Hyperspectral and Multispectral Image Fusion based on a Sparse Representation

Qi Wei, *Student Member, IEEE*, José Bioucas-Dias, *Senior Member, IEEE*, Nicolas Dobigeon, *Senior Member, IEEE*, and Jean-Yves Tourneret, *Senior Member, IEEE*

TECHNICAL REPORT – 2014, October

University of Toulouse, IRIT/INP-ENSEEIHT

2 rue Camichel, BP 7122, 31071 Toulouse cedex 7, France

Abstract

This paper presents a variational based approach for fusing hyperspectral and multispectral images. The fusion problem is formulated as an inverse problem whose solution is the target image assumed to live in a much lower dimensional subspace. A sparse regularization term is carefully designed, relying on a decomposition of the scene on a set of dictionaries. The dictionary atoms and the supports of the corresponding active coding coefficients are learned from the observed images. Then, conditionally on these dictionaries and supports, the fusion problem is solved via alternating optimization with respect to the target image (using the alternating direction method of multipliers) and the coding coefficients. Simulation results demonstrate the efficiency of the proposed algorithm when compared with state-of-the-art fusion methods.

Index Terms

Image fusion, hyperspectral image, multispectral image, sparse representation, dictionary, alternating direction method of multipliers (ADMM)

I. INTRODUCTION

Fusion of multi-sensor images has been explored during recent years and is still a very active research area [2]. A popular fusion problem in remote sensing consists of merging a high spatial resolution panchromatic (PAN) image and a low spatial resolution multispectral (MS) image. Many solutions have been proposed in the literature to solve this problem, known as *pansharpening* [2]–[5]. More recently, hyperspectral (HS) imaging acquiring a scene in several hundreds of contiguous spectral bands has opened a new range of relevant applications such as target detection [6] and spectral unmixing [7]. However, while HS sensors provide abundant spectral information, their spatial resolution is generally more limited [8]. To obtain images with good spectral and spatial resolutions, the remote sensing community has been devoting increasing research efforts to the problem of fusing HS with MS or PAN images [9], [10]. From an application point of view, this problem is also important as motivated by recent national programs, e.g., the Japanese next-generation space-borne hyperspectral image suite (HISUI), which fuses co-registered MS and HS images acquired over the same scene under the same conditions [11].

The fusion of HS and MS differs from pansharpening since both spatial and spectral information is contained in multi-band images. Therefore, a lot of pansharpening methods, such as component substitution [12] and relative spectral contribution [13] are inapplicable or inefficient for the fusion of HS and MS images. Since the fusion problem is generally ill-posed, Bayesian inference offers a convenient way to regularize the problem by defining an appropriate generic prior for the scene of interest. Following this strategy, Gaussian or ℓ_2-norm priors have been considered to build various estimators, in the image domain [14]–[16] or in a transformed domain [17]. Recently, the fusion of HS and MS images based on spectral unmixing has also been explored [18], [19].

Besides, sparse representations have received a considerable interest in recent years exploiting the self-similarity properties of natural images [20]–[23]. Using this property, a sparse constraint has been proposed in [24], [25] to regularize various ill-posed super-resolution and/or fusion problems. The linear decomposition of an image using a few atoms of a redundant dictionary

learned from this image (instead of a predefined dictionary, e.g., of wavelets) has recently been used for several problems related to low-level image processing tasks such as denoising [26] and classification [27], demonstrating the ability of sparse representations to model natural images. Learning a dictionary from the image of interest is commonly referred to as dictionary learning (DL). Liu *et al.* recently proposed to solve the pansharpening problem based on DL [5]. DL has also been investigated to restore HS images [28]. More precisely, a Bayesian scheme was introduced in [28] to learn a dictionary from an HS image, which imposes a self-consistency of the dictionary by using Beta-Bernoulli processes. This method provided interesting results at the price of a high computational complexity. Fusing multiple images using a sparse regularization based on the decomposition of these images into high and low frequency components was considered in [25]. However, the method developed in [25] required a training dataset to learn the dictionaries. The references mentioned before proposed to solve the corresponding sparse coding problem either by using greedy algorithms such as matching pursuit (MP) and orthogonal MP [29] or by relaxing the ℓ_0-norm to an ℓ_1-norm to take advantage of the last absolute shrinkage and selection operator (LASSO) [30].

In this paper, we propose to fuse HS and MS images within a constrained optimization framework, by incorporating a sparse regularization using dictionaries learned from the observed images. Knowing the trained dictionaries and the corresponding supports of the codes circumvents the difficulties inherent to the sparse coding step. The optimization problem can then be solved by optimizing alternatively with respect to (w.r.t.) the projected target image and the sparse code. The optimization w.r.t. the image is achieved by the split augmented Lagrangian shrinkage algorithm (SALSA) [31], which is an instance of the alternating direction method of multipliers (ADMM). By a suitable choice of variable splittings, SALSA enables a huge non-diagonalizable quadratic problem to be decomposed into a sequence of convolutions and pixel decoupled problems, which can be solved efficiently. The coding step is performed using a standard least-square (LS) algorithm which is possible because the supports have been fixed a priori.

The paper is organized as follows. Section II formulates the fusion problem within a con-

strained optimization framework. Section III presents the proposed sparse regularization and the method used to learn the dictionaries and the code support. The strategy investigated to solve the resulting optimization problem is detailed in Section IV. Simulation results are presented in Section V whereas conclusions are reported in Section VI.

II. PROBLEM FORMULATION

A. Notations and observation model

This paper considers the fusion of HS and MS images. The HS image is supposed to be a blurred and down-sampled version of the target image whereas the MS image is a spectrally degraded version of the target image. Both images are contaminated by white Gaussian noises. Instead of resorting to the totally vectorized notations used in [14], [15], [17], the HS and MS images are reshaped band-by-band to build $m_\lambda \times m$ and $n_\lambda \times n$ matrices, respectively, where m_λ is the number of HS bands, $n_\lambda < m_\lambda$ is the number of MS bands, n is the number of pixels in each band of the MS image and m is the number of pixels in each band of the HS image. The resulting observation models associated with the HS and MS images can be written as follows [14], [32], [33]

$$\mathbf{Y}_H = \mathbf{XBS} + \mathbf{N}_H$$
$$\mathbf{Y}_M = \mathbf{RX} + \mathbf{N}_M$$

(1)

where

- $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m_\lambda \times n}$ is the full resolution target image with m_λ bands and n pixels,
- $\mathbf{Y}_H \in \mathbb{R}^{m_\lambda \times m}$ and $\mathbf{Y}_M \in \mathbb{R}^{n_\lambda \times n}$ are the observed HS and MS images,
- $\mathbf{B} \in \mathbb{R}^{n \times n}$ is a cyclic convolution operator acting on the bands,
- $\mathbf{S} \in \mathbb{R}^{n \times m}$ is a down-sampling matrix (with down-sampling factor denoted as d),
- $\mathbf{R} \in \mathbb{R}^{n_\lambda \times m_\lambda}$ is the spectral response of the MS sensor,
- $\mathbf{N}_H$ and $\mathbf{N}_M$ are the HS and MS noises.

Note that $\mathbf{B}$ is a sparse symmetric Toeplitz matrix for a symmetric convolution kernel and $m = n/d^2$, where d is an integer standing for the downsampling factor. In this work, each

column of the noise matrices $\mathbf{N}_\mathrm{H} = [\mathbf{n}_{\mathrm{H},1}, \ldots \mathbf{n}_{\mathrm{H},m}]$ and $\mathbf{N}_\mathrm{M} = [\mathbf{n}_{\mathrm{M},1}, \ldots \mathbf{n}_{\mathrm{M},n}]$ is assumed to be a band-dependent Gaussian noise vector, i.e., $\mathbf{n}_{\mathrm{H},i} \sim \mathcal{N}\left(\mathbf{0}_{m_\lambda}, \mathbf{\Lambda}_\mathrm{H}\right) (i = 1, \ldots, m)$ and $\mathbf{n}_{\mathrm{M},i} \sim \mathcal{N}\left(\mathbf{0}_{n_\lambda}, \mathbf{\Lambda}_\mathrm{M}\right) (i = 1, \ldots, n)$ where $\mathbf{0}_a$ is the $a \times 1$ vector of zeros, $\mathbf{\Lambda}_\mathrm{H} = \mathrm{diag}(s_{\mathrm{H},1}^2, \ldots, s_{\mathrm{H},m_\lambda}^2) \in \mathbb{R}^{m_\lambda \times m_\lambda}$ and $\mathbf{\Lambda}_\mathrm{M} = \mathrm{diag}(s_{\mathrm{M},1}^2, \ldots, s_{\mathrm{M},n_\lambda}^2) \in \mathbb{R}^{n_\lambda \times n_\lambda}$ are diagonal matrices. Note that the Gaussian noise assumption used in this paper is quite popular in image processing [34]–[36] as it facilitates the formulation of the likelihood and the associated optimization algorithms. By denoting the Frobenius norm as $\|.\|_F$, the signal to noise ratios (SNRs) of each band in the two images (expressed in decibels) are defined as

$$\mathrm{SNR}_{\mathrm{H},i} = 10 \log \left(\frac{\|(\mathbf{XBS})_i\|_F^2}{s_{\mathrm{H},i}^2} \right), i = 1, \ldots, m_\lambda$$
$$\mathrm{SNR}_{\mathrm{M},j} = 10 \log \left(\frac{\|(\mathbf{RX})_j\|_F^2}{s_{\mathrm{M},j}^2} \right), j = 1, \ldots, n_\lambda.$$

B. Subspace learning

The unknown image is $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ where $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m_\lambda}]^T$ is the $m_\lambda \times 1$ vector corresponding to the ith spatial location (with $i = 1, \ldots, n$). As the bands of the HS data are generally spectrally dependent, the HS vector $\boldsymbol{x}_i$ usually lives in a subspace whose dimension is much smaller than the number of bands m_λ [37], [38], i.e.,

$$\boldsymbol{x}_i = \mathbf{H} \boldsymbol{u}_i \tag{2}$$

where $\boldsymbol{u}_i$ is the projection of the vector $\boldsymbol{x}_i$ onto the subspace spanned by the columns of $\mathbf{H} \in \mathbb{R}^{m_\lambda \times \widetilde{m}_\lambda}$ ($\mathbf{H}$ is an orthogonal matrix such that $\mathbf{H}^T \mathbf{H} = \mathbf{I}_{\widetilde{m}_\lambda}$). Using the notation $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$, we have $\mathbf{X} = \mathbf{HU}$ where $\mathbf{U} \in \mathbb{R}^{\widetilde{m}_\lambda \times n}$. Moreover, $\mathbf{U} = \mathbf{H}^T \mathbf{X}$ since $\mathbf{H}$ is an orthogonal matrix. In this case, the fusion problem (1) can be reformulated as estimating the unknown matrix $\mathbf{U}$ from the following observation equations

$$\mathbf{Y}_\mathrm{H} = \mathbf{HUBS} + \mathbf{N}_\mathrm{H}$$
$$\mathbf{Y}_\mathrm{M} = \mathbf{RHU} + \mathbf{N}_\mathrm{M}. \tag{3}$$

The dimension of the subspace $\widetilde{m}_\lambda$ is generally much smaller than the number of HS bands, i.e., $\widetilde{m}_\lambda \ll m_\lambda$. As a consequence, inferring in the subspace $\mathbb{R}^{\widetilde{m}_\lambda \times 1}$ greatly decreases the computational burden of the fusion algorithm. Another motivation for working in the subspace associated with $\mathbf{U}$ is to bypass the possible matrix singularity caused by the spectral dependency of the HS data. Note that each column of the orthogonal matrix $\mathbf{H}$ can be interpreted as a basis of the subspace of interest. In this paper, the subspace transform matrix $\mathbf{H}$ has been determined from a principal component analysis (PCA) of the HS data $\mathbf{Y}_H = [\mathbf{y}_{H,1}, \ldots, \mathbf{y}_{H,m}]$ (see step 7 of Algorithm 1). Note that instead of modifying the principal components directly as in the substitution based method [39], [40], the PCA is only employed to learn the subspace where the fusion problem is solved.

