternating between a hierarchical sparse coding step and a multiscale dictionary update procedure, dictionary is learned to sparsify and finely adapted to the training data; besides, a tree-structured sparsity prior is enforced to organize the learned atoms in a prespecified dendriform fashion, with larger atoms close to the root whereas the smaller near the leaves.

2.1. Adaptive Quadtree Decomposition

To achieve a variable size partition while avoiding the cost of more sophisticated techniques, an efficient quadtree decomposition is employed due to its effective balance between adaptivity of segmentation and simplicity of implementation. As for our setting, the primary purpose is to isolate the high detail regions into small sizes while grouping the low frequency regions into patches as large as possible, expecting to enhance the potential expressive force of the dictionary. Inspired by [16], local residual mean and variance values are jointly used as a simple yet effective measurement to assess the amount of details in a patch.

Given an input image, it is broken into fully overlapping patches of $\sqrt{m} \times \sqrt{m}$ pixels which are treated as independent

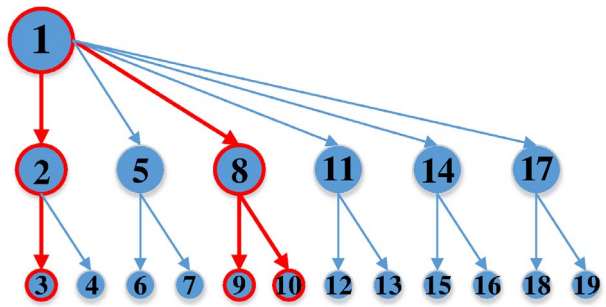


Fig. 2. Illustration of a tree-structured dictionary with $p = 19$. Atoms are embedded in nodes with decreasing sizes from root to leaves. Every node together with its descendants compose the tree-structured set of groups $\mathcal{G}$. The nonzero atoms form a connected and rooted subtree (in red contour) and the remaining nodes respect the constraint (2).

root nodes of quadtrees. For each root-patch, the local mean is calculated and removed to form the residual. At this stage, a binary decision is made whether to terminate the process or descend further into the tree. For this purpose, both the local variance and magnitudes of means for all $\sqrt{s}\times\sqrt{s}$ subpatches within the current $\sqrt{m}\times\sqrt{m}$ residual root-patch are computed and compared with corresponding thresholds T_v and T_m. If either the variance is sufficiently small or *all* the $\sqrt{s}\times\sqrt{s}$ means drop below T_m, this patch is identified as a leaf node. Otherwise, the mean-removed residual becomes a new parent node divided further into four children of size $\frac{m}{4}$ pixels. In turn, all the internal nodes are processed in the same manner, until obtaining leaves or reaching the allowed minimum size $\sqrt{s}\times\sqrt{s}$ in the tree.

The rational behind this strategy is that natural images typically contain large smooth areas as well as strong discontinuities such as textures, sharp edges and corners. While the patch variances consistently reflect the visual saliency, the means of image intensity usually vary quite slowly and even keep fairly constant over large smooth regions. Therefore for the latter case, if every sample in such a homogeneous region is minus the average sample amplitude, the ultimate mean value of an arbitrary patch in the mean-removed residual region will approach to zero.

2.2. Hierarchical Sparse Representation

In order to infer a dictionary to simultaneously capture frequency information from the different-sized examples, we propose to encode a tree-structured prior across multiple scales, for an intriguing property of increasing frequencies from root to leaves.

In Fig. 2, dictionary atoms $\{\mathbf{d}_1,\cdots,\mathbf{d}_p\}\in\mathbf{D}$ are embedded in a directed tree $\mathcal{T}$ of p nodes with sizes presenting in decreasing order from root to leaves. For an input signal $\mathbf{x}\in\mathbb{R}^m$, we expect its sparse decomposition vector $\alpha\in\mathbb{R}^p$ w.r.t. the dictionary admits a specific form of nonzero pattern: a rooted and connected subtree of $\mathcal{T}$. Define $descendants(j)\subseteq\{1,\cdots,p\}$ consists of the node j and all its descendants, such a constraint can be formulated as follow

$$\alpha_j=0\Rightarrow[\alpha_k=0 \text{ for all } k\in descendants(j)], \quad (2)$$

with a description that if a dictionary atom is not used in the decomposition, its descendants in the tree should not appear in the decomposition either. Moreover, a convex relaxation has been proposed and applied in different contexts [17, 18, 19]. Denoting $2^{\{1,\cdots,p\}}$ the power-set composed of all the 2^p subsets of $\{1,\cdots,p\}$, we firstly give the follow definition.

