LETTER

Learning Deep Dictionary for Hyperspectral Image Denoising*

Leigang HUO[†,††], **Xiangchu FENG**[†], *Nonmembers*, **Chunlei HUO**[††a)], *Member*, *and* **Chunhong PAN**[††], *Nonmember*

SUMMARY Using traditional single-layer dictionary learning methods, it is difficult to reveal the complex structures hidden in the hyperspectral images. Motivated by deep learning technique, a deep dictionary learning approach is proposed for hyperspectral image denoising, which consists of hierarchical dictionary learning, feature denoising and fine-tuning. Hierarchical dictionary learning is helpful for uncovering the hidden factors in the spectral dimension, and fine-tuning is beneficial for preserving the spectral structure. Experiments demonstrate the effectiveness of the proposed approach.

key words: *dictionary learning, deep learning, hyperspectral image denoising, fine-tuning*

1. Introduction

Hyperspectral image denoising is an important preprocessing step for the subsequent procedures such as classification and target recognition. However, hyperspectral image denoising is more challenging than the multispectral images due to the complex noise caused by the high spectral bands.

Dictionary learning is an effective tool for image denoising. As for the hyperspectral images, in order to avoid the prohibitive computation and improve the separability between the signal and noise, the dictionaries are usually applied on the reduced feature space instead of the original high dimensional spectral space. For instance, Chen [1], [2] proposed to keep the first few significant components and denoise the components of low energies individually by bivariate wavelet thresholding [1] and BM3D [2], [3] respectively, followed by 1-d wavelet shrinkage along the spectral dimension. Lam [4] suggested projecting the hyperspectral image into the feature space spanned by the first few eigvectors and denoising the transformed components separately by the bilateral filtering. Lam's approach outperforms Chen's approaches [1], [2] due to the difficulty of the latter technique in selecting the proper "clean" components and balancing the fine features and noise. The other way to learn the dictionary is by matrix factorization. For instance, Cerra [5] proposed to denoise the hyperspectral images simultaneously with the unmixing procedure, where the endmember acts as the role of the dictionary.

Despite of the effectiveness in removing noise, the traditional dimension reduction and dictionary learning framework is limited for hyperspectral images due to the "shallow" architecture. In other words, the complex hidden relationship between the original high-dimension spectral space and the reduced low-dimension feature space is difficult to be revealed using the simple single-layer dictionary learning configuration. To address this problem, a novel hyperspectral image denoising approach is proposed based on deep dictionary learning. The main difference between the proposed approach and the traditional ones lies in the hierarchical dictionary learning architecture and the back-propagation mechanism. By the hierarchical dictionary learning architecture, the separation between the signals and noises is improved progressively. By taking advantages of the back-propagation mechanism, the spectral structure and the useful detail can be preserved.

2. The Proposed Approach

As illustrated by Fig. 1, three following steps are involved in the proposed approach, and we will elaborate the detail step by step below.

2.1 Hierarchical Dictionary Learning

Given the observed hyperspectral image $X \in R^{p\times n}$ and the clean version $Y \in R^{p\times n}$, the hyperspectral image denoising problem can be modeled as $X = Y + e$, where p is the band number, $n = h \times w$, h and w denote the row and column number of each band, the ith row of X is extracted from the ith band by the row-wise manner. e is the

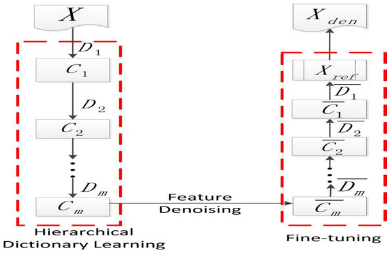

Fig. 1 The flowchart of the proposed approach

Manuscript received December 8, 2014.
Manuscript revised March 16, 2015.
Manuscript publicized April 20, 2015.

[†] The authors are with the School of Mathematics and Statistics, Xidian University, Xi'an, China.

[††] The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

[*] This work was supported by the Natural Science Foundation of China under Grants 61271294, 61375024, 91338202, 91438105.

a) E-mail: clhuo@nlpr.ia.ac.cn

DOI: 10.1587/transinf.2014EDL8246

additive signal-independent noise mainly caused by electronic devices. Dictionary learning is to learn a group of bases(dictionary) $D \in R^{p \times k}$ and approximate the hyperspectral image by $X = DC$. C denotes the decomposition coefficient.