III. PROPOSED FUSION RULE FOR MULTISPECTRAL AND HYPERSPECTRAL IMAGES

A. Ill-posed inverse problem

As shown in (3), recovering the projected high-spectral and high-spatial resolution image $\mathbf{U}$ from the observations $\mathbf{Y}_H$ and $\mathbf{Y}_M$ is a linear inverse problem (LIP) [31]. In most single-image restoration problems (using either $\mathbf{Y}_H$ or $\mathbf{Y}_M$), the inverse problem is ill-posed or under-constrained [24], which requires regularization or prior information (in the Bayesian terminology). However, for multi-source image fusion, the inverse problem can be ill-posed or well-posed, depending on the dimension of the subspace and the number of spectral bands. If the matrix $\mathbf{RH}$ has full column rank and is well conditioned, which is seldom the case, the estimation of $\mathbf{U}$ according to (3) is an over-determined problem instead of an under-determined problem [41]. In this case, it is redundant to introduce regularizations. Conversely, if there are fewer MS bands than the subspace dimension $\widetilde{m}_\lambda$ (e.g., the MS image degrades to a PAN image), the matrix $\mathbf{RH}$ cannot have full column rank, which means that the fusion problem is an ill-posed LIP. In this paper, we focus on the under-determined case. Note however that the over-determined problem can be viewed as a special case with a regularization term set to zero. Another motivation for studying the under-determined problem is that it includes an archetypal fusion task referred to as pansharpening [2].

Based on the model (3) and the noise assumption, the distributions of $\mathbf{Y}_{\mathrm{H}}$ and $\mathbf{Y}_{\mathrm{M}}$ are

$$
\begin{aligned}
\mathbf{Y}_{\mathrm{H}}|\mathbf{U} &\sim \mathcal{MN}_{m_\lambda,m}(\mathbf{HUBS}, \boldsymbol{\Lambda}_{\mathrm{H}}, \mathbf{I}_m), \\
\mathbf{Y}_{\mathrm{M}}|\mathbf{U} &\sim \mathcal{MN}_{n_\lambda,n}(\mathbf{RHU}, \boldsymbol{\Lambda}_{\mathrm{M}}, \mathbf{I}_n).
\end{aligned}
\tag{4}
$$

where $\mathcal{MN}$ represents the matrix normal distribution. The probability density function of a matrix normal distribution $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c)$ is defined by

$$
p(\mathbf{X}|\mathbf{M}, \boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M})^T\boldsymbol{\Sigma}_r^{-1}(\mathbf{X}-\mathbf{M})\right]\right)}{(2\pi)^{np/2}|\boldsymbol{\Sigma}_c|^{n/2}|\boldsymbol{\Sigma}_r|^{p/2}}
$$

where $\mathbf{M}$ denotes the mean and $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_c$ are two matrices denoting row and column covariance matrices.

According to Bayes' theorem and using the fact that the noises $\mathbf{N}_H$ and $\mathbf{N}_M$ are independent, the posterior distribution of $\mathbf{U}$ can be written as

$$
p\left(\mathbf{U}|\mathbf{Y}_{\mathrm{H}}, \mathbf{Y}_{\mathrm{M}}\right) \propto p\left(\mathbf{Y}_{\mathrm{H}}|\mathbf{U}\right) p\left(\mathbf{Y}_{\mathrm{M}}|\mathbf{U}\right) p\left(\mathbf{U}\right).
\tag{5}
$$

In this work, we compute the maximum a posteriori (MAP) estimator using an optimization framework to solve the fusion problem. Taking the negative logarithm of the posterior distribution, maximizing the posterior distribution is equivalent to solving the following minimization problem

$$
\min_{\mathbf{U}} \underbrace{\frac{1}{2}\left\|\boldsymbol{\Lambda}_{\mathrm{H}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{H}} - \mathbf{HUBS})\right\|_F^2}_{\substack{\text{HS data term} \\ \propto \ln p(\mathbf{Y}_{\mathrm{H}}|\mathbf{U})}} + \underbrace{\frac{1}{2}\left\|\boldsymbol{\Lambda}_{\mathrm{M}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{M}} - \mathbf{RHU})\right\|_F^2}_{\substack{\text{MS data term} \\ \propto \ln p(\mathbf{Y}_{\mathrm{M}}|\mathbf{U})}}
$$
$$
+ \underbrace{\lambda\phi(\mathbf{U})}_{\substack{\text{regularizer} \\ \propto \ln p(\mathbf{U})}}
\tag{6}
$$

where the two first terms are associated with the MS and HS images (data fidelity terms) and the last term is a penalty ensuring appropriate regularization. Note that λ is a parameter adjusting the importance of regularization w.r.t. the data fidelity terms. It is also noteworthy that the MAP estimator is equivalent to the minimum mean square error (MMSE) estimator when $\phi(\mathbf{U})$ has a quadratic form, which is the case in our approach detailed below.

B. Sparse Regularization

Based on the self-similarity property of natural images, modeling image patches with a sparse representation has been shown to be very effective in many signal processing applications [24], [42], [43]. Instead of incorporating a simple Gaussian prior or smooth regularization for the fusion of HS and MS images [14], [16], [17], a sparse representation is introduced to regularize the fusion problem. More specifically, image patches of the target image projected into a subspace are represented as a sparse linear combination of elements from an appropriately chosen over-complete dictionary with columns referred to as atoms. In this paper, the atoms of the dictionary are tuned to the input images, leading to much better results than predefined dictionaries. More specifically, the goal of sparse regularization is to represent the patches of the target image as a weighted linear combination of a few elementary basis vectors or atoms, chosen from a learned over-complete dictionary. The proposed sparse regularization is defined as

$$\phi(\mathbf{U}) = \frac{1}{2} \sum_{i=1}^{\widetilde{m}_\lambda} \left\| \mathbf{U}_i - \mathcal{P}\left(\bar{\mathbf{D}}_i \bar{\mathbf{A}}_i\right) \right\|_F^2 \tag{7}$$

where

- $\mathbf{U}_i \in \mathbb{R}^n$ is the ith band (or row) of $\mathbf{U} \in \mathbb{R}^{\widetilde{m}_\lambda \times n}$, with $i = 1, \ldots, \widetilde{m}_\lambda$,
- $\mathcal{P}(\cdot) : \mathbb{R}^{n_p \times n_{pat}} \mapsto \mathbb{R}^{n \times 1}$ is a linear operator that averages the overlapping patches[1] of each band,
- $\bar{\mathbf{D}}_i \in \mathbb{R}^{n_p \times n_{at}}$ is an overcomplete dictionary whose columns are basis elements of size n_p (corresponding to the size of a patch),
- $\bar{\mathbf{A}}_i \in \mathbb{R}^{n_{at} \times n_{pat}}$ is the ith band code (n_{at} is the number of atoms and n_{pat} is the number of patches associated with the ith band).

Note that there are $\widetilde{m}_\lambda$ vectors $\mathbf{U}_i \in \mathbb{R}^n$ since the dimension of the hyperspectral subspace in which the observed vectors $\boldsymbol{x}_i$ have been projected is $\widetilde{m}_\lambda$. The operation decomposing each band into overlapping patches of size $\sqrt{n_p} \times \sqrt{n_p}$ is denoted as $\mathcal{P}^*(\cdot) : \mathbb{R}^{n \times 1} \mapsto \mathbb{R}^{n_p \times n_{pat}}$, which is the adjoint operation of $\mathcal{P}(\cdot)$, i.e., $\mathcal{P}\left[\mathcal{P}^*(\mathbf{X})\right] = \mathbf{X}$.

---

[1]Note that the overlapping decomposition adopted here is to prevent block artifacts [44].

C. Dictionary learning

The DL strategy advocated in this paper consists of learning the dictionaries $\bar{\mathbf{D}}_i$ and an associated sparse code $\bar{\mathbf{A}}_i$ for each band of a rough estimation of $\mathbf{U}$ using the observed HS and MS images. A rough estimation of $\mathbf{U}$, referred as $\tilde{\mathbf{U}}$ is constructed using the MS image $\mathbf{Y}_{\mathrm{M}}$ and the HS image $\mathbf{Y}_{\mathrm{H}}$, following the strategy initially studied in [14]. Note that other estimation methods might also be used to compute a rough estimation of $\mathbf{U}$ (see step 1 in Algorithm 1). Then each band $\tilde{\mathbf{U}}_i$ of $\tilde{\mathbf{U}}$ is decomposed into n_{pat} overlapping patches of size $\sqrt{n_{\mathrm{p}}} \times \sqrt{n_{\mathrm{p}}}$ forming a patch matrix $\mathcal{P}^*(\tilde{\mathbf{U}}_i) \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{pat}}}$.

Many DL methods have been studied in the recent literature. These methods are for instance based on K-SVD [45], online dictionary learning (ODL) [22] or Bayesian learning [28]. In this study, we propose to learn the set $\bar{\mathbf{D}} \triangleq [\bar{\mathbf{D}}_1, \ldots, \bar{\mathbf{D}}_{\tilde{m}_\lambda}]$ of over-complete dictionaries using ODL since it is effective from the computational point of view and has empirically demonstrated to provide more relevant representations. More specifically, the dictionary $\mathbf{D}_i$ associated with the band $\mathbf{U}_i$ is trained by solving the following optimization problem (see step 3 in Algorithm 1).

$$\left\{\bar{\mathbf{D}}_i, \tilde{\mathbf{A}}_i\right\} = \operatorname*{argmin}_{\mathbf{D}_i, \mathbf{A}_i} \frac{1}{2} \left[\left\| \mathcal{P}^*(\tilde{\mathbf{U}}_i) - \mathbf{D}_i \mathbf{A}_i \right\|_F^2 + \mu \left\| \mathbf{A}_i \right\|_1 \right]. \tag{8}$$

Then, to provide a more compact representation, we propose to re-estimate the sparse code

$$\bar{\mathbf{A}}_i = \operatorname*{argmin}_{\mathbf{A}_i} \frac{1}{2} \left\| \mathcal{P}^*(\tilde{\mathbf{U}}_i) - \bar{\mathbf{D}}_i \mathbf{A}_i \right\|_F^2, \text{ s.t. } \left\| \mathbf{A}_i \right\|_0 \leq K \tag{9}$$

where K is a given maximum number of atoms, for each patch of $\mathbf{U}_i$. This ℓ_0-norm constrained regression problem can be addressed using greedy algorithms, e.g., orthogonal matching pursuit (OMP). Generally, the maximum number of atoms K is set much smaller than the number of atoms in the dictionary, i.e., $K \ll n_{\mathrm{at}}$. The positions of the non-zero elements of the code $\bar{\mathbf{A}}_i$, namely the supports denoted as $\bar{\mathbf{\Omega}}_i \triangleq \left\{ (j,k) | \bar{\mathbf{A}}_i(j,k) \neq 0 \right\}$, are also identified (see steps 4 and 5 in Algorithm 1).

