Hyperspectral Image Denoising using Dictionary Learning

Cassio F. Dantas, Jérémy E Cohen, Rémi Gribonval

HYPERSPECTRAL IMAGE DENOISING USING DICTIONARY LEARNING

Cássio F. Dantas, Jérémy E. Cohen, Rémi Gribonval

Univ Rennes, Inria, CNRS, IRISA
Rennes, France

ABSTRACT

Hyperspectral images are corrupted by noise during their acquisition. In this work, we propose to efficiently denoise hyperspectral images under two assumptions: (i) noiseless hyperspectral images in matrix form are low-rank, and (ii) image patches are sparse in a proper representation domain defined through a dictionary. These two assumptions have already led to state-of-the-art denoising methods using fixed Wavelet transforms. We propose to rather learn the dictionary from hyperspectral images, a task commonly known as dictionary learning. We show that the dictionary learning approach is more efficient to denoise hyperspectral images than state-of-the-art methods with fixed dictionaries, at the cost of a larger computation time.

***Index Terms*—** Hyperspectral image, denoising, sparsity, low-rank, dictionary learning.

1. INTRODUCTION

Hyperspectral imaging has become a major image modality over the last years, largely thanks to the development of spectral sensors. Hyperspectral images collect reflectance spectra with significant spectral resolution (~200 wavelengths) for a large number of pixels (~1000×1000 pixels). These images contain a lot of information regarding the composition of the scene, and can therefore be used in remote sensing for monitoring forest or coastal regions, or in chemometrics to study the composition of chemical compounds [1].

However, Hyperspectral Images (HSI) are very often corrupted by various types of noise. In particular, for remotely acquired HSI, at least two kinds of noise are of importance: noise due to the spectral sensor sensitivity, and noise due to the swiping pattern of the sensors which yields stripes. Moreover, in the presence of clouds or other atmospheric perturbation, missing data may be present as well [2]. In this work, we will suppose that HSI are only corrupted with anisotropic Gaussian noise in order to simplify the analysis of our denoising method, but other types of noise as well as missing data can be tackled with similar tools.

The authors are with PANAMA research group. This work was supported by the Becose ANR project (ANR-15-CE23-0021).

Removing noise in HSI is an important pre-processing step for any learning task such as segmentation, detection or spectral unmixing, and has therefore been studied extensively in the literature. State-of-the-art techniques exploit two key properties [2]: HSI typically are low rank matrices once vectorized in the spatial dimensions; and HSI (or small patches of it) are sparse in some well chosen bases such as wavelets. In particular, the state-of-the-art method HyRes [3] makes use solely of these two assumptions to very efficiently denoise remotely acquired HSI with minimal computation time.

In this work, we propose to push further the boundaries of HSI denoising performance by switching from a fixed dictionary setting (wavelets) to a learned dictionary setting. Dictionary Learning (DL) is well established in HSI applications such as unmixing [4] and classification [5]. It has also been previously applied to HSI denoising in several occasions (see [6] and references therein) but rarely coupled with a low-rank decomposition. Although the low-rankness of the HSI is enforced via nuclear norm in [7], the dimensionality reduction promoted by an explicit low-rank factorization is not exploited at all as the full-size HSI is manipulated. Moreover, in [7], like in most DL methods for traditional image denoising [8], the noise variance is supposed known, while here we use an efficient heuristic to estimate it.

In what follows, we first formalize the denoising task and the low-rank and image sparsity assumptions in Section 2. In Section 3 we present our generic DL-based algorithm for HSI, and instantiate it with two DL algorithms, K-SVD [9] and SuKro [10]. Finally, in Section 4, we show on artificially noised HSIs that the proposed approach outperforms state-of-the-art methods for HSI denoising at the cost of increased computation time that is reasonable for offline applications.

2. SPARSE AND LOW-RANK MODELING

A hyperspetral image is composed of two spatial dimensions (image space) and one spectral dimension, and thus naturally represented as a 3 dimensional cube of data, say $\mathcal{H} \in \mathbb{R}^{n_1 \times n_2 \times p}$, where n_1 and n_2 correspond to the spatial dimensions and p is the number of spectral bands.