Definition 1 (Tree-Structured Set of Groups)**.** *Given a directed tree $\mathcal{T}$ containing p nodes, a tree-structured set of groups $\mathcal{G}\triangleq\{g\}_{g\in\mathcal{G}}\subseteq 2^{\{1,\cdots,p\}}$ consists of all the paths starting from every node in the tree down to leaves. For such a set of groups, it satisfies that $|\mathcal{G}|=p$ and $\bigcup_{g\in\mathcal{G}}=\{1,\cdots,p\}$, moreover, for any two groups $g,h\in\mathcal{G}$, if $g\cap h\neq\emptyset$, it must holds that either $g\subset h$ or $h\subset g$.*

For a tree-structured set of groups $\mathcal{G}$, the hierarchical sparsity-inducing norm $\mathcal{H}$ is defined as

$$\mathcal{H}(\alpha)\triangleq\sum_{g\in\mathcal{G}}\omega_g\|\alpha_g\|_2, \quad (3)$$

where $\alpha_g\in\mathbb{R}^{|g|}$ is the sub-vector consisting of the entries of α indexed by g, and ω_g denotes a positive scalar weight for group g. Indeed, when regularizing by $\mathcal{H}$, some of the sub-vectors α_g are set to zero for some groups $g\in\mathcal{G}$, implying the corresponding nodes in some complete subtrees of $\mathcal{T}$ are removed from the sparse linear combination. Consequently, the rest of nodes exactly form the desired connected and rooted subtree-structured sparsity pattern.

Since the dictionary is shared by all the training patches, the root-atom must appear in every sparse decomposition, gathering mostly the low frequencies; conversely, the deeper the atoms in the tree, the more specific they become, and the more high frequencies are involved. This makes sense in light of a hierarchical representation: by linearly combining atoms from root to leaves, large atoms provide low-frequency information, whereas abundant details are progressively refined by finer scales in deeper layers.

2.3. Multiscale Dictionary Learning

Employing the hierarchical regularization $\mathcal{H}$ in learning process means that the learned atoms will self-organize to match the tree-structured prior. In our scheme, we propose to alternate a *joint* hierarchical sparse coding procedure with a *separate* multiscale dictionary update stage to cope with learning across several existing scales.

Algorithm 1 summarizes the proposed learning scheme. Before sparse coding, an appropriate pre-processing of zero-padding is applied to balance the discrepancy across scales.

Algorithm 1 Multiscale Dictionary Learning.