D. Including the sparse code into the estimation framework

Since the regularization term (7) exhibits separable terms w.r.t. each image $\mathbf{U}_i$ in band $\#i$, it can be easily interpreted in a Bayesian framework as the joint prior distribution of the images $\mathbf{U}_i$ ($i = 1, \ldots, \widetilde{m}_\lambda$) assumed to be a priori independent, where each marginal prior $p(\mathbf{U}_i)$ is a Gaussian distribution with mean $\mathcal{P}(\bar{\mathbf{D}}_i \bar{\mathbf{A}}_i)$. More formally, by denoting $\bar{\mathbf{A}} \triangleq [\bar{\mathbf{A}}_1, \ldots, \bar{\mathbf{A}}_{\widetilde{m}_\lambda}]$, the prior distribution for $\mathbf{U}$ associated with the regularization (7) can be written

$$p(\mathbf{U}|\bar{\mathbf{D}}, \bar{\mathbf{A}}) = \prod_{i=1}^{\widetilde{m}_\lambda} p(\mathbf{U}_i|\bar{\mathbf{D}}_i, \bar{\mathbf{A}}_i). \tag{10}$$

In a standard approach, the hyperparameters $\bar{\mathbf{D}}$ and $\bar{\mathbf{A}}$ can be a priori fixed, e.g., based on the dictionary learning step detailed in the previous section. However, this choice can drastically impact the accuracy of the representation and therefore the relevance of the regularization term. Inspired by hierarchical models frequently encountered in Bayesian inference [46], we propose to add a second level in the Bayesian paradigm by fixing the dictionaries $\bar{\mathbf{D}}$ and the set of supports $\bar{\mathbf{\Omega}} \triangleq \{\bar{\mathbf{\Omega}}_1, \ldots, \bar{\mathbf{\Omega}}_{\widetilde{m}_\lambda}\}$, but by including the code $\mathbf{A}$ within the estimation process. The associated joint prior can be written as follows

$$p(\mathbf{U}, \mathbf{A}|\bar{\mathbf{D}}, \bar{\mathbf{A}}) = \prod_{i=1}^{\widetilde{m}_\lambda} p(\mathbf{U}_i|\bar{\mathbf{D}}_i, \mathbf{A}_i) p(\mathbf{A}_i|\bar{\mathbf{A}}_i) \tag{11}$$

where $\bar{\mathbf{\Omega}}$ is derived from $\bar{\mathbf{A}}$. Therefore, the regularization term (7) reduces to

$$\phi(\mathbf{U}, \mathbf{A}) = \frac{1}{2} \sum_{i=1}^{\widetilde{m}_\lambda} \left\|\mathbf{U}_i - \mathcal{P}(\bar{\mathbf{D}}_i \mathbf{A}_i)\right\|_F^2 = \frac{1}{2}\left\|\mathbf{U} - \bar{\mathbf{U}}\right\|_F^2,$$
$$\text{s.t. } \left\{\mathbf{A}_{i, \backslash \bar{\mathbf{\Omega}}_i} = 0\right\}_{i=1}^{\widetilde{m}_\lambda} \tag{12}$$

where $\bar{\mathbf{U}} \triangleq [\mathcal{P}(\bar{\mathbf{D}}_1 \mathbf{A}_1), \ldots, \mathcal{P}(\bar{\mathbf{D}}_{\widetilde{m}_\lambda} \mathbf{A}_{\widetilde{m}_\lambda})]$ and $\mathbf{A}_{i, \backslash \bar{\mathbf{\Omega}}_i} = \{\mathbf{A}_i(j, k) \mid (j, k) \notin \bar{\mathbf{\Omega}}_i\}$. It is worthy to note that i) the regularization term in (12) is still separable w.r.t. each band $\mathbf{U}_i$, and ii) the optimization of (12) w.r.t. $\mathbf{A}_i$ reduces to an ℓ_2-norm optimization task w.r.t. the non-zero elements in $\mathbf{A}_i$, which can be solved easily. The hierarchical structure of the observed data, parameters and hyperparameters is summarized in Fig. 1.

Finally, substituting (12) into (6), the optimization problem to be solved can be expressed as

follows

$$\min_{\mathbf{U},\mathbf{A}} L(\mathbf{U}, \mathbf{A}) \triangleq \tfrac{1}{2}\big\|\boldsymbol{\Lambda}_{\mathrm{H}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{H}} - \mathbf{HUBS})\big\|_F^2 +$$

$$\tfrac{1}{2}\big\|\boldsymbol{\Lambda}_{\mathrm{M}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{M}} - \mathbf{RHU})\big\|_F^2 + \tfrac{\lambda}{2}\big\|\mathbf{U} - \bar{\mathbf{U}}\big\|_F^2, \tag{13}$$

$$\text{s.t. } \big\{\mathbf{A}_{i,\backslash\bar{\boldsymbol{\Omega}}_i} = 0\big\}_{i=1}^{\widetilde{m}_\lambda}.$$

Note that the set of constraints $\big\{\mathbf{A}_{i,\backslash\bar{\boldsymbol{\Omega}}_i} = 0\big\}_{i=1}^{\widetilde{m}_\lambda}$ could have been removed. In this case, to ensure a sparse representation of $\mathbf{U}_i$ ($i = 1, \dots, \widetilde{m}_\lambda$), sparse constraints on the codes $\mathbf{A}_i$ ($i = 1, \dots, \widetilde{m}_\lambda$), such as $\{\|\mathbf{A}_i\|_0 < K\}_{i=1}^{\widetilde{m}_\lambda}$ or sparsity promoting penalties, e.g., $\sum_{i=1}^{\widetilde{m}_\lambda} \|\mathbf{A}_i\|_1$ should have been included into the object function (13). This would have resulted in a much more computationally intensive algorithm.

IV. ALTERNATE OPTIMIZATION

Once $\bar{\mathbf{D}}$, $\bar{\boldsymbol{\Omega}}$ and $\mathbf{H}$ have been learned from the observed data, (13) reduces to a standard constrained quadratic optimization problem w.r.t. $\mathbf{U}$ and $\mathbf{A}$. However, this problem is difficult to solve due to its large dimension and the fact that the operators $\mathbf{H}(\cdot)\mathbf{BD}$ and $\mathcal{P}(\cdot)$ cannot be easily diagonalized. To cope with this difficulty, we propose an optimization technique that alternates optimization w.r.t. $\mathbf{U}$ and $\mathbf{A}$, which is a simple version of a block coordinate descent algorithm.

The optimization w.r.t. $\mathbf{U}$ conditional on $\mathbf{A}$ (or equivalent on $\bar{\mathbf{U}}$) can be achieved efficiently with the ADMM [47] whose convergence has been proved in the convex case. The optimization w.r.t. $\mathbf{A}$ with the support constraint $\mathbf{A}_{i,\backslash\bar{\boldsymbol{\Omega}}_i} = 0$ ($i = 1, 2, \dots, \widetilde{m}_\lambda$) conditional on $\mathbf{U}$ is a least squares (LS) regression problem for the non-zero elements of $\mathbf{A}$, which can be solved easily. The resulting scheme including learning $\bar{\mathbf{D}}$, $\bar{\boldsymbol{\Omega}}$ and $\mathbf{H}$ is detailed in Algo. 1. The alternating ADMM and LS steps are detailed in what follows.

Note that the objective function is convex w.r.t $\mathbf{U}$ and $\mathbf{A}$ separately, although being non-convex with respect to the pair $(\mathbf{U}, \mathbf{A})$. In practice, a very simple way to ensure convergence is to add the quadratic terms $\mu_a\|A\|_F^2$, with very small μ_a. In this case, the solution of Algorithm 1 is unique and the ADMM algorithm converges linearly [48]. In practice, we have noticed that even when μ_a is zero, the solution of Algorithm 1 always converges to a unique point.

Algorithm 1: Fusion of HS and MS based on a sparse representation

---

    Input: $\mathbf{Y}_{\mathrm{H}}$, $\mathbf{Y}_{\mathrm{M}}$, $\mathrm{SNR}_{\mathrm{H}}$, $\mathrm{SNR}_{\mathrm{M}}$, $\widetilde{m}_\lambda$, $\mathbf{R}$, $\mathbf{B}$, $\mathbf{S}$, K

    % *Propose a rough estimation $\tilde{\mathbf{U}}$ of $\mathbf{U}$*

1 Compute $\tilde{\mathbf{U}} \triangleq \hat{\boldsymbol{\mu}}_{\mathbf{U}|\mathbf{Y}_{\mathrm{M}}}$ following the method in [14] ;

2 **for** $i = 1$ **to** $\widetilde{m}_\lambda$ **do**

        % *Online dictionary learning*

3 $\bar{\mathbf{D}}_i \leftarrow \mathsf{ODL}(\tilde{\mathbf{U}}_i)$;

        % *Sparse coding*

4 $\bar{\mathbf{A}}_i \leftarrow \mathsf{OMP}(\bar{\mathbf{D}}_i, \tilde{\mathbf{U}}_i, K)$;

        % *Identify the supports*

5   Set $\bar{\mathbf{\Omega}}_i = \left\{(j,k) | \bar{\mathbf{A}}_i(j,k) \neq 0\right\}$;

6 **end**

    % *Identify the hyperspectral subspace*

7 $\hat{\mathbf{H}} \leftarrow \mathsf{PCA}(\mathbf{Y}_{\mathrm{H}}, \widetilde{m}_\lambda)$;

    % *Start alternate optimization*

8 **for** $t = 1, 2, \ldots$ **to** *stopping rule* **do**

        % *Optimize w.r.t. $\mathbf{U}$ using SALSA (see Algorithm 2)*

9 $\hat{\mathbf{U}}^{(t)} \in \{\mathbf{U}|\, L(\mathbf{U}, \hat{\mathbf{A}}^{(t-1)}) \leq L(\hat{\mathbf{U}}^{(t-1)}, \hat{\mathbf{A}}^{(t-1)})\}$;

        % *Optimize w.r.t. $\mathbf{A}$ (LS regression)*

10 $\hat{\mathbf{A}}^{(t)} \in \{\mathbf{A}|\, L(\hat{\mathbf{U}}^{(t)}, \mathbf{A}) \leq L(\hat{\mathbf{U}}^{(t)}, \hat{\mathbf{A}}^{(t-1)})\}$;

11 **end**

12 Set $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{U}}$;

    Output: $\hat{\mathbf{X}}$ (high resolution HS image)

A. ADMM Step

The function to be minimized w.r.t. $\mathbf{U}$ conditionally on $\mathbf{A}$ (or $\bar{\mathbf{U}}$) is

$$\tfrac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{H}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{H}} - \mathbf{HUBS})\right\|_F^2 + \tfrac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{M}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{M}} - \mathbf{RHU})\right\|_F^2 + \tag{14}$$

$$\tfrac{\lambda}{2}\left\|\mathbf{U} - \bar{\mathbf{U}}\right\|_F^2.$$

By introducing the splittings $\mathbf{V}_1 = \mathbf{UB}$, $\mathbf{V}_2 = \mathbf{U}$ and $\mathbf{V}_3 = \mathbf{U}$ and the respective scaled Lagrange multipliers $\mathbf{G}_1, \mathbf{G}_2$ and $\mathbf{G}_3$, the augmented Lagrangian associated with the optimization

of $\mathbf{U}$ can be written as

$$L(\mathbf{U}, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3) =$$
$$\frac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{H}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{H}} - \mathbf{H}\mathbf{V}_1\mathbf{S})\right\|_F^2 + \frac{\mu}{2}\left\|\mathbf{U}\mathbf{B} - \mathbf{V}_1 - \mathbf{G}_1\right\|_F^2 +$$
$$\frac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{M}}^{-\frac{1}{2}}(\mathbf{Y}_{\mathrm{M}} - \mathbf{R}\mathbf{H}\mathbf{V}_2)\right\|_F^2 + \frac{\mu}{2}\left\|\mathbf{U} - \mathbf{V}_2 - \mathbf{G}_2\right\|_F^2 +$$
$$\frac{1}{2}\left\|\bar{\mathbf{U}} - \mathbf{V}_3\right\|_F^2 + \frac{\mu}{2}\left\|\mathbf{U} - \mathbf{V}_3 - \mathbf{G}_3\right\|_F^2.$$

The updates of $\mathbf{U}, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{G}_1, \mathbf{G}_2$ and $\mathbf{G}_3$ are obtained with the SALSA algorithm [31], [49], which is an instance of the ADMM algorithm with guaranteed convergence. The SALSA scheme is summarized in Algorithm 2. Note that the optimization w.r.t. to $\mathbf{U}$ (step 5) can be efficiently solved in the Fourier domain.