Nevertheless, a hyperspectral image is also often represented in its matrix (2 dimensional) form by vectorizing its spatial dimensions at each spectral band. In an additive noise

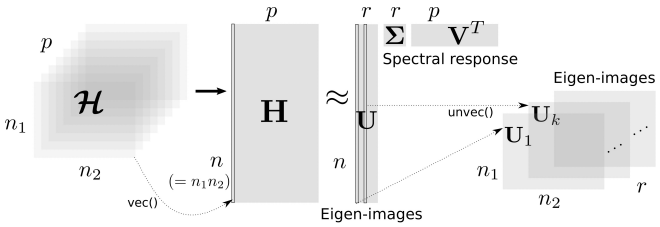


Fig. 1: Hyperspectral image modeling: 3D data cube (left), 2D representation with vectorized spatial dimensions (center) and low-rank factorization model (right).

model, this leads to:

$$\mathbf{H} = \mathbf{X} + \mathbf{N} \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{n \times p}$ (with $n = n_1 n_2$) containing as its j-th column the vectorized observed image at band j. $\mathbf{X}$ and $\mathbf{N}$ are respectively the noise-free unknown signal, and the noise matrix, both $(n \times p)$ matrices.

2.1. Low-rank assumption

An HSI $\mathbf{X}$ is commonly modeled as a low-rank matrix in the literature [3, 11]. The classical linear mixture model [12] is itself a low-rank model $\mathbf{X}{=}\mathbf{A}\mathbf{S}^T$, in which the image is decomposed into a few subregions called source images (as the columns of $\mathbf{A}$) each containing a certain material with distinct spectral signature (in the columns of $\mathbf{S}$). The HSI rank r in this case is the number of materials present in the image.

The singular value decomposition (SVD) is a way of obtaining an analogous low-rank decomposition of $\mathbf{X}$, even though it does not, in principle, promote spectral unmixing. It leads to the noisy low-rank model

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T + \mathbf{N} \tag{2}$$

where the columns of $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are respectively a set of vectorized *eigen-images* and the associated spectral components, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal singular value matrix. Figure 1 illustrates the described models and notations.

2.2. Sparsity assumption

The eigen-images $\mathbf{U}_k = \mathrm{unvec}(\mathbf{u}_k)$ (with $\mathbf{u}_k$ the k-th column of $\mathbf{U}$) can, then, be sparsely represented in a well-suited base –or, more generally, representation domain. In the noisy case, sparse reconstruction thus works as a form of denoising. Indeed, sparse modeling for images has been used for decades now [8, 13], with very good performance in denoising tasks.

State-of-the art techniques in Hyperspectral image restauration [3] use a 2-D (Kronecker) orthogonal wavelet basis $\mathbf{B}$ in which the vectorized eigen-images $\mathbf{u}_k$ are represented as a sparse set of wavelet coefficients $\mathbf{w}_k$ (k-th column of sparse coefficient matrix $\mathbf{W}$): $\mathbf{U} = \mathbf{B}\mathbf{W}$. The full model becomes:

$$\mathbf{H} = \mathbf{B}\mathbf{W}\mathbf{\Sigma}\mathbf{V}^T + \mathbf{N} \tag{3}$$

2.3. A typical processing pipeline

In view of model (3), a usual denoising approach given the corrupted signal $\mathbf{H}$ –which is no longer low-rank due to the noise– consists in estimating $\hat{\mathbf{U}}$,$\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{V}}$ via SVD truncation and futher denoising $\hat{\mathbf{U}}$ via 2D-Wavelet shrinkage. The spectral component of the low-rank model $\hat{\mathbf{V}}$ can also be denoised (for instance, [14] uses a 1D-Wavelet shrinkage).

3. PROPOSED APPROACH

A first part of our contribution consists in replacing the classic orthogonal Wavelet basis in the described pipeline by a dictionary both 1) overcomplete and 2) learned from data.