Input: Multiple-sized training samples collected from QT decomposition $\mathbf{X} = \{\mathbf{X}^1, \cdots, \mathbf{X}^K\}$, pre-defined tree-structured set of groups $\mathcal{G} = \{g\}_{g \in \mathcal{G}}$, positive weights $\{\omega_g\}_{g \in \mathcal{G}}$, regularization parameter λ, number of scales K, number of iterations T.
Initialization: $\mathbf{D}^0 \in \mathbb{R}^{m \times p}$.
for $t = 1$ to T **do**
  Randomly sample root-patch $\mathbf{x}_t^1 \in \mathbb{R}^m$ from $\mathbf{X}^1$.
  Joint hierarchical sparse coding:
  Proximal gradient method to solve $\alpha_t \triangleq \arg\min_{\alpha \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_t^1 - \mathbf{D}_{t-1}\alpha\|_2^2 + \lambda\mathcal{H}(\alpha)$.
  Separate multiscale dictionary update:
  for $k = 1$ to K **do**
    Select sub-patches $[\mathbf{x}_1^k, \cdots, \mathbf{x}_t^k] \in \mathbf{X}^k$ used during previous iterations.
    Extract columns $\mathbf{D}_{t-1}^k = [\mathbf{d}_1^k, \cdots, \mathbf{d}_{p_k}^k] \in \mathbb{R}^{m_k \times p_k}$ with zero-padding removed.
    Block-coordinate descent with warm restart $\mathbf{D}_{t-1}^k$ to optimize $\mathbf{D}_t^k \triangleq \arg\min_{\mathbf{D}^k \in \mathcal{C}^k} \frac{1}{t}\sum_{i=1}^t \left\{\frac{1}{2}\|\mathbf{x}_i^k - \mathbf{D}^k\alpha_i^k\|_2^2 + \lambda\mathcal{H}(\alpha_i)\right\}$.
  end for
  Add zero-padding to $\mathbf{D}_t^k$ for all $k \in \{1, \cdots, K\}$.
  Put back each $\mathbf{D}_t^k$ in position to obtain $\mathbf{D}_t$.
end for
Return $\mathbf{D}_T$.

While $\sqrt{m} \times \sqrt{m}$ root-atoms keep unchanged, each small atom of $\frac{m}{4^{k-1}}$ pixels in layer k is randomly embedded into a big patch of size $\sqrt{m} \times \sqrt{m}$ with zero-padding elsewhere. In this way, the tree-structured dictionary with multiple-sized nodes can be simply used as a single-scale dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$.

Joint Hierarchical Sparse Coding

For sparse coding, we follow the proximal gradient methods due to both optimal convergence rate for the class of first-order techniques and capability to handle large nonsmooth convex problems [9]. Regularized by the tree-structured norm $\mathcal{H}$ in (3), objective function of the sparse decomposition problem is formalized as follow

$$\min_{\alpha \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda\mathcal{H}(\alpha). \quad (4)$$

Here $\mathbf{x}$ denotes the input signal of dimension m, $\mathbf{D} \in \mathbb{R}^{m \times p}$ is the learned dictionary, and λ is a non-negative regularization parameter. Linearizing the smooth convex square loss $f(\alpha) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\alpha\|_2^2$ around the current estimate $\hat{\alpha}$ according to the first-order Taylor approximation, proximal problem is given below

$$\min_{\alpha \in \mathbb{R}^p} f(\hat{\alpha}) + \nabla f(\hat{\alpha})^T(\alpha - \hat{\alpha}) + \lambda\mathcal{H}(\alpha) + \frac{L}{2}\|\alpha - \hat{\alpha}\|_2^2, \quad (5)$$

where $L > 0$ is a parameter upper-bounding the Lipschitz constant of ∇f. The added quadratic term keeps the update in a neighborhood of $\hat{\alpha}$ where f stays close to its linear approximation.

In effect, problem (5) can be viewed as a special case of proximal operators associated with the tree-structured norm $\lambda\mathcal{H}$, which admits closed solutions by a dual approach. It has been further proved that the computation of the dual formulation amounts to calculating a composition of elementary proximal operators, which can be solved efficiently via accelerated gradient techniques. Complexity is close to linear in the number of dictionary atoms p, implying almost the same cost as traditional ℓ_1-norm regularized problems.

Separate Multiscale Dictionary Update

The procedure for updating dictionary atoms is based on block-coordinate descent with warm restarts [5]. Concretely, training set of patches collected from QT decomposition is divided into K subsets as $\mathbf{X} = \{\mathbf{X}^1, \cdots, \mathbf{X}^K\}$ where K is the number of scales. Each subset $\mathbf{X}^k = [\mathbf{x}_1^k, \cdots, \mathbf{x}_{n_k}^k] \in \mathbb{R}^{m_k \times n_k}$ with $1 \leq k \leq K$ includes sub-patches with size of $m_k = \frac{m}{4^{k-1}}$ pixels (as a reminder, m is the size of root node in the quadtree).