B. Patchwise Sparse Coding

The optimization w.r.t. $\mathbf{A}$ conditional on $\mathbf{U}$ is

$$\hat{\mathbf{A}}_i = \arg\min_{\mathbf{A}_i} \left\|\mathbf{U}_i - \mathcal{P}(\bar{\mathbf{D}}_i\mathbf{A}_i)\right\|_F^2, \text{ s.t. } \mathbf{A}_{i,\backslash\bar{\mathbf{\Omega}}_i} = 0, \tag{15}$$

where $i = 1, \ldots, \widetilde{m}_\lambda$. Since the operator $\mathcal{P}(\cdot)$ is a linear mapping from patches to images and $\mathcal{P}\left[\mathcal{P}^*\left(\mathbf{X}\right)\right] = \mathbf{X}$, the problem (15) can be rewritten as

$$\hat{\mathbf{A}}_i = \arg\min_{\mathbf{A}_i} \left\|\mathcal{P}\left(\mathcal{P}^*(\mathbf{U}_i) - \bar{\mathbf{D}}_i\mathbf{A}_i\right)\right\|_F^2, \text{ s.t. } \mathbf{A}_{i,\backslash\bar{\mathbf{\Omega}}_i} = 0. \tag{16}$$

The solution of (16) can be approximated by solving

$$\hat{\mathbf{A}}_i = \arg\min_{\mathbf{A}_i} \left\|\mathcal{P}^*(\mathbf{U}_i) - \bar{\mathbf{D}}_i\mathbf{A}_i\right\|_F^2, \text{ s.t. } \mathbf{A}_{i,\backslash\bar{\mathbf{\Omega}}_i} = 0. \tag{17}$$

Note that using the sub-optimal solution instead of the optimal solution does not affect the convergence of the alternating optimization. Tackling the support constraint consists of only updating the non-zero elements of each column of $\mathbf{A}_i$. The jth vectorized column of $\mathcal{P}^*(\mathbf{U}_i)$ is denoted as $\mathbf{p}_{i,j}$, the vector composed of the K non-zero elements of the jth column of $\mathbf{A}_i$ is denoted as $\mathbf{a}_{\bar{\mathbf{\Omega}}_i^j}$, and the corresponding column of $\bar{\mathbf{D}}_i$ is denoted as $\bar{\mathbf{D}}_{\bar{\mathbf{\Omega}}_i^j}$. Then the $\widetilde{m}_\lambda$ problems

---

Algorithm 2: SALSA sub-iterations

Input: $\hat{\mathbf{U}}^{(t)}$, $\bar{\mathbf{D}}$, $\hat{\mathbf{A}}^{(t)}$, $\mathbf{Y}_{\mathrm{H}}$, $\mathbf{Y}_{\mathrm{M}}$, $\mathrm{SNR}_{\mathrm{H}}$, $\mathrm{SNR}_{\mathrm{M}}$, $\mathbf{H}$, $\mathbf{R}$, $\mathbf{B}$, $\mathbf{S}$, λ and μ (SALSA parameter)

1   Set $\bar{\mathbf{U}} = \left[\mathcal{P}\left(\bar{\mathbf{D}}_1\hat{\mathbf{A}}_1^{(t)}\right), \ldots, \mathcal{P}\left(\bar{\mathbf{D}}_{\widetilde{m}_\lambda}\hat{\mathbf{A}}_{\widetilde{m}_\lambda}^{(t)}\right)\right]$;

2   Set $\boldsymbol{\delta} \in \{0,1\}^n$ such that $\boldsymbol{\delta}(i) = \begin{cases} 1 & \text{if pixel } i \text{ is sampled,} \\ 0 & \text{otherwise;} \end{cases}$

3   **Initialization**: $\mathbf{V}_1^{(0)}, \mathbf{V}_2^{(0)}, \mathbf{V}_3^{(0)}$, $\mathbf{G}_1^{(0)}$, $\mathbf{G}_2^{(0)}, \mathbf{G}_3^{(0)}$;

4   **for** $k = 0$ **to** n_{it} **do**

     % *Optimize w.r.t* **U**

5      $\hat{\mathbf{U}}^{(t,k+1)} \leftarrow \left[(\mathbf{V}_1^{(k)} + \mathbf{G}_1^{(k)})\mathbf{B}^T + (\mathbf{V}_2^{(k)} + \mathbf{G}_2^{(k)}) + (\mathbf{V}_3^{(k)} + \mathbf{G}_3^{(k)})\right]$ $\left(\mathbf{B}\mathbf{B}^T + 2\mathbf{I}\right)^{-1}$;

     % *Optimize* **V**$_1$ *(according to down-sampling)*

6      $\boldsymbol{\nu}_1 \leftarrow (\hat{\mathbf{U}}^{(t,k+1)}\mathbf{B} - \mathbf{G}_1^{(k)})$;

7      $\mathbf{V}_1^{(k+1)}(:,\boldsymbol{\delta}) \leftarrow \left(\mathbf{H}^T\boldsymbol{\Lambda}_{\mathrm{H}}^{-1}\mathbf{H} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{H}^T\boldsymbol{\Lambda}_{\mathrm{H}}^{-1}\mathbf{Y}_{\mathrm{H}} + \boldsymbol{\nu}_1(:,\boldsymbol{\delta})\right)$;

8      $\mathbf{V}_1^{(k+1)}(:,1-\boldsymbol{\delta}) \leftarrow \boldsymbol{\nu}_1(:,1-\boldsymbol{\delta})$;

     % *Optimize w.r.t* **V**$_2$

9      $\boldsymbol{\nu}_2 \leftarrow (\hat{\mathbf{U}}^{(t,k+1)} - \mathbf{G}_2^{(k)})$;

10     $\mathbf{V}_2^{(k+1)} \leftarrow \left(\mathbf{H}^T\mathbf{R}^T\boldsymbol{\Lambda}_{\mathrm{M}}^{-1}\mathbf{R}\mathbf{H} + \mu\mathbf{I}\right)^{-1}\left(\mathbf{H}^T\mathbf{R}^T\boldsymbol{\Lambda}_{\mathrm{M}}^{-1}\mathbf{Y}_{\mathrm{M}} + \mu\boldsymbol{\nu}_2\right)$;

     % *Optimize w.r.t* **V**$_3$

11     $\boldsymbol{\nu}_3 \leftarrow (\hat{\mathbf{U}}^{(t,k+1)} - \mathbf{G}_3^{(k)})$;

12     $\mathbf{V}_3^{(k+1)} \leftarrow (\lambda + \mu)^{-1}\left(\lambda\bar{\mathbf{U}} + \mu\boldsymbol{\nu}_3\right)$;

     % *Update Lagrange multipliers*

13     $\mathbf{G}_1^{(k+1)} \leftarrow (-\boldsymbol{\nu}_1 + \mathbf{V}_1^{(k+1)})$;

14     $\mathbf{G}_2^{(k+1)} \leftarrow (-\boldsymbol{\nu}_2 + \mathbf{V}_2^{(k+1)})$;

15     $\mathbf{G}_3^{(k+1)} \leftarrow (-\boldsymbol{\nu}_3 + \mathbf{V}_3^{(k+1)})$;

16   **end**

17   Set $\hat{\mathbf{U}}^{(t+1)} = \hat{\mathbf{U}}^{(t,n_{\mathrm{it}})}$;

     Output: $\hat{\mathbf{U}}^{(t+1)}$

---

in (17) reduce to $\widetilde{m}_\lambda \times n_{\mathrm{pat}}$ sub-problems

$$\hat{\mathbf{a}}_{\bar{\boldsymbol{\Omega}}_i^j} = \arg\min_{\mathbf{a}_{\bar{\boldsymbol{\Omega}}_i^j}} \left\|\mathbf{p}_{i,j} - \bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}\mathbf{a}_{\bar{\boldsymbol{\Omega}}_i^j}\right\|_F^2, \tag{18}$$

whose solutions $\hat{\mathbf{a}}_{\bar{\boldsymbol{\Omega}}_i^j} = (\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}^T\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j})^{-1}\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}^T\mathbf{p}_{i,j}$ (for $i = 1, \ldots, \widetilde{m}_\lambda, j = 1, \ldots, n_{\mathrm{pat}}$) can be explicitly computed in parallel. The corresponding patch estimate is $\hat{\mathbf{p}}_{i,j} \triangleq \mathbf{T}_{i,j}\mathbf{p}_{i,j}$, with $\mathbf{T}_{i,j} = \bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}(\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}^T\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j})^{-1}\bar{\mathbf{D}}_{\bar{\boldsymbol{\Omega}}_i^j}^T$. These patches are used to build $\bar{\mathbf{U}}$ (i.e., equivalently, $\mathcal{P}\left(\bar{\mathbf{D}}_i\mathbf{A}_i\right)$) required

in the optimization w.r.t. $\mathbf{U}$ (see section IV-A). Note that $\mathbf{T}_{i,j}$ is a projection operator, and hence is symmetric ($\mathbf{T}_{i,j}^T = \mathbf{T}_{i,j}$) and idempotent ($\mathbf{T}_{i,j}^2 = \mathbf{T}_{i,j}$). Note also that $\mathbf{T}_{i,j}$ needs to be calculated only once, given the learned dictionaries and associated supports.

C. Complexity Analysis

The SALSA algorithm has a complexity of the order $\mathcal{O}\left(n_{\mathrm{it}} \widetilde{m}_\lambda n \log\left(\widetilde{m}_\lambda n\right)\right)$ [31], where n_{it} is the number of SALSA iterations. The computational complexity of the patchwise sparse coding is $\mathcal{O}\left(K n_{\mathrm{p}} n_{\mathrm{pat}} \widetilde{m}_\lambda\right)$. Conducting the fusion in a subspace of dimension $\widetilde{m}_\lambda$ instead of working with the initial space of dimension m_λ greatly decreases the complexity of both SALSA and sparse coding steps.

V. SIMULATION RESULTS ON SYNTHETIC DATA

This section studies the performance of the proposed sparse representation-based fusion algorithm. The reference image considered here as the high spectral and high spectral image is a $128 \times 128 \times 93$ HS image with spatial resolution of 1.3m acquired by the reflective optics system imaging spectrometer (ROSIS) optical sensor over the urban area of the University of Pavia, Italy. The flight was operated by the Deutsches Zentrum für Luft- und Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the European Union. This image was initially composed of 115 bands that have been reduced to 93 bands after removing the water vapor absorption bands (with spectral range from 0.43 to 0.86m). It has received a lot of attention in the remote sensing literature [50]–[52]. A composite color image, formed by selecting the red, green and blue bands of the reference image is shown in the left of Fig. 2.

A. Simulation Scenario

We propose to reconstruct the reference HS image from two lower resolved images. A high-spectral low-spatial resolution HS image has been constructed by applying a 5×5 exponential degrading spatial filter on each band of the reference image and down-sampling every 4 pixels in

both horizontal and vertical directions. In a second step, we have generated a 4-band MS image by filtering the reference image with the IKONOS-like reflectance spectral responses depicted in Fig. 3. The HS and MS images are both contaminated by zero-mean additive Gaussian noises. Our simulations have been conducted with $\text{SNR}_{1,\cdot} = 35\text{dB}$ for the first 43 bands and $\text{SNR}_{1,\cdot} = 30\text{dB}$ for the remaining 50 bands of the HS image. For the MS image, $\text{SNR}_{2,\cdot}$ is 30dB for all bands. The noise-contaminated HS and MS images are depicted in the middle and right of Fig. 2 (the HS image has been interpolated for better visualization).