The aim of hierarchical dictionary learning is to learn the hidden structures of the original data progressively. As illustrated in Fig. 1, hierarchical dictionary learning is realized by the successive dictionary learning. In detail, at the first layer, the hyperspectral image X is approximated as

$$X \approx D_1 C_1 \tag{1}$$

Where $D_1 \in R^{p \times k_1}$, $C_1 \in R^{k_1 \times n}$. p is the number of dictionary atoms, and k_1 the dimension of each atom. At the ith($2 \leq i \leq m$) layer, the feature(the decomposition coefficient) C_{i-1} achieved at the $i-1$th layer is decomposed as

$$C_{i-1} \approx D_i C_i \tag{2}$$

Where m is the levels of dictionary learning, $C_{i-1} \in R^{k_{i-1} \times n}$, $D_i \in R^{k_{i-1} \times p_i}$, $C_i \in R^{p_i \times n}$. The hierarchical dictionary learning problem can then be modeled as the concatenative matrix multiplication:

$$X \approx D_1 D_2 \cdots D_m C_m \tag{3}$$

The hierarchical dictionary learning procedure can be considered as the progressive dimension reduction layer by layer. For the traditional single-layer dimension reduction technique such as PCA, the dimensions of the features are reduced dramatically, and the real useful information may be killed and difficult to be recovered.

Various dictionary learning techniques can be used for this step without problem. Since the focus of this paper is validating the advantages of the deep dictionary learning framework, for simplicity, the dictionaries are learned by the convex semi-nonnegative matrix factorization [6], where the decomposition coefficient C_i contains only non-negative elements. Noting the role of C_i in representing the probability(abundance) that each voxel belongs to certain material types(end-member), this constraint is reasonable.

2.2 Feature Denoising

After hierarchical dictionary learning, the noises are mainly gathered at the mth layer $C_m \in R^{p_m \times n}$. In consequence, feature denoising is applied on C_m. In this paper, BM3D [3] is used to denoise each component of C_m. In detail, C_m is firstly arranged as the data-cube G with the size $p_m \times h \times w$. Then, each component in G, $G_i \in R^{h \times w}$, is denoised individually by BM3D, and the denoised data-cube is reorganized as 2-D matrix of the size $p_m \times n$. For convenience, the denoised version of C_m is still denoted as C_m.

2.3 Fine Tuning

Hierarchical dictionary learning is a biased feature learning procedure, the spectral structure will be impacted by various factors(such as the decomposition level, dictionary size and dictionary dimension) if the reconstructed version $\overline{X} = D_1 D_2 \cdots D_m C_m$ is used as the denoised result directly. To address this problem, the dictionary and decomposition coefficient at each layer are adjusted iteratively driven by the reference denoised hyperspectral image X_{ref}. Considering the facts that the ideally clean hyperspectral image is difficult to obtain and that the denoised hyperspectral images by the promising approaches are similar in performances and formed to a cluster, it is not necessary for X_{ref} to have the very high denoising performance. The role of X_{ref} is to direct the fine-tuning procedure, and the hyperspectral image denoising approaches in the literature can be used for generating X_{ref}. In other words, the fine-tuning procedure is a weakly supervised back-propagation, and the denoising performance will be enhanced gradually.

The objective of fine-tuning is to minimize the error between the reference hyperspectral image X_{ref} and the reconstructed denoised version by tuning the dictionary and the decomposition coefficient layer-by-layer. The cost function of the fine-tuning procedure is formulated as follows:

$$\begin{aligned} Error &= \frac{1}{2} \| X_{ref} - \overline{D_1}\,\overline{D_2} \cdots \overline{D_m}\,\overline{C_m} \|_F^2 \\ &= tr[X_{ref}^T X_{ref} - 2X_{ref}^T \overline{D_1}\,\overline{D_2} \cdots \overline{D_m}\,\overline{C_m}] + \\ &\quad \overline{C_m}^T\, \overline{D_m}^T \cdots \overline{D_1}^T\, \overline{D_1}\,\overline{D_2} \cdots \overline{D_m}\,\overline{C_m}. \end{aligned} \tag{4}$$

Where $\overline{D_i}$ and $\overline{C_i}$ denote the adjusted version of D_i and C_i respectively. By setting $\frac{\partial Error}{\partial \overline{D_i}} = 0$, the update rule of $\overline{D_i}$ can be achieved as follows:

$$\overline{D_i} = (A^T A)^{-1} A^T X_{ref} \overline{C_i}^T (\overline{C_i}\,\overline{C_i}^T)^{-1} = A^{\dagger} X_{ref} \overline{C_i}^{\dagger} \tag{5}$$