At each iteration t, an arbitrary root-sample $\mathbf{x}_t^1 \in \mathbb{R}^m$ is randomly picked from $\mathbf{X}^1$ for computing hierarchical sparse vector α_t over the previous dictionary estimate $\mathbf{D}_{t-1}$. For each scale k, the set of sub-patches used during previous iterations is chosen from $\mathbf{X}^k$ as $[\mathbf{x}_1^k, \cdots, \mathbf{x}_t^k]$. Dictionary columns are extracted with zero-padding removed to form $\mathbf{D}_{t-1}^k = [\mathbf{d}_1^k, \cdots, \mathbf{d}_{p_k}^k]$. Their corresponding positions in $\mathbf{D}_{t-1}$ is recorded as well. The new atoms from $\mathbf{D}_t^k$ are updated by minimizing the following cost function, which is a good estimate of the desired *expected* cost when t tends to infinity.

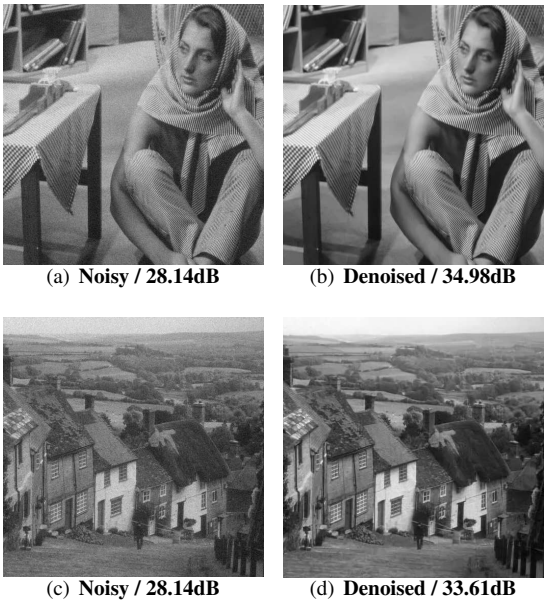
$$\min_{\mathbf{D}^k \in \mathcal{C}^k} \frac{1}{t}\sum_{i=1}^t \left\{\frac{1}{2}\|\mathbf{x}_i^k - \mathbf{D}^k\alpha_i^k\|_2^2 + \lambda\mathcal{H}(\alpha_i)\right\}, \quad (6)$$

where $\mathcal{C}^k \triangleq \{\mathbf{D}^k \in \mathbb{R}^{m_k \times p_k},\ \|\mathbf{d}_j^k\|_2 \leq 1 \text{ for all } j \in \{1, \cdots, p_k\}\}$ denotes a constraint set to avoid any degenerate solutions, and $\alpha_i^k \in \mathbb{R}^{p_k}$ is the sparse sub-vector of α_i corresponding to columns in $\mathbf{D}^k$. Since this problem admits separable constraints in updated blocks $\mathbf{d}_j^k$, a global optimum is given by iterating block-coordinate descent sequentially over columns along with an orthogonal projection onto the ℓ_2-ball.

After a few iterations, taking $\mathbf{D}_{t-1}^k$ as a warm restart for computing $\mathbf{D}_t^k$ has found to be effective. Accordingly, the above procedure is applied separately to each scale k to update atoms in $\mathbf{D}^k$, which should be repeated K times to achieve the complete update for dictionary $\mathbf{D}$. In fact, what makes the online dictionary update appealing is that it is significantly faster than batch alternatives such as K-SVD, yet it does not require a careful learning rate tuning like regular stochastic gradient descent methods.

3. EXPERIMENTS

In this section, we present experiments on denoising natural images compared to related methods, demonstrating the ap-



(a) **Noisy / 28.14dB** (b) **Denoised / 34.98dB**

(c) **Noisy / 28.14dB** (d) **Denoised / 33.61dB**

Fig. 3. Visual performance for images *barbara* and *hill* with noise level $\sigma = 10$.

plicability and potential of the proposed dictionary learning methods for improved representation of image content.