B. Learning the Subspace, the Dictionaries and the Code Supports

1) Subspace: To learn the transform matrix $\mathbf{H}$, we used the principal component analysis (PCA) as in [16]. Note that PCA is a classical dimensionality reduction technique used in hyperspectral imagery. The empirical correlation matrix $\mathbf{\Upsilon} = \mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right]$ of the HS pixel vectors is diagonalized leading to

$$\mathbf{W}^T \mathbf{\Upsilon} \mathbf{W} = \mathbf{\Gamma} \tag{19}$$

where $\mathbf{W}$ is an $m_\lambda \times m_\lambda$ unitary matrix ($\mathbf{W}^T = \mathbf{W}^{-1}$) and $\mathbf{\Gamma}$ is a diagonal matrix whose diagonal elements are the ordered eigenvalues of $\mathbf{\Upsilon}$ denoted as $d_1 \geq d_2 \geq ... \geq d_{m_\lambda}$. The top $\widetilde{m}_\lambda$ components are selected and the matrix $\mathbf{H}$ is then constructed as the eigenvectors associated with the $\widetilde{m}_\lambda$ largest eigenvalues of $\mathbf{\Upsilon}$. In practice, the selection of the number of principal components $\widetilde{m}_\lambda$ depends on how many materials (or endmembers) the target image contains. If the number of truncated principal components is smaller than the dimension of the subspace spanned by the target image vectors, the projection will lead to a loss of information. On the contrary, if the number of principal components is larger than the real dimension, the over-fitting problem may arise leading to a degradation of the fusion performance. As an illustration, the eigenvalues of $\mathbf{\Upsilon}$ for the Pavia image are displayed in Fig. 4. For this example, the $\widetilde{m}_\lambda = 5$ eigenvectors contain 99.9% of the information and have been chosen to build the subspace of interest.

2) Dictionaries: As explained before, the target high resolution image is assumed to live in a lower dimensional subspace. Firstly, a rough estimation of the projected image is obtained

with the method proposed in [14]. In a second step, $\widetilde{m}_\lambda = 5$ dictionaries are learned from the rough estimation of the projected image using the ODL method. As $n_{\text{at}} \gg n_{\text{p}}$, the dictionary is over-complete. There is no unique rule to select the dictionary size n_{p} and the number of atoms n_{at}. However, two limiting cases can be identified

- The patch reduces to a single pixel, which means $n_{\text{p}} = 1$. In this case, the sparsity is not necessary to be introduced since only one 1D dictionary atom (which is a constant) is enough to represent any target patch.

- The patch is as large as the whole image, which means only one atom is needed to represent the image. In this case, the atom is too "specialized" to describe any other image.

More generally, the smaller the patches, the more objects the atoms can approximate. However, too small patches are not efficient to properly capture the textures, edges, etc. With larger patch size, a larger number of atoms are required to guarantee the over-completeness (which requires larger computation cost). In general, the size of patches is selected empirically. For the ODL algorithm used in this study, this size has been fixed to $n_{\text{p}} = 6 \times 6$ and the number of atoms is $n_{\text{at}} = 256$. The learned dictionaries for the first three bands of $\tilde{\mathbf{U}}$ are displayed in Fig. A. This figure shows that the spatial properties of the target image have been captured by the atoms of the dictionaries.

3) Code Supports: Based on the dictionaries learned following the strategy presented in Section V-B2, the codes are re-estimated by solving (9) with OMP. Note that the target sparsity K represents the maximum number of atoms used to represent one patch, which also determines the number of non-zeros elements of $\mathbf{A}$ estimated jointly with the projected image $\mathbf{U}$. If K is too large, the optimization w.r.t. $\mathbf{U}$ and $\mathbf{A}$ leads to over-fitting, which means there are too many parameters to estimate while the sample size is too small. The training supports for the first three bands are displayed in the right column of Fig. A. The number of rows is 256, which represents the number of atoms in each dictionary $\bar{\mathbf{D}}_i$ $(i = 1, \ldots, \widetilde{m}_\lambda)$. The white dots in the jth column indicate which atoms are used for reconstructing the jth patch $(j = 1, \ldots, n_{\text{pat}})$. The sparsity is clearly observed in this figure. Note that some atoms are frequently used whereas some others are not. The most popular atoms represent spatial details that are quite common in images. The

other atoms represent details that are characteristics of specific patches.

C. Fusion Quality Metrics

To evaluate the quality of the proposed fusion strategy, several image quality measures have been employed. Referring to [17], we propose to use RMSE, SAM, UIQI, ERGAS and DD as defined below.

1) RMSE: The root mean square error (RMSE) is a similarity measure between the target image $\mathbf{X}$ and fused image $\hat{\mathbf{X}}$ defined as

$$\text{RMSE}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{n m_\lambda} \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2.$$

The smaller RMSE, the better the fusion quality.

2) SAM: The spectral angle mapper (SAM) measures the spectral distortion between the actual and estimated images. The SAM of two spectral vectors $\boldsymbol{x}_n$ and $\hat{\boldsymbol{x}}_n$ is defined as

$$\text{SAM}(\boldsymbol{x}_n, \hat{\boldsymbol{x}}_n) = \arccos\left(\frac{\langle \boldsymbol{x}_n, \hat{\boldsymbol{x}}_n \rangle}{\|\boldsymbol{x}_n\|_2 \|\hat{\boldsymbol{x}}_n\|_2}\right).$$

The overall SAM is finally obtained by averaging the SAMs computed from all image pixels. Note that the value of SAM is expressed in degrees and thus belongs to $(-90, 90]$. The smaller the absolute value of SAM, the less important the spectral distortion.

3) UIQI: The universal image quality index (UIQI) was proposed in [53] for evaluating the similarity between two single band images. It is related to the correlation, luminance distortion and contrast distortion of the estimated image w.r.t. the reference image. The UIQI between two single-band images $\mathbf{a} = [a_1, a_2, \ldots, a_N]$ and $\hat{\mathbf{a}} = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_N]$ is defined as

$$\text{UIQI}(\mathbf{a}, \hat{\mathbf{a}}) = \frac{4\sigma_{a\hat{a}}^2 \mu_a \mu_{\hat{a}}}{(\sigma_a^2 + \sigma_{\hat{a}}^2)(\mu_a^2 + \mu_{\hat{a}}^2)}$$

where $(\mu_a, \mu_{\hat{a}}, \sigma_a^2, \sigma_{\hat{a}}^2)$ are the sample means and variances of a and $\hat{a}$, and $\sigma_{a\hat{a}}^2$ is the sample covariance of $(a, \hat{a})$. The range of UIQI is $[-1, 1]$ and $\text{UIQI}(\mathbf{a}, \hat{\mathbf{a}}) = 1$ when $\mathbf{a} = \hat{\mathbf{a}}$. For multi-band images, the overall UIQI can be computed by averaging the UIQI computed band-by-band.

4) ERGAS: The relative dimensionless global error in synthesis (ERGAS) calculates the amount of spectral distortion in the image [54]. This measure of fusion quality is defined as

$$\text{ERGAS} = 100 \times \frac{m}{n} \sqrt{\frac{1}{m_\lambda} \sum_{i=1}^{m_\lambda} \left(\frac{\text{RMSE}(i)}{\mu_i} \right)^2}$$

where m/n is the ratio between the pixel sizes of the MS and HS images, μ_i is the mean of the ith band of the HS image, and m_λ is the number of HS bands. The smaller ERGAS, the smaller the spectral distortion.

5) DD: The degree of distortion (DD) between two images $\mathbf{X}$ and $\hat{\mathbf{X}}$ is defined as

$$\text{DD}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{nm_\lambda} \|\text{vec}(\mathbf{X}) - \text{vec}(\hat{\mathbf{X}})\|_1.$$

The smaller DD, the better the fusion.

D. Comparison with other fusion methods

This section compares the proposed fusion method with four other state-of-the-art fusion algorithms for MS and HS images [14], [16]–[18]. The parameters used for the proposed fusion algorithm have been specified as follows

- The regularization parameter used in the SALSA method is $\mu = \frac{0.05}{\|\mathbf{N}_H\|_F}$. The selection of this parameter μ is still an open issue even if there are some strategies to tune it to accelerate convergence [31]. According to the convergence theory [55], for any $\mu > 0$, if the minimization of (14) has a solution, say $\mathbf{U}^\star$, then the sequence $\left\{ \mathbf{U}^{(t,k)} \right\}_{k=1}^\infty$ converges to $\mathbf{U}^\star$. If the minimization of (14) has no solution, then at least one of the sequences $\left\{ \mathbf{U}^{(t,k)} \right\}_{k=1}^\infty$ or $\left\{ \mathbf{G}^{(t,k)} \right\}_{k=1}^\infty$ diverges. Simulations have shown that the choice of μ does not affect significantly the fusion performance as long as μ is positive.

- The regularization coefficient is $\lambda = 25$. The choice of this parameter will be discussed in Section V-E.

All the algorithms have been implemented using MATLAB R2013A on a computer with Intel(R) Core(TM) i7-2600 CPU@3.40GHz and 8GB RAM. The fusion results obtained with the

different algorithms are depicted in Fig. 7. Visually, the proposed method performs competitively with other state-of-the-art methods. To better illustrate the difference of the fusion results, quantitative results are reported in Table I which shows the RMSE, UIQI, SAM, ERGAS and DD for all methods. It can be seen that the proposed method always provides the best results.

E. Selection of the regularization parameter λ

To select an appropriate value of λ, the performance of the proposed algorithm has been evaluated as a function of λ. The results are displayed in Fig. 8 showing that there is no optimal value of λ for all the quality measures. In the simulation of Section V-D, we have chosen $\lambda = 25$ which provides the best fusion results in terms of RMSE. It is noteworthy that in a wide range of λ, the proposed method always outperforms the other four methods.

F. Selection of subspace dimension $\widetilde{m}_\lambda$

As discussed in Section V-B1, an appropriate selection of subspace dimension is important for keeping spectral information while avoiding over-fitting problem. To illustrate the influence of the subspace dimension, another simulation considering top 10 components in PCA was implemented. The qualitative and quantitative results are reported in Fig. 9 and Table II.

Comparing Table II and Table I, it can be seen that most of the fusion methods degraded in this case which illustrates that selecting ten top components had caused over-fitting problem. How to adaptively evaluate the number of subspace is still an open question and is to be explored in the future work.

G. Test with other datasets

1) Fusion of PAVIA data and MS data (whole version): The fusion results, including qualitative and quantitative ones, of whole PAVIA dataset are also given in Fig. 10 and Table III. It is consistent with the results in Section V-D.

2) Fusion of AVIRIS data and MS data: The proposed fusion method has been tested with another dataset. The reference image is a $128 \times 128 \times 176$ hyperspectral image acquired over Moffett field, CA, in 1994 by the JPL/NASA airborne visible/infrared imaging spectrometer (AVIRIS) [56]. The blurring kernel $\mathbf{B}$, down-sampling operator $\mathbf{S}$ and SNRs for the two images are the same as in Section V-B. The reference image is filtered using the LANDSAT-like spectral responses depicted in Fig. 11, to obtain a 4-band MS image. For the dictionaries and supports, the number and size of atoms and the sparsity of the code are the same as in Section V-B. The proposed fusion method has been applied to the observed HS and MS images with a subspace of dimension $\widetilde{m}_\lambda = 10$. The regularization parameter has been selected by cross-validation to get the best performance in terms of RMSE. The images (reference, MS and MS) and the fusion results obtained with the different methods are shown in Fig. 12. More quantitative results are reported in Table IV. These results are in good agreement with what we obtained with the previous image, proving that the proposed sparse representation based fusion algorithms improves the fusion quality.

3) Pansharpening of AVIRIS data: The only difference with the Section V-G2 is that the MS image is replaced with a PAN image obtained by averaging all the bands of the reference image (contaminated by Gaussian noise with SNR = 30dB). The quantitative results are given in Table V and are again in favor of the proposed fusion method. The only difference with the Section V-G2 is that the MS image is replaced with a PAN image obtained by averaging all the bands of the reference image (contaminated by Gaussian noise with SNR = 30dB). The qualitative results are displayed in Fig. 13. The quantitative results are given in Table V and are again in favor of the proposed fusion method.