Where $A = \overline{D_1} \cdots \overline{D_{i-1}}$, $A^{\dagger} = (A^T A)^{-1} A^T$, $\overline{C_i}^{\dagger} = \overline{C_i}^T (\overline{C_i}\,\overline{C_i}^T)^{-1}$. Considering the non-negative constraint on C_i, the update rule of $\overline{C_i}$ is represented as follows based on Proposition 4 in [6]:

$$\overline{C_i}^{jk} = \overline{C_i}^{jk} \sqrt{\frac{[A^T X_{ref}]_+^{jk} + [A^T A]_-^{jk} \overline{C_i}}{[A^T X_{ref}]_-^{jk} + [A^T A]_+^{jk} \overline{C_i}}} \tag{6}$$

Where $P_+ = \frac{(|P|+P)}{2}$ and $P_- = \frac{(|P|-P)}{2}$ are the positive and negative part of the matrix P, respectively. P^{jk} denotes the element of P at the jth row and kth column.

The solution of fine-tuning is the alternative update of $\overline{D_i}$ and $\overline{C_i}$, whose initial values are D_i and C_i respectively. The fine-turning procedure starts from the first layer, and the alternative updates are repeated layer-by-layer until the stopping criterion is reached. In this paper, the stopping criterion means reaching the maximized iterations(e.g., 100) or the relative error between two successive iterations is less than a given threshold(e.g., 0.001).

After fine tuning, the final denoised hyperspectral image X_{den} is approximated by $X_{den} = \overline{D_1}\,\overline{D_2} \cdots \overline{D_m}\,\overline{C_m}$.

3. Experiments

To validate the effectiveness of the proposed approach, three datasets are illustrated in this paper, one is the synthetic hyperspectral image over Indian Pines generated using the ground truth of Indian Pines data [7] and the spectral signatures extracted from the USGS digital spectral library [8], and other two datasets are taken over Pavia Center by ROSIS [9] and Office by SOC710[10] respectively. Some bad bands are removed before denoising, and the datasets are described in detail in Table 1.

The novelties of the proposed approach lie in the hierarchical deep learning and fine-tuning. By experiments, we found that satisfied performances can be achieved when the layers are set to be 3, and the performances improved by more layers can be neglected. The layer configurations for each dataset are listed in Table 1. For convenience, the proposed approach is denoted as **DDL3+FT**. To validate the novelties of the proposed approach, other five related approaches are used for comparison:

(1) **PCA+BM3D**. PCA framework [4] is utilized, but BM3D [11] is used for denoising each component instead of the bilateral filter. And the results taken by PCA+BM3D is selected as the reference hyperspectral image X_{ref} for DDL3+FT.

(2) **rPCA+saBM3D**. Robust PCA framework [12] is utilized, and shape-adaptive BM3D [11] is used for denoising each component.

(3) **BM4D**[13]. BM4D is one of the state-of-arts approaches for the volumetric data denoising.

(4) **DDL3**. DDL3 is same as DDL3+FT in hierarchical deep learning, but fine-tuning is omitted. DDL3 is used to investigate the importance of fine-tuning.

(5) **DDL1+FT**. DDL1+FT is similar to DDL3+FT, but single-layer dictionary learning is used. DDL1+FT is utilized to demonstrate the advantages of multi-layer learning.

For the synthetic dataset, the spectral responses are normalized to $[0, 1]$, and the noises of varying standard variances($\sigma = 0.1, \cdots, 0.5$) are added to the synthetic data. $PSNR$(Peak Signal-to-Noise Ratio) between the denoised hypersectral image and the ground truth is used for performance evaluation. The performances on the synthetic data are listed in Table 2. For the real datasets, two metrics are

Table 1 Datasets description

Dataset	Size	Band number	Layer Configuration
Indian Pines	145*145	151	[80,40,20]
Office	640*640	120	[80,60,40]
Pavia Center	646*485	102	[90,60,30]

Table 2 Performance comparison on the synthetic dataset

σ	0.1	0.2	0.3	0.4	0.5
PCA+BM3D	39.5	31.8	26.5	22.8	18.3
rPCA+saBM3D	41.1	31.9	26.2	22.1	18.1
BM4D	36.0	30.8	27.7	25.7	22.8
DDL3+FT	40.9	31.9	28.2	26.5	23.1
DDL1+FT	39.6	31.9	26.6	23.0	18.6
DDL3	25.2	22.4	22.0	17.2	14.9