3.1. Implementation Details

In our experiments, denoising have carried out with 12 standard benchmark images, each of which is corrupted by synthetic white Gaussian noises with standard deviation σ in $\{5, 10, 15, 20, 25, 50, 100\}$ for pixel values in the range $[0; 255]$. For quadtree decomposition, the root node size $\sqrt{m}$ and the leaf node size $\sqrt{s}$ are chosen as 16 and 8 to conform with commonplace in dictionary learning. Threshold values T_m for means of all 8×8 patches and T_v for local variances with patch sizes $v \in \{16, 8\}$ are empirically set to $16, 80, 80$, respectively, to make an effective distinction between low- and high-details regions. In hierarchical sparse coding, a predefined balanced tree structure with 3 levels is tested. Branch numbers for nodes in level 0 and 1 are set to $\{128, 2\}$, respectively, and each child-node within depth 1 and 2 contains 2 and 1 atoms (with the root node contain no atom). These settings imply a dictionary with a total of 512 atoms at two different scales 16×16 and 8×8. For dictionary learning, initializations are either a standard DCT one or a generic dictionary learned on 10^6 natural image patches extracted randomly from Pascal VOC'12 database.

3.2. Results

Table 1 reports the results obtained on each image for different levels of noises, and Table 2 compares the average PSNR on 12 images with performance achieved by several state-of-the-arts approaches of the literatures, namely, GSM [20], K-SVD [21], BM3D [22], EPLL [23] and Plow [24]. Visual examples are shown in Fig. 3. For further validation, we also compared our results with another multiscale dictionary learning work [12] in Table 3, in which denoising was conducted on eight test images taken from our benchmark including *barbara*, *boat*, *cameraman*, *couple*, *hill*, *house*, *lena* and *peppers*. For lower noise ($\sigma \leq 10$), their algorithm performs slightly better than ours, but for other noise levels, we outperform theirs with a marginal improvement of 0.04dB. Note [12] is essentially a multiple-dictionary learning scheme which handles different areas of input image separately. Since our method only depends on a single dictionary, such a comparison is a bit unfair as more dictionaries have a built-in advantages over only one.

From these observations, conclusions are drawing as follows: First, our model consistently yields better performance than others in the whole range of noise level, except for the BM3D, which is well-known for outstanding denoising performance. Second, the gap between our results with BM3D is insignificant with a average of 0.04dB. One explanation for such a degradation is that the truly effective number of atoms in our dictionary is $\sum_{k=1}^{K} p_k/4^{k-1}$ with p_k the number of

Table 1. Denoising performance in PSNR(dB) on 12 standard images. 7 different noise levels of σ between 5 and 100 are tested.

σ	5	10	15	20	25	50	100
barbara	38.45	34.98	33.00	31.59	30.58	26.99	23.31
boat	37.32	33.89	32.10	30.82	29.87	26.66	23.70
bridge	35.59	31.10	28.84	27.44	26.43	23.70	21.43
cameraman	37.96	33.89	31.69	30.29	29.25	26.05	22.85
couple	37.36	33.86	31.92	30.61	29.58	26.24	23.12
fingerprint	36.61	32.50	30.30	28.79	27.65	24.35	21.21
flinstones	36.00	32.24	30.39	29.19	28.30	24.90	21.08
hill	37.10	33.61	31.84	30.66	29.78	27.00	24.20
house	39.91	36.84	34.96	33.71	32.78	29.56	24.96
lena	38.67	35.88	34.20	32.99	32.03	28.79	25.40
man	37.86	33.95	31.94	30.62	29.64	26.65	23.86
peppers	38.14	34.57	32.53	31.23	30.09	26.59	22.86
Average	37.58	33.94	31.98	30.66	29.67	26.46	23.17

Table 2. Quantitative comparative evaluation with GSM [20], K-SVD [21], BM3D [22], EPLL [23] and Plow [24]. PSNR results are averaged on 12 benchmark images, with best shown in bold.