3.1. Overcomplete learned dictionary

Dictionary learning is a consolidated research topic [8–10]. Both the overcompleteness and the fact of learning a dictionary from a targeted type of data has been shown to improve the performance of natural image denoising tasks [8] when compared to the classical orthogonal Wavelet-based approach. Any existing dictionary learning approach could be used for our purpose here. We propose to use two methods:

1. K-SVD [9] as a standard (baseline) algorithm for learning unstructured dictionaries;

2. SuKro [10] as an algorithm for learning Kronecker-structured dictionaries. This type of structure arises naturally when dealing with vectorized 2D-signals (the eigen-images in the columns of $\hat{\mathbf{U}}$). The usual Wavelet approaches (e.g. in the state-of-the-art HyRes [3]) are themselves Kronecker-structured.

An important drawback of the overcompleteness (and consequently, non-orthogonality) of the dictionary can be pointed out: it significantly complicates the estimation of the sparse coefficients when compared to the orthogonal case (for instance, in HyRes [3], this step comes down to a simple thresholding operation). Here, we will need to resort to the so-called sparse coding algorithms [13] which are a sub-optimal way of solving this –now ill-posed– problem.

Consequently, computational complexity (and running time) is significantly bigger for the proposed method when compared to the literature. However, we will see in Section 4 that a performance gain is obtained. Therefore, except in very constrained cases, the proposed technique is worthy.

3.2. A patch-based denoising approach

In order to avoid having a prohibitively large dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ (with $m > n$), which would be very hard to train, we rely on a patch-based approach as generally used in the DL literature for denoising tasks [8]. The dictionary is learned over

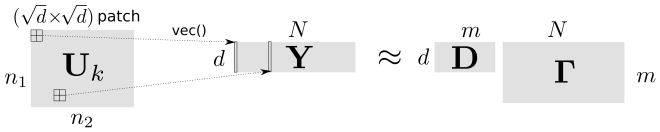


Fig. 2: N patches $(\sqrt{d} \times \sqrt{d})$ are extracted from a eigen-image $\mathbf{U}_k$ to form the training data matrix $\mathbf{Y}$, which is used to learn a dictionary $\mathbf{D}$ and a sparse reconstruction matrix $\mathbf{\Gamma}$.

small patches of the noisy image itself, making the algorithm completely self-contained.

Given an eigen-image $\hat{\mathbf{U}}_k \in \mathbb{R}^{n_1 \times n_2}$ we (sub)sample N equally-spaced training patches with dimensions $(\sqrt{d} \times \sqrt{d})$. The patches are vectorized and stacked as the columns of a training matrix $\mathbf{Y} \in \mathbb{R}^{d \times N}$. The corresponding dictionary $\mathbf{D} \in \mathbb{R}^{d \times m}$ with $m \geq d$ and sparse coefficient matrix $\mathbf{\Gamma} \in \mathbb{R}^{m \times N}$ are obtained by jointly optimizing the following cost function:

$$\{\hat{\mathbf{D}}, \hat{\mathbf{\Gamma}}\} = \text{argmin}_{\mathbf{D},\mathbf{\Gamma}} \|\mathbf{D}\mathbf{\Gamma} - \mathbf{Y}\|_F^2 + g(\mathbf{\Gamma}) \qquad (4)$$

where $g(\cdot)$ is a sparsity-inducing penalty function (for instance an ℓ_0 or ℓ_1 norm regularization on the columns of $\mathbf{\Gamma}$). Figure 2 summarizes the data structure and adopted notation.

The proposed denoising approach is described in Algorithm 1. Note that a new dictionary is learned for each eigen-image $\mathbf{U}_k$. To fully characterize the algorithm, some elements are still to be defined: the rank and noise estimation, sparse coding, dictionary update and denoising functions (respectively in lines 4, 5, 13, 14 and 16) are further detailed in sections 3.3 and 3.4.