Table 3 Performance comparison on the real data sets

approach	Office	Pavia Center
PCA+BM3D	(0.83, 0.07)	(0.38, 0.05)
rPCA+saBM3D	(0.84, 0.08)	(0.37, 0.04)
BM4D	(0.84, 0.07)	(0.35, 0.04)
DDL3+FT	(0.86, 0.06)	(0.41, 0.04)
DDL1+FT	(0.84, 0.07)	(0.22, 0.14)
DDL3	(0.45, 0.29)	(0.20,0.31)

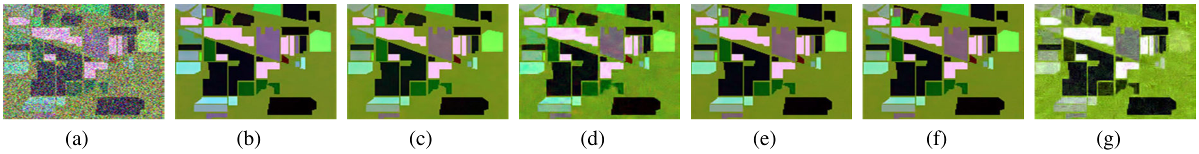

Fig. 2 Results comparison on Indian Pines, $\sigma = 0.3$. (a): noisy pseudo-color image at bands(1,76,151), (b): PCA+BM3D, (c): rPCA+saBM3D, (d): BM4D, (e): DDL3+FT, (f): DDL1+FT, (g):DDL3.

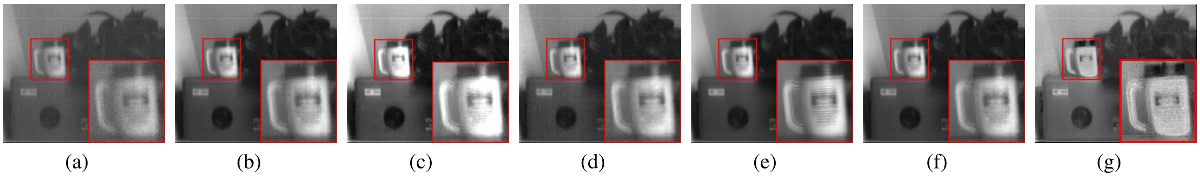

Fig. 3 Results comparison on Office. (a): noisy image at band 1, (b): PCA+BM3D, (c): rPCA+saBM3D, (d): BM4D, (e): DDL3+FT, (f): DDL1+FT, (g): DDL3.

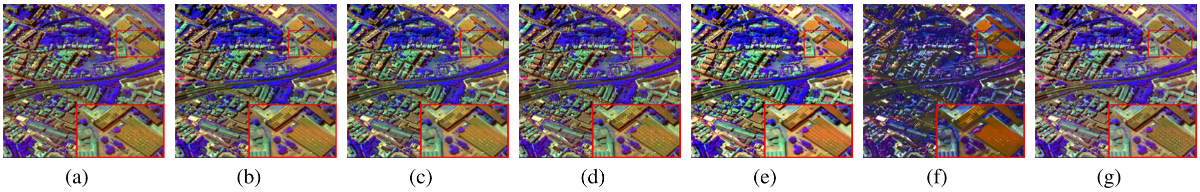

Fig. 4 Results comparison on Pavia Center. (a): noisy pseudo-color image at bands (1,51,102), (b): PCA+BM3D, (c): rPCA+saBM3D, (d): BM4D, (e): DDL3+FT, (f): DDL1+FT, (g): DDL3.

utilized for performance evaluation: the blind image quality index Q_{bqi}[14] and spectral angle mapper Q_{sam}[15]. Q_{bqi} lies within the range $(-1, 1)$, and it is used to measure the denoised image quality. The larger Q_{iad} is, the better denoising performance will be. Q_{sam} is used to evaluate the spectral structure preservation confidence. Smaller Q_{sam} means the better performance. In Table 3, the performances $(\overline{Q_{bqi}}, Q_{sam})$ on the real datasets by different approaches are listed, where $\overline{Q_{bqi}}$ means the averaged Q_{bqi} over bands.

From Table 2, it can be indicated that the proposed approach, DDL3+FT, outperforms other techniques. For instance, at $\sigma = 0.1$, PSNR is improved from 39.5 by PCA+BM3D and 36.0 by BM4D to 40.9 by DDL3+FT. DDL3+FT is superior to PCA+BM3D and BM4D due to the multi-layer dictionary learning architecture and back-propagation mechanism. In detail, DDL3 is inferior to PCA+BM3D due to the lack of fine-tuning, and DDL1+FT is comparable with or inferior to BM4D. However, by taking advantages of multi-layer dictionary learning and fine-tuning, DDL3+FT exceed other approaches, including PCA+BM3D who generates the reference for DDL3+FT. In other words, the improvements cannot be achieved if only one of the above two factors is considered, hierarchical dictionary learning or fine-tuning. The advantages of DDL3+FT can be validated visually by Fig. 2, where the results by different approaches are shown. Noting that rPCA+saBM3D outperforms best at $\sigma = 0.1$ due to the shape-adaptive neighborhoods employed in saBM3D, however, its performances reduce rapidly with the increasing σ even with the help of robust PCA, the underlying reason lies in the low confidence of the shape-adaptive neighborhoods extracted from the noisy components.