σ	5	10	15	20	25	50	100
GSM	37.05	33.34	31.31	29.91	28.84	25.66	22.80
K-SVD	37.42	33.62	31.58	30.18	29.10	25.61	22.10
EPLL	37.36	33.64	31.67	30.32	29.29	26.12	23.03
Plow	37.38	32.98	31.38	30.13	29.30	26.38	23.24
BM3D	**37.62**	**34.00**	**32.05**	**30.73**	**29.72**	26.38	**23.25**
proposed	37.58	33.94	31.98	30.66	29.67	**26.46**	23.17

Table 3. Comparison with Multiscale K-SVD [12], with best shown in bold.

σ	5	10	15	20	25	50	100
[12]	**38.2**	**34.74**	**32.78**	31.46	30.45	**27.24**	23.67
proposed	38.11	34.69	**32.78**	**31.49**	**30.49**	**27.24**	**23.80**

atoms at scale k, since some atoms are mostly zeros; another possibility may derive from the too rigid structure (connected and rooted subtree) to flexibly select the most suitable atom in terms of error reduction for signal approximation. Further work is required to modify our scheme to achieve better gain.

4. CONCLUSION

This paper designs a multiscale dictionary learning scheme for hierarchical sparse representation. Quadtree decomposition based on local features is leveraged to partition training images into smooth and rich-details areas. Regularized by a tree-structured penalty, objective function is efficiently optimized by alternating a proximal gradient method with a block-coordinate descent. The learned dictionary naturally organizes atoms in a hierarchy with a descending order for node-sizes and increasing frequencies from root to leaves. Natural signals are sparsely decomposed along a connected and rooted subtree, with large atoms near the root provide low-frequency information, whereas abundant details are hierarchically refined by small atoms in finer scales. With application to image denoising, experimental results show that the proposed method is competitive with the state-of-the-art results over several natural images with various noise realizations.

5. REFERENCES

[1] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2-3, pp. 85–283, 2014.

[2] M. Aharon, M. Elad, and A. Bruckstein, "The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[3] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," 2007, pp. 801–808.

[4] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM J. Imaging Sci.*, vol. 1, no. 3, pp. 228–247, Jul. 2008.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[6] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.

[7] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Stat. Sci.*, vol. 27, no. 4, pp. 450–468, 2012.

[8] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Adv. Neural Inf. Process. Syst. (NIPS)*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, pp. 82–89, Curran Associates, Inc.

[9] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," *J. Mach. Learn. Res.*, vol. 12, pp. 2297–2334, 2011.

[10] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Convex and network flow optimization for structured sparsity," *J.Mach. Learn. Res.*, vol. 12, pp. 2681–2720, 2011.

[11] Z. Szabó, B. Póczos, and A. Lőrincz, "Online group-structured dictionary learning," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2011, pp. 2865–2872.

[12] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Modeling Simul.*, vol. 7, no. 1, pp. 214–241, Apr 2008.

[13] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, March 2010.

[14] B. Ophir, M. Lustig, and M. Elad, "Multi-scale dictionary learning using wavelets," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1014–1024, Sept 2011.

[15] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3180–3193, Jun. 2016.

[16] J. Vaisey and A. Gersho, "Image compression with variable block size segmentation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 2040–2060, Aug 1992.

[17] G. Rocha P. Zhao and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Ann. Stat.*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[18] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010.

[19] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion, "Multiscale mining of fmri data with hierarchical structured sparsity," *SIAM J. Imaging Sci.*, vol. 5, no. 3, pp. 835–856, 2012.

[20] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli., "Image denoising using scale mixtures of gaussians in the wavelet domain.," *IEEE Trans. on Image Process.*, vol. 12, no. 11, pp. 1338–1351, 2003.

[21] 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.

[22] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian., "Image denoising by sparse 3d transform-domain collaborative filtering.," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 600–616, 2007.

[23] D. Zoran and Y. Weiss., "From learning models of natural image patches to whole image restoration.," in *International Conference on Computer Vision (ICCV)*, 2011.

[24] P. Chatterjee and P. Milanfar., "Patch-based near-optimal image denoising.," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1635–1649, 2012.