Algorithm 1 Dictionary-based HSI denoising approach

1: **INPUT:** Noisy hyperspectral image $\mathbf{H}$
2: ▷ Low-rank step
3: $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \text{SVD}(\mathbf{H})$
4: $r = \text{EstimateRank}(\mathbf{\Sigma})$
5: $[\sigma_1, \ldots, \sigma_r] = \text{EstimateNoise}(\mathbf{\Sigma})$
6: $\hat{\mathbf{U}} = \mathbf{U}(:, 1:r)\,, \quad \hat{\mathbf{V}} = \mathbf{V}(:, 1:r)\mathbf{\Sigma}(1:r, 1:r)$
7: ▷ *Sparse step*
8: **for** $k = [1, \ldots, r]$ **do**
9: Extract N patches from $\hat{\mathbf{U}}_k$ to form $\mathbf{Y}$
10: ▷ *Dictionary learning*
11: Initialize $\mathbf{D}_0$
12: **for** $i = [1, \ldots, n_{\text{it}}]$ **do**
13: $\mathbf{\Gamma}_i = \text{SparseCoding}(\mathbf{Y}, \mathbf{D}_{i-1}, \sigma_k)$
14: $\mathbf{D}_i = \text{DictionaryUpdate}(\mathbf{Y}, \mathbf{\Gamma}_i)$
15: **end for**
16: $\hat{\mathbf{U}}_k = \text{Denoise}(\hat{\mathbf{U}}_k, \mathbf{D}_{n_{\text{it}}}, \sigma_k)$
17: **end for**
18: $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_r]$
19: **OUTPUT:** Recovered image $\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{V}}^T$

3.3. Dictionary Learning Algorithms

The customary DL framework to tackle problem (4) consists in iteratively alternating two optimization steps:

1. Sparse coding: updating $\mathbf{\Gamma}$ with fixed $\mathbf{D}$.
2. Dictionary update: optimizing for $\mathbf{D}$ with fixed $\mathbf{\Gamma}$.

Sparse coding There is a extensive literature on how to address this problem [13]. Although in theory any of the existing algorithms would serve our purposes, we use the traditional Othogonal Matching Pursuit (OMP) algorithm [15] which allows us to directly control the reconstruction precision of the input data. As a stopping criterion, we will use a threshold τ on the reconstruction error, i.e. the energy of the residual, proportional to the noise standard deviation σ (see Section 3.4 for more details on the noise level estimation) $\tau = a\sigma\sqrt{d}$, with a=1.16 an empirically-set gain following [8].

Dictionary update This is the step which distinguishes one DL algorithm from another. Two dictionary learning approaches are used:

1. K-SVD [9]: the columns of $\mathbf{D}$ are updated one at a time via a rank-one SVD approximation of the error matrix with respect to the current atom.
2. SuKro [10]: it learns structured dictionaries which can be written as a sum of a few Kronecker products (denoted $\boxtimes$) of smaller matrices $\mathbf{D}_{j,q} \in \mathbb{R}^{\sqrt{d} \times \sqrt{m}}$.

$$\mathbf{D} = \sum_{q=1}^{n_{\text{kron}}} \mathbf{D}_{1,q} \boxtimes \mathbf{D}_{2,q}. \qquad (5)$$

For the SuKro dictionary update step we use an Alternating Least-Squares (ALS) procedure where the blocks $D_{j,q}$ are optimized alternatively with a closed-form optimal update in each step. This algorithm, which is more efficient than the one originally proposed in [10], was proposed in [16] for a higher-order generalization of SuKro, i.e. for sums of Kronecker products of any number of blocks, and we apply it for the particular case of two blocks described in (5).

Computation of the denoised image Once the learning process is finished, the denoised version of the eigen-image $\hat{\mathbf{U}}_k$ is computed as follows. All possible patches from the noisy image –i.e. not only the N training patches, but all $(n_1 - \sqrt{d}+1)(n_2 - \sqrt{d}+1)$ overlapping patches with a one-pixel translation from its neighbors– are reconstructed with the learned dictionary using OMP with the same threshold τ. Each pixel in the final denoised image is then calculated as the weighted average of all reconstructed patches covering that pixel (each with weight one) and its value in the noisy image (with weight $\lambda = (\sigma\sqrt{n})^{-1}$). See [8] for details on why, in some sense, this is an optimal recovering procedure.