For the real data, as can be informed from Table 3, DDL3+FT still performs best with respect to image quality improvement and spectral structure preservation. Similar to the synthetic data, DDL3 performs worst in two metrics. Take Office dataset as an example, $\overline{Q_{bqi}}$ is 0.45 and Q_{sam} 0.29, both metrics are inferior to other approaches, which can be verified by Fig. 3. The underlying reason lies in the fact that dictionary learning is the biased feature leaning procedure, and the bias will be accumulated by the multi-layer configuration if no reliable supervised information can be utilized. In contrast, directed by the reference information and back-propagation mechanism, DDL1+FT is superior to DDL3 and even competitive to PCA+BM3D. This comparison demonstrates the importance of (weakly) supervised information and back-propagation mechanism in improving the denoising performance. Nevertheless, as illustrated by the performances on Pavia Center dataset, the improvement by single-layer dictionary learning equipped with fine-tuning is limited due to its inability in revealing the complex relations between the features spaces before and after dimension reduction. For the similar reason, DDL3+FT outperforms PCA+BM3D and BM4D. In short, the combination of deep dictionary learning and fine-tuning is very important for hyperspectral image denoising. From Fig. 3 and Fig. 4, the above remarks can be validated in detail.

4. Conclusion

Hyperspectral image denoising is challenging due to the complex noises and the limitation of the traditional single-layer strategy in discovering the complex relationship hidden in the dimension reduction procedure. A novel denoising approach is proposed for hyperspectral images based on the deep dictionary learning architecture and the fine-tuning mechanism. The former is helpful in revealing the complex hidden factors and improving the image quality, and the latter plays an important role in preserving the spectral structure. The future developments are mainly related to the extension of newly proposed denoising approaches into the deep dictionary learning framework.

References

[1] G. Chen and S.-E. Qian, "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage," IEEE Trans. Geosci. Remote Sens., vol.49, no.3, pp.973–980, 2011.

[2] G. Chen, S.-E. Qian, and S. Gleason, "Denoising of hyperspectral imagery by combining pca with block-matching 3-d filtering," Canadian J. Remote Sensing, vol.37, no.6, pp.590–595, 2012.

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

[4] A. Lam, I. Sato, and Y. Sato, "Denoising hyperspectral images using spectral domain statistics," ICPR, 2012.

[5] D. Cerra, R. Muller, and P. Reinartz, "Noise reduction in hyperspectral images through spectral unmixing," IEEE Geosci. Remote Sens. Lett., vol.11, no.1, pp.109–113, 2014.

[6] C. Ding, T. Li, and M.I. Jordan, "Convex and semi-nonnegative matrix factorizations," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.1, pp.45–55, 2010.

[7] G. Camps-Valls, L. Gomez-Chov, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," IEEE Geosci. Remote Sens. Lett., vol.3, no.1, pp.93–97, 2006.

[8] R. Clark, G. Swayze, and R. Wise, "Usgs digital spectral library splib06a." Online, 2007.

[9] "Rosis." http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_scene

[10] "Soc710." http://surfaceoptics.com/products/hyperspectral-imaging/soc710-portable-hyperspectral-camera/.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Bm3d image denoising with shape-adaptive principal component analysis," SPARS, 2009.

[12] H. Xu, C. Caramanis, and S. Mannor, "Outlier-robust pca: the high-dimensional case," IEEE Trans. Inf. Theory, vol.59, no.1, pp.546–572, 2013.

[13] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi, "Nonlocal transform domain filter for volumetric data denoising and reconstruction," IEEE Trans. Image Process., vol.22, no.1, pp.119–133, 2013.

[14] X. Kong, K. Li, Q. Yang, L. Wenyin, and M.-H. Yang, "A new image quality metric for image auto-denoising," ICCV, pp.2888–2895, 2013.

[15] R.H. Yuhas, J.W. Boardman, and A.F. Goetz, "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," Summaries of the 4th Annual JPL Airborne Geoscience Workshop, 1993.