VI. Conclusions

In this paper, we proposed a novel method for hyperspectral and multispectral image fusion based on a sparse representation. The sparse representation ensured that the target image was well represented by atoms of dictionaries a priori learned from the observations. Identifying the supports jointly with the dictionaries circumvented the difficulty inherent to sparse coding. An

alternate optimization algorithm, consisting of an alternating direction method of multipliers and a least square regression, was designed to minimize the target function. Compared with other state-of-the-art fusion methods, the proposed fusion method offered smaller spatial error and smaller spectral distortion with a reasonable computation complexity. This improvement was attributed to the specific sparse prior designed to regularize the resulting inverse problem. Future works include the estimation the regularization parameter λ within the fusion scheme. Updating the dictionary jointly with the target image would also deserve some attention.

APPENDIX

Since $\mathbf{U}$ is a priori distributed according to a Gaussian distribution, and $\mathbf{Y}_{\mathrm{M}}$ is also Gaussian distributed conditional on $\mathbf{U}$ (as the noise is i.i.d Gaussian), $\mathbf{U}$ is also Gaussian distributed conditional on $\mathbf{Y}_{\mathrm{M}}$ [57]. Let define $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $\mathbf{Y}_{\mathrm{M}} = \left[\boldsymbol{y}_{\mathrm{M},1}, \ldots, \boldsymbol{y}_{\mathrm{M},n}\right]$, where the vectors $\boldsymbol{u}_i$ represent the pixels of the target image $\mathbf{U}$ and vectors $\boldsymbol{y}_{\mathrm{M},i}$ represent the pixels of MS image. Assuming that the vectors $\boldsymbol{u}_i$ are priori spatially decorrelated, the distribution of $\mathbf{U}$ conditional on $\mathbf{Y}_{\mathrm{M}}$ is

$$p(\mathbf{U}|\mathbf{Y}_{\mathrm{M}}) = \prod_{i=1}^{n} p(\boldsymbol{u}_i|\boldsymbol{y}_{\mathrm{M},i}) = \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_i|\boldsymbol{y}_{\mathrm{M},i}}, \mathbf{C}_{\boldsymbol{u}_i|\boldsymbol{y}_{\mathrm{M},i}})$$

We propose to define a rough approximation $\tilde{\mathbf{U}}$ of the projected target image $\mathbf{U}$ as the conditional mean $\boldsymbol{\mu}_{\mathbf{U}|\mathbf{Y}_{\mathrm{M}}} = \left[\boldsymbol{\mu}_{\boldsymbol{u}_1|\boldsymbol{y}_{\mathrm{M},1}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{u}_n|\boldsymbol{y}_{\mathrm{M},n}}\right] \in \mathbb{R}^{m_\lambda \times n}$. Straightforward computations lead to

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i|\boldsymbol{y}_{\mathrm{M},i}} = \mathrm{E}\left[\boldsymbol{u}_i\right] + \mathbf{C}_{\boldsymbol{u}_i,\boldsymbol{y}_{\mathrm{M},i}} \mathbf{C}_{\boldsymbol{y}_{\mathrm{M},i},\boldsymbol{y}_{\mathrm{M},i}}^{-1} \left[\boldsymbol{y}_{\mathrm{M},i} - \mathrm{E}\left[\boldsymbol{y}_{\mathrm{M},i}\right]\right]. \tag{20}$$

It can be seen that the computation of $\hat{\boldsymbol{\mu}}_{\boldsymbol{u}_i|\boldsymbol{y}_{\mathrm{M},i}}$ requires to determine $\mathrm{E}\left[\boldsymbol{u}_i\right]$, $\mathbf{C}_{\boldsymbol{u}_i,\boldsymbol{y}_{\mathrm{M},i}}$, $\mathbf{C}_{\boldsymbol{y}_{\mathrm{M},i},\boldsymbol{y}_{\mathrm{M},i}}$ and $\mathrm{E}\left[\boldsymbol{y}_{\mathrm{M},i}\right]$. The prior mean of projected image $\mathrm{E}\left[\boldsymbol{u}_i\right]$ is approximated by spatially interpolated observed HS imagery and the prior mean of MS image $\mathrm{E}\left[\boldsymbol{y}_{\mathrm{M},i}\right]$ is approximated by a spatially smoothed version of the MS image bands. These approximates are able to capture most of the nonstationarity exhibited by most natural images [58]. The estimation of the covariance matrix is conducted at the lower resolution of the observed HS imagery which is quite simple. More details about these approximations can be found in [14].

REFERENCES

[1] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fusion of multispetral and hyperspectral images based on sparse representation," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Lisbon, Portugal, 2014.

[2] I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, "A survey of classical methods and new trends in pansharpening of multispectral images," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–22, 2011.

[3] M. González-Audícana, J. L. Saleta, R. G. Catalán, and R. García, "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition," *IEEE Trans. Geosci. and Remote Sens.*, vol. 42, no. 6, pp. 1291–1299, 2004.

[4] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. and Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb 2011.

[5] D. Liu and P. T. Boufounos, "Dictionary learning based pan-sharpening," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 2397–2400.

[6] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 29–43, jan 2002.

[7] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[8] C.-I. Chang, *Hyperspectral data exploitation: theory and applications*. John Wiley & Sons, 2007.

[9] G. Chen, S.-E. Qian, J.-P. Ardouin, and W. Xie, "Super-resolution of hyperspectral imagery using complex ridgelet transform," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 10, no. 03, 2012.

[10] X. He, L. Condat, J. Bioucas-Dias, J. Chanussot, and J. Xia, "A new pansharpening method based on spatial and spectral sparsity priors," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4160–4174, Sept 2014.

[11] N. Yokoya and A. Iwasaki, "Hyperspectral and multispectral data fusion mission on hyperspectral imager suite (HISUI)," in *Proc. IEEE Int. Conf. Geosci. Remote Sens. (IGARSS)*, Melbourne, Australia, July 2013, pp. 4086–4089.

[12] V. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogramm. Eng. Remote Sens.*, vol. 58, no. 5, pp. 561–567, 1992.

[13] J. Zhou, D. Civco, and J. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[14] R. C. Hardie, M. T. Eismann, and G. L. Wilson, "MAP estimation for hyperspectral image resolution enhancement using an auxiliary sensor," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1174–1184, Sept. 2004.

[15] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multi-band images," *arXiv preprint arXiv:1307.5996*, 2013.

[16] ——, "Bayesian fusion of hyperspectral and multispectral images," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[17] Y. Zhang, S. De Backer, and P. Scheunders, "Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images," *IEEE Trans. Geosci. and Remote Sens.*, vol. 47, no. 11, pp. 3834 –3843, Nov. 2009.

[18] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. and Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.

[19] N. Yokoya, J. Chanussot, and A. Iwasaki, "Hyperspectral and multispectral data fusion based on nonlinear unmixing," in *Proc. IEEE GRSS Workshop Hyperspectral Image SIgnal Process.: Evolution in Remote Sens. (WHISPERS)*, Shanghai, China, June 2012, pp. 1–4.

[20] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recognition (CVPR)*, Minnesota, USA, 2007, pp. 1–8.

[21] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, 2008.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Machine Learning (ICML)*, Montreal, Canada, 2009, pp. 689–696.

[23] T. Deselaers and V. Ferrari, "Global and efficient self-similarity for object classification and detection," in *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2010, pp. 1633–1640.

[24] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[25] H. Yin, S. Li, and L. Fang, "Simultaneous image fusion and super-resolution using sparse representation," *Information Fusion*, vol. 14, no. 3, pp. 229–240, 2013.

[26] 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, 2006.

[27] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2010, pp. 3501–3508.

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

[29] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, pp. 267–288, 1996.

[31] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems." *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–95, 2011.

[32] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, 1999.

[33] R. Molina, M. Vega, J. Mateos, and A. K. Katsaggelos, "Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images," *Applied and Computational Harmonic Analysis*, vol. 24, no. 2, pp. 251 – 267, 2008.

[34] A. Jalobeanu, L. Blanc-Feraud, and J. Zerubia, "An adaptive Gaussian model for satellite image deblurring," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 613–621, 2004.

[35] A. Duijster, P. Scheunders, and S. De Backer, "Wavelet-based em algorithm for multispectral-image restoration," *IEEE Trans. Geosci. and Remote Sens.*, vol. 47, no. 11, pp. 3892–3898, 2009.

[36] M. Xu, H. Chen, and P. K. Varshney, "An image fusion approach based on Markov random fields," *IEEE Trans. Geosci. and Remote Sens.*, vol. 49, no. 12, pp. 5116–5127, 2011.

[37] C.-I. Chang, X.-L. Zhao, M. L. Althouse, and J. J. Pan, "Least squares subspace projection approach to mixed pixel classification for hyperspectral images," *IEEE Trans. Geosci. and Remote Sens.*, vol. 36, no. 3, pp. 898–912, 1998.

[38] J. M. Bioucas-Dias and J. M. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. and Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, 2008.

[39] T. Tu, S. Su, H. Shyu, and P. S. Huang, "A new look at IHS-like image fusion methods," *Information Fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[40] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. and Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, 2008.

[41] G. Dahlquist and Å. Björck, *Numerical Methods in Scientific Computing*, ser. Numerical Methods in Scientific Computing. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2008, vol. 1.

[42] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[43] 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.

[44] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising – part i: theory," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 539–554, 2006.

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

[46] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed. Boca Raton, FL: CRC press, 2013.

[47] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[48] Z.-Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 72, no. 1, pp. 7–35, 1992.

[49] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, 2010.

[50] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, Supplement 1, pp. S110–S122, 2009.

[51] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Trans. Geosci. and Remote Sens.*, vol. 7, no. 4, pp. 736–740, 2010.

[52] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. and Remote Sens.*, vol. 51, no. 2, pp. 844–856, 2013.

[53] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[54] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. Int. Conf. Fusion of Earth Data*, Nice, France, Jan 2000, pp. 99–103.

[55] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[56] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis *et al.*, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. of Environment*, vol. 65, no. 3, pp. 227–248, 1998.

[57] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[58] B. Hunt and T. Cannon, "Nonstationary assumptions for Gaussian models of images," *IEEE Trans. Systems, Man, Cybernet.*, vol. 6, no. 12, pp. 876–882, 1976.
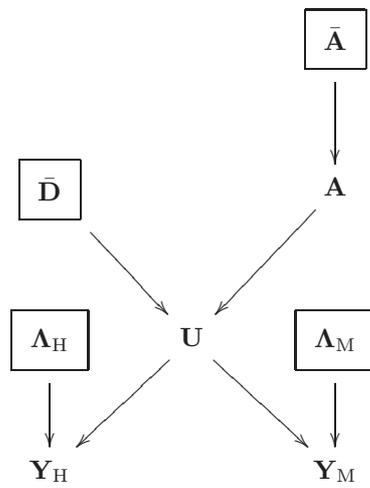
Fig. 1. DAG for the data, parameters and hyperparameters (the fixed parameters appear in boxes).