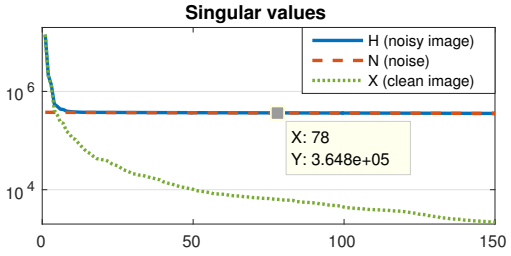


Fig. 3: Singular values of the San Diego HSI before (dotted green line) and after the noise addition (full blue line), as well as the noise singular values (dashed red) in logarithmic scale. Note that the noise singular values are very close to a constant $\sigma\sqrt{n}$ ($= 912 \times 400 = 3.648 \cdot 10^5$) as predicted by the theory.

3.4. Noise and Rank Estimation

In this subsection we show how the noise level can be efficiently estimated in practice under the assumption that the original image is low-rank and for a Gaussian distribution on the entries of the noise matrix $\mathbf{N}$.

First of all, as $\frac{n}{p} \gg 1$, the singular values of the noise matrix $\mathbf{N}$ are very close to a constant. Indeed, using the limit case of the Marčenko-Pastur distribution [17] when the column dimension is large in front of the row dimension, we get that all singular values of $\mathbf{N}$ are close to $\sigma\sqrt{n}$.

Moreover, as both the noiseless matrix $\mathbf{X}$ and the noise $\mathbf{N}$ are tall matrices generated by *a priori* independent processes, they are likely to be almost orthogonal to each other. Then,

$$\mathbf{H}^T\mathbf{H} \approx \mathbf{X}^T\mathbf{X} + \mathbf{N}^T\mathbf{N} \approx \mathbf{X}^T\mathbf{X} + \sigma^2 n\mathbf{I} \qquad (6)$$

and we get that the eigenvalues of $\mathbf{H}^T\mathbf{H}$ are the sum of those of $\mathbf{X}^T\mathbf{X}$ and a constant term $\sigma^2 n$.

Finally, since $\mathbf{X}$ is (approximately) a rank r matrix, any singular value of $\mathbf{H}$ after the r-th index is (approximately) equal to $\sigma\sqrt{n}$. Therefore, the noise level can be estimated by looking at the tail value ς_{tail} of the singular values of $\mathbf{H}$. Figure 3 shows that we indeed observe such a singular value profile in practice when adding Gaussian noise to a real HSI.

While $\hat{\sigma} = \varsigma_{\text{tail}}/\sqrt{n}$ estimates the noise level in $\mathbf{H}$, we are rather interested in the noise level σ_k at an eigen-image $\mathbf{U}_k$. Considering that each column $\mathbf{u}_k$ of $\mathbf{U}$ is scaled to unit-norm regardless of the associated singular value (denoted ς_k for the k-th largest singular value), we observed that σ_k is inversely proportional to ς_k and can be estimated as follows:

$$\hat{\sigma}_k = \sigma/\varsigma_k = \varsigma_{\text{tail}}/(\varsigma_k\sqrt{n}) \qquad (7)$$

The rank estimation is based on this same observation: the singular values saturate as the noise dominates. A suitable approximate rank r is obtained by detecting this saturation. We set $\hat{r}$ equals to the first singular value ς_k such that $(\varsigma_k - \varsigma_{\text{tail}})/\varsigma_{\text{tail}} < \epsilon$.

$$\text{EstimateRank}(\mathbf{\Sigma}) = \min\{k \mid (\varsigma_k - \varsigma_{\text{tail}})/\varsigma_{\text{tail}} < \epsilon\} \qquad (8)$$

where ϵ is a threshold empirically calibrated to $\epsilon = 3\times 10^{-2}$.