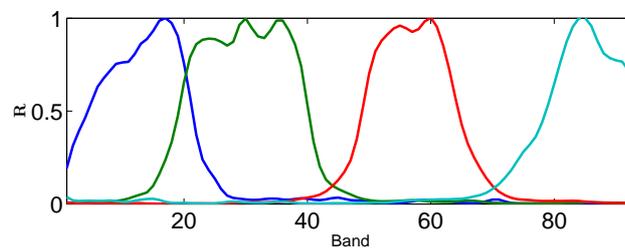Fig. 2. (left) Reference image. (middle) HS Image. (right) MS Image.
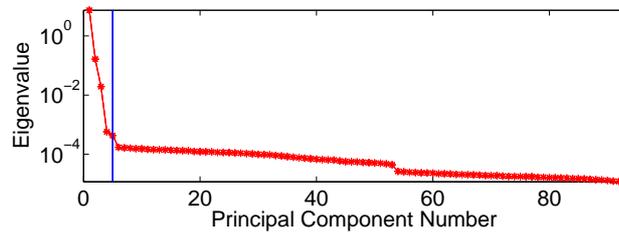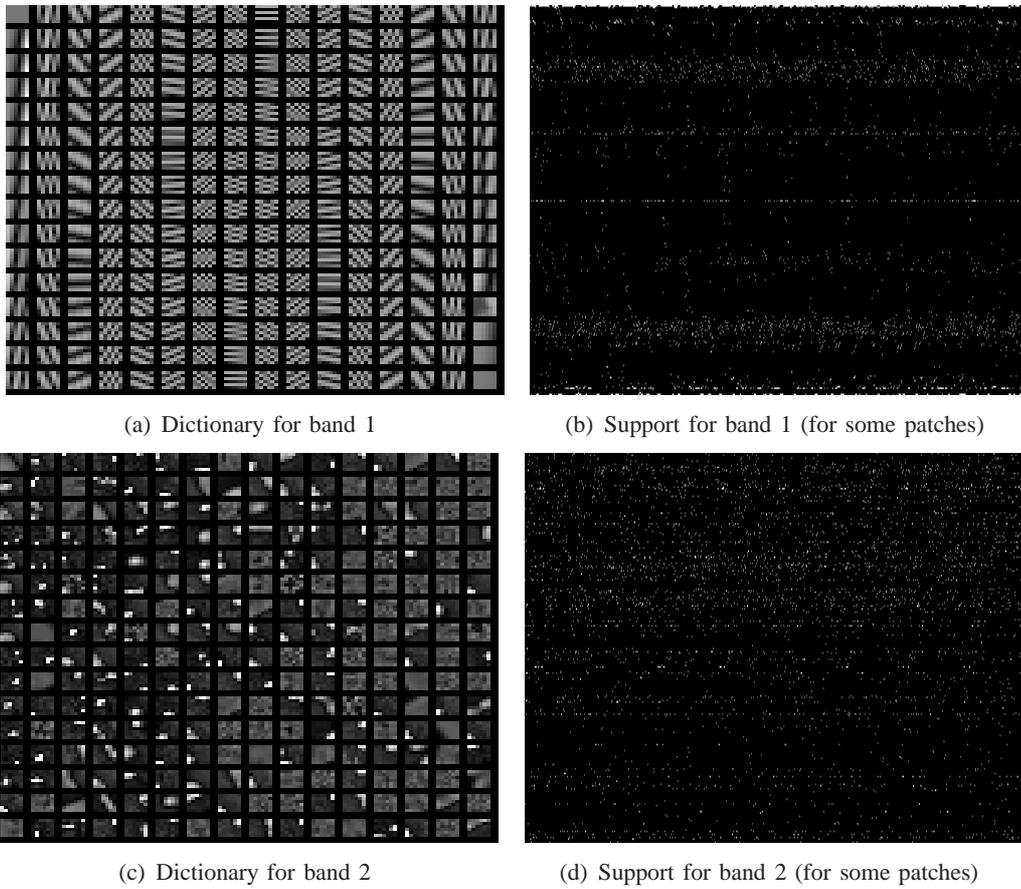

Fig. 3. IKONOS-like spectral responses.

Fig. 4. Eigenvalues of Υ for the Pavia HS image.

(a) Dictionary for band 1

(b) Support for band 1 (for some patches)

(c) Dictionary for band 2

(d) Support for band 2 (for some patches)

(e) Dictionary for band 3

(f) Support for band 3 (for some patches)

Fig. 5. Learned dictionaries (left) and corresponding supports (right).

(a) Dictionary for band 4

(b) Support for band 4 (for some patches)

(c) Dictionary for band 5
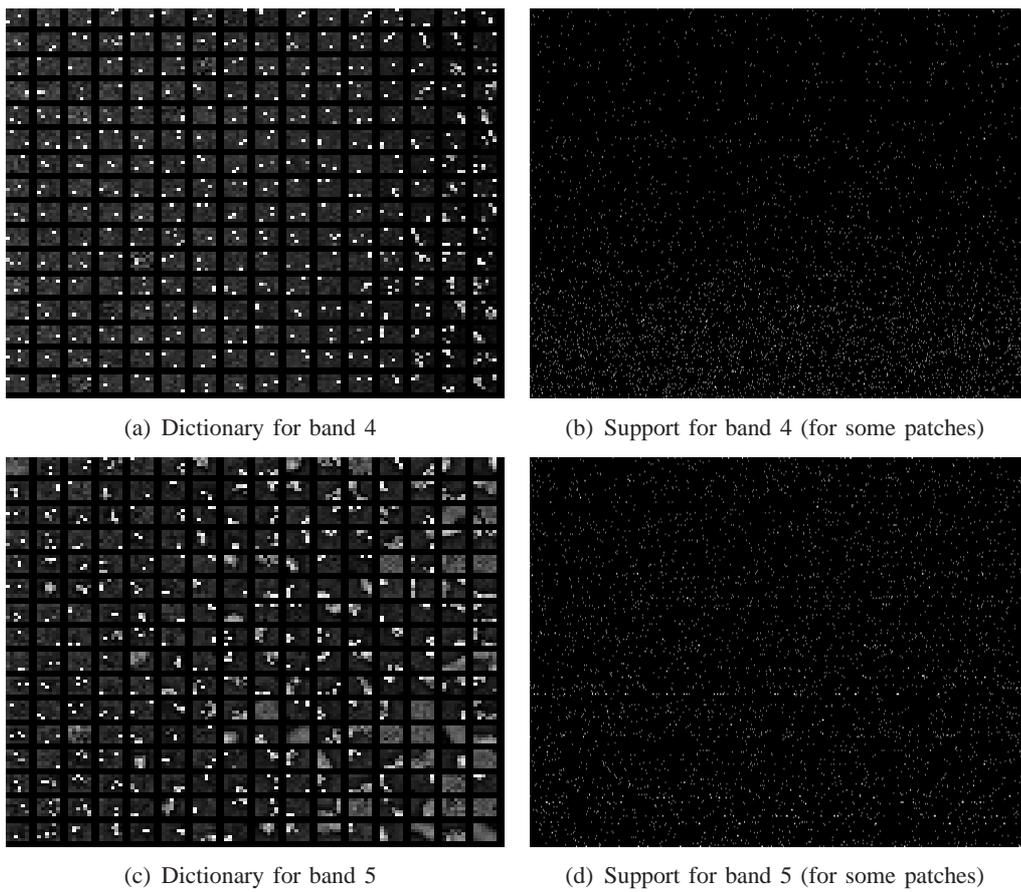
(d) Support for band 5 (for some patches)
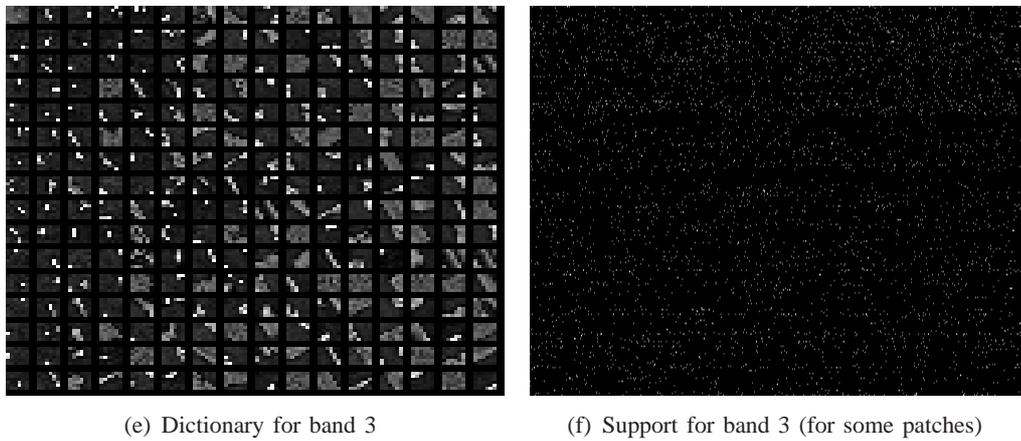
Fig. 6. Learned dictionaries (left) and corresponding supports (right).

Fig. 7. Pavia dataset: (Top 1) Reference. (Top 2) HS. (Top 3) MS. (Middle 1) MAP [14]. (Middle 2) Wavelet MAP [17]. (Middle 3) CNMF fusion [18]. (Bottom 1) MMSE estimator [16]. (Bottom 2) Rough estimation $\tilde{U}$. (Bottom 3) Proposed method.
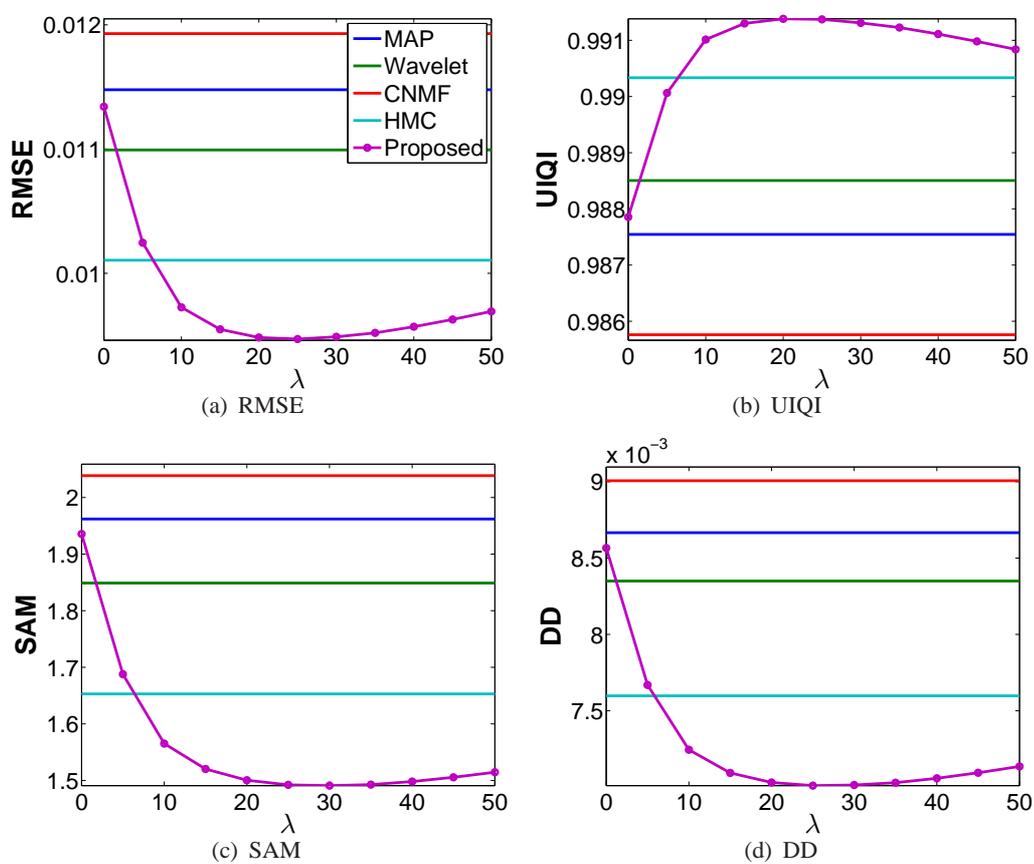
Fig. 8.   Performance of the proposed fusion algorithm versus λ.