4. EXPERIMENTAL RESULTS

Four hyperspectral images were used in the experiments with $(n_1 \times n_2 \times p)$ pixels: *San Diego* ($400 \times 400 \times 158$), *Houston* ($349 \times 1905 \times 144$), *Washington DC* ($1280 \times 256 \times 191$), *Urban* ($307 \times 307 \times 162$). The images were corrupted with additive white Gaussian noise with standard deviation σ uniform over all spectral modes and adjusted to give the desired input signal-to-noise ratio (SNR).

For each eigen-image (columns of $\mathbf{U}_k$), we collect $N = 20000$ patches with (6×6) pixels from the noisy image to form the training data. As a pre-processing, each patch is centralized to zero mean in its pixel values. The patches are taken uniformly-spaced and partially overlapping. A dictionary is learned from this data with $n_{\text{it}} = 20$ iterations of alternating optimization. Dictionaries are initialized with an Overcomplete (cropped) 2D-DCT (ODCT). In the case of SuKro, the remaining terms are initialized with unit-norm Gaussian random matrices, with no relevant impact on its performance.

The results reported in Table 1 are averaged over 10 runs with independent noise realizations. Standard deviations are usually around 0.005dB (and never bigger than 0.02dB).

Three variations of the proposed method are tested, with three different types of dictionary: K-SVD unstructured learned dictionary, SuKro learned structured dictionary with $n_{\text{kron}} = 3$ terms and an overcomplete but fixed ODCT dictionary (i.e. the learning process is skipped). The proposed methods are compared to: a Wavelet 3D approach [18] which uses only the sparsity prior by applying a 3D-wavelet basis directly in the data cube $\mathcal{H}$; and the HyRes [3] method, which combines both sparse and low-rank assumptions. The latter is a state-of-the-art approach (see [3] for a detailed comparison of this method with other methods in the literature).

The proposed methods with K-SVD and SuKro dictionaries consistently outperform the state-of-the-art HyRes approach by about 1.5dB. A performance gain of about 1.3dB is already obtained by replacing the orthogonal Wavelet basis in HyRes by a fixed overcomplete DCT dictionary allied to the patch-based procedure. The learning process then brings an additional 0.2dB in denoising performance. A visual comparison is provided in Figure 4.

Execution times are reported in Table 2. The literature methods HyRes and Wavelet 3D take about one and two minutes respectively, independently of the noise level. The running times for the proposed method, on the other hand, increase as the noise level decreases. That's because the OMP reconstruction threshold gets smaller, requiring more iterations. The execution times range from 120 to 760s in the ODCT case and from 260 to 1600s with a learned dictionary (actually, SuKro is slightly faster than K-SVD). Therefore, the proposed technique with dictionary learning takes around 5 to 30 times longer than HyRes.

Table 1: Output SNR [dB] comparison with literature

	Method	Input SNR [dB]					
		5	10	15	20	25	30
San Diego	Wav. 3D	20.19	23.36	26.51	29.48	32.33	35.08
	HyRes	23.26	26.37	29.63	32.61	35.44	37.92
	ODCT	24.79	27.77	30.79	33.77	36.57	38.88
	K-SVD	24.92	27.86	30.89	33.88	36.68	**39.00**
	SuKro	**24.95**	**27.93**	**30.94**	**33.89**	**36.69**	38.99
Houston	Wav. 3D	18.28	21.59	24.90	28.19	31.45	34.64
	HyRes	22.86	26.45	29.76	33.00	36.08	39.49
	ODCT	24.18	27.49	30.87	34.13	37.30	39.85
	K-SVD	24.38	**27.64**	**31.01**	**34.27**	**37.45**	**39.99**
	SuKro	**24.39**	27.63	30.98	34.23	37.42	39.97
Washington	Wav. 3D	18.87	21.92	25.03	28.25	31.51	34.77
	HyRes	23.25	26.54	29.78	33.30	36.35	38.76
	ODCT	24.58	27.92	31.27	34.57	37.47	39.80
	K-SVD	24.74	28.04	31.37	34.68	**37.59**	**39.89**
	SuKro	**24.76**	**28.05**	**31.38**	**34.69**	**37.59**	39.88
Urban	Wav. 3D	18.37	21.58	24.78	27.85	30.82	33.68
	HyRes	22.02	25.38	28.45	31.29	33.95	36.01
	ODCT	23.34	26.52	29.54	32.34	34.76	36.54
	K-SVD	**23.45**	**26.63**	**29.65**	**32.44**	**34.86**	**36.61**
	SuKro	**23.45**	**26.63**	**29.65**	32.43	34.83	36.58