Fig. 9. Pavia dataset (Top 10 components): (Top 1) Reference. (Top 2) HS. (Top 3) MS. (Middle 1) MAP [14]. (Middle 2) Wavelet MAP [17]. (Middle 3) CNMF fusion [18]. (Bottom 1) MMSE estimator [16]. (Bottom 2) Rough estimation $\breve{U}$. (Bottom 3) Proposed method.
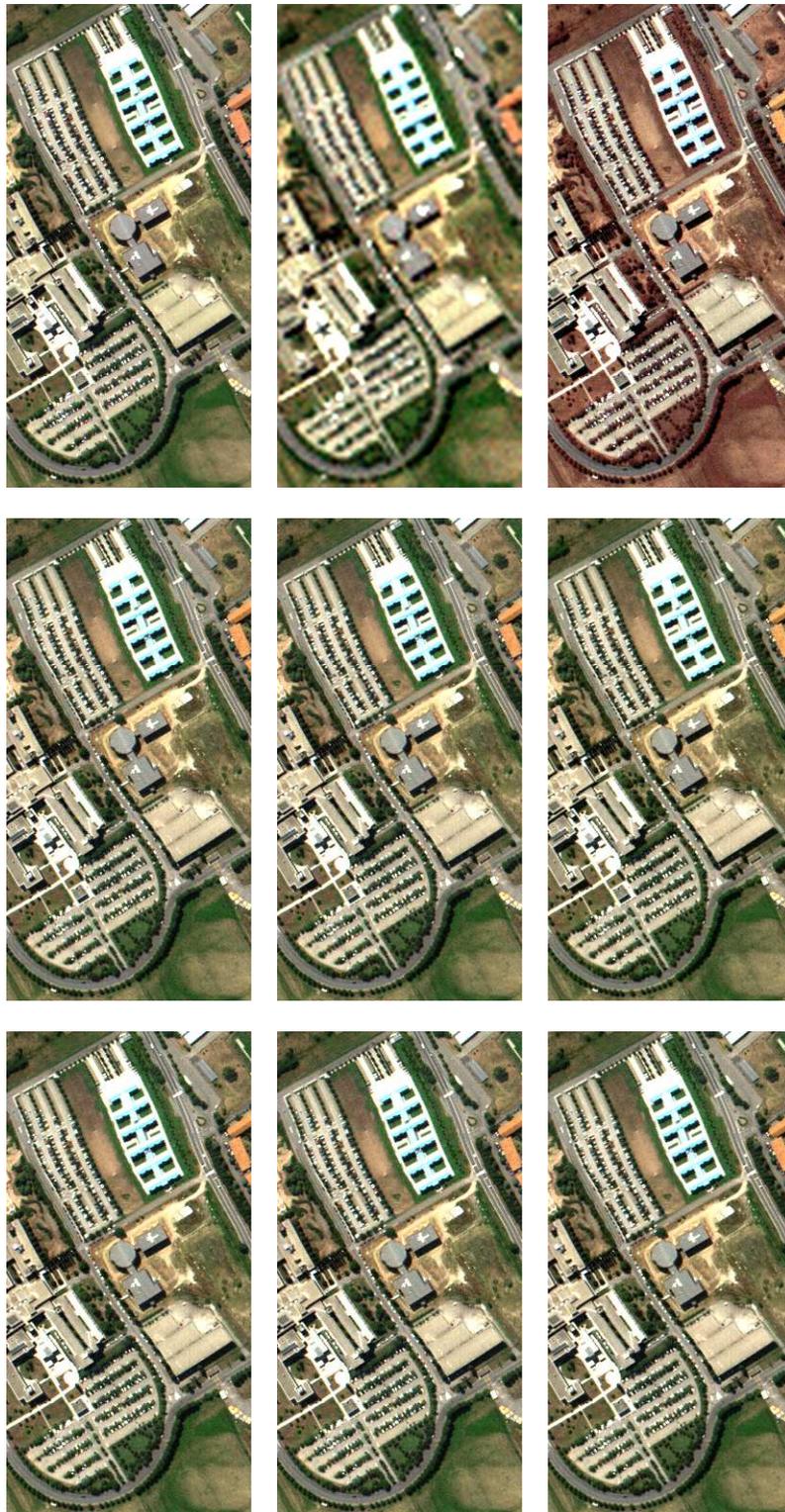
Fig. 10. Whole Pavia dataset: (Top 1) Reference. (Top 2) HS. (Top 3) MS. (Middle 1) MAP [14]. (Middle 2) Wavelet MAP [17]. (Middle 3) CNMF fusion [18]. (Bottom 1) MMSE estimator [16]. (Bottom 2) Rough estimation $\tilde{U}$. (Bottom 3) Proposed method.
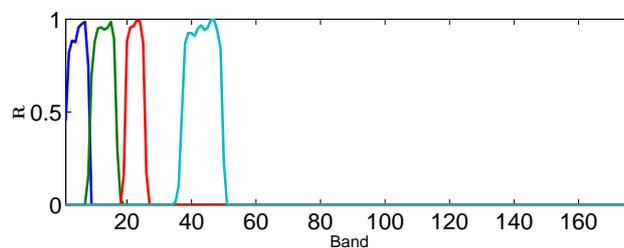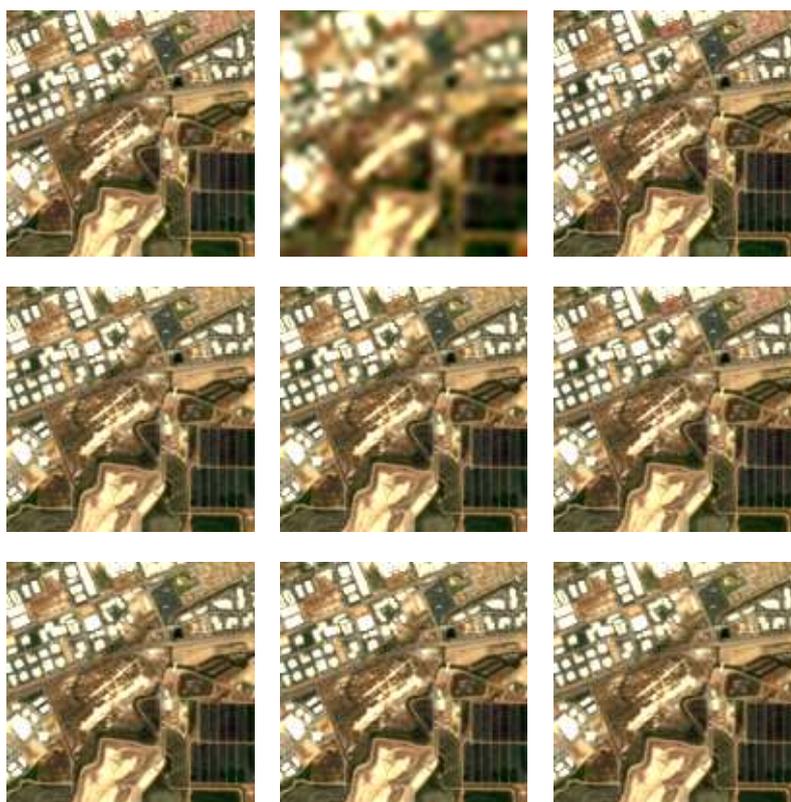
Fig. 11. LANDSAT spectral responses.

Fig. 12. Moffett dataset (HS+MS): (Top 1) Reference. (Top 2) HS. (Top 3) MS. (Middle 1) MAP [14]. (Middle 2) Wavelet MAP [17]. (Middle 3) CNMF fusion [18]. (Bottom 1) MMSE estimator [16]. (Bottom 2) Rough estimation $\tilde{\mathbf{U}}$. (Bottom 3) Proposed method.
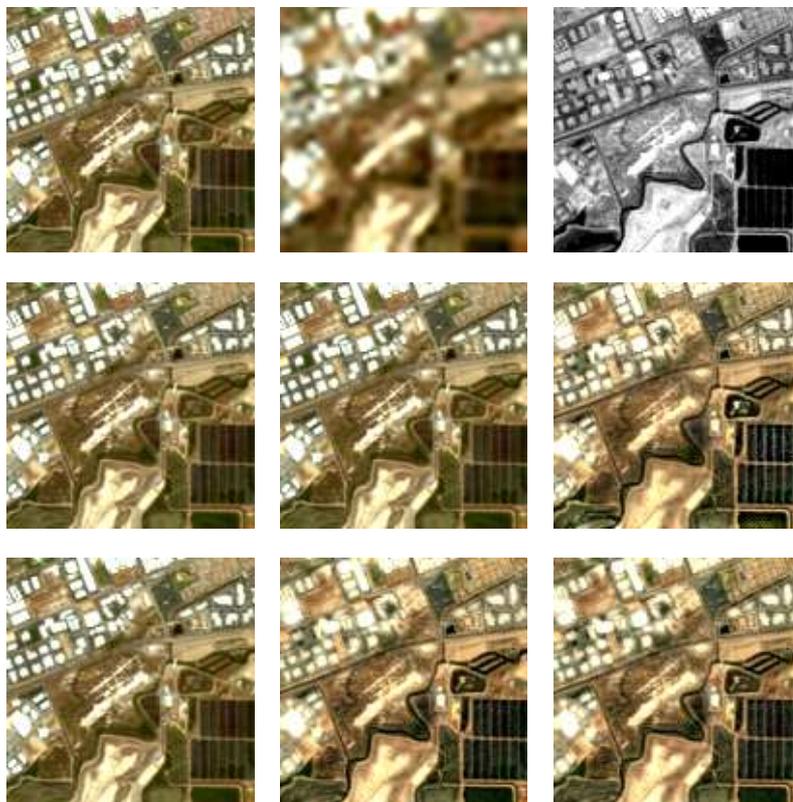
Fig. 13. Moffett dataset (HS+PAN): (Top 1) Reference. (Top 2) HS. (Top 3) PAN. (Middle 1) MAP [14]. (Middle 2) Wavelet MAP [17]. (Middle 3) CNMF fusion [18]. (Bottom 1) MMSE estimator [16]. (Bottom 2) Rough estimation $\tilde{\mathbf{U}}$. (Bottom 3) Proposed method.

TABLE I

PERFORMANCE OF DIFFERENT MS + HS FUSION METHODS (PAVIA DATASET): RMSE (IN 10^{-2}), UIQI, SAM (IN DEGREE), ERGAS, DD (IN 10^{-3}) AND TIME (IN SECOND).

Methods	RMSE	UIQI	SAM	ERGAS	DD	Time
MAP [14]	1.148	0.9875	1.962	1.029	8.666	**3**
Wavelet MAP [17]	1.099	0.9885	1.849	0.994	8.349	75
CNMF [18]	1.119	0.9857	2.039	1.089	9.007	14
HMC [16]	1.011	0.9903	1.653	0.911	7.598	6003
Rough $\tilde{U}$	1.136	0.9878	1.939	1.019	8.586	\
Proposed	**0.947**	**0.9913**	**1.492**	**0.850**	**7.010**	282

TABLE II

PERFORMANCE OF DIFFERENT MS + HS FUSION METHODS (PAVIA DATASET TOP 10 COMPONENTS): RMSE (IN 10^{-2}), UIQI, SAM (IN DEGREE), ERGAS, DD (IN 10^{-3}) AND TIME (IN SECOND).

Methods	RMSE	UIQI	SAM	ERGAS	DD	Time
MAP [14]	1.145	0.9876	1.958	1.030	8.671	**3**
Wavelet MAP [17]	1.173	0.9870	2.042	1.034	8.761	68
CNMF [18]	1.246	0.9851	2.134	1.111	9.136	13
HMC [16]	1.036	0.9899	1.736	0.924	7.765	11617
Rough $\tilde{U}$	1.156	0.9874	2.004	1.033	8.744	\
Proposed	**0.954**	**0.9912**	**1.516**	**0.853**	**7.059**	346

TABLE III

PERFORMANCE OF DIFFERENT MS + HS FUSION METHODS (WHOLE PAVIA DATASET): RMSE (IN 10^{-2}), UIQI, SAM (IN DEGREE), ERGAS, DD (IN 10^{-3}) AND TIME (IN SECOND