Table 2: Execution times (in seconds) for Washington image

	Method	Input SNR [dB]					
		5	10	15	20	25	30
Washington	Wav. 3D	125	120	125	125	120	124
	HyRes	55	50	50	49	48	50
	ODCT	128	201	296	408	533	761
	K-SVD	273	445	677	980	1315	1650
	SuKro	262	394	592	857	1143	1433

5. CONCLUSION

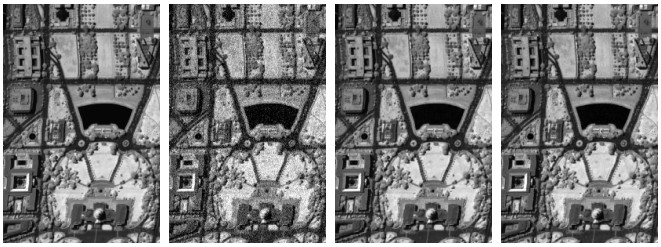
In this paper, we proposed to replace fixed sparsifying transforms by overcomplete learned dictionaries to improve state-of-the-art performance in the HSI denoising task.

Although the proposed patch-based approach is markedly slower than other previous methods, running times remain reasonable for an offline application, even for large images. A potential improvement would be to also denoise the right term of the low-rank decomposition (spectral component $\hat{\mathbf{V}}$).

6. REFERENCES

[1] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, Dec 2017.

[2] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, "Noise reduction in hyperspectral imagery: Overview and application," *Remote Sensing*, vol. 10, no. 3, 2018.



(a) Original (b) Noisy: 10dB (c) SuKro (d) HyRes

Fig. 4: Zoom in Washington image (100^{th} spectral band).

[3] B. Rasti, M. O. Ulfarsson, and P. Ghamisi, "Automatic hyperspectral image restoration using sparse and low-rank modeling," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, Dec 2017.

[4] A. Castrodad, Z. Xing, J. B. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse representations for modeling, source separation, and mapping of hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, Nov 2011.

[5] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, Oct 2011.

[6] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin, "Dictionary learning for noisy and incomplete hyperspectral images," *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 33–56, 2012.

[7] Y. Zhao and J. Yang, "Hyperspectral image denoising via sparse representation and low-rank constraint," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 296–308, Jan 2015.

[8] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, Nov 2006.

[10] C. F. Dantas, M. N. da Costa, and R. d. R. Lopes, "Learning dictionaries as a sum of kronecker products," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 559–563, May 2017.

[11] M. Golbabaee and P. Vandergheynst, "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *IEEE ICASSP*, Mar 2012, pp. 2741–2744.

[12] D. Manolakis, C. Siracusa, and G. Shaw, "Hyperspectral subpixel target detection using the linear mixing model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, Jul 2001.

[13] Alfred M. Bruckstein, David L. Donoho, and Michael Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb 2009.

[14] G. Chen and S. Qian, "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, Mar 2011.

[15] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[16] C. F. Dantas, J. E. Cohen, and R. Gribonval, "Learning Tensor-structured Dictionaries with Application to Hyperspectral Image Denoising," in *27th EUSIPCO*, A Coruña, Spain, Sep 2019.

[17] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, Apr 1967.

[18] Abdullah A Basuhail and Samuel Peter Kozaitis, "Wavelet-based noise reduction in multispectral imagery," in *Algorithms for multispectral and hyperspectral imagery IV*. International Society for Optics and Photonics, 1998, vol. 3372, pp. 234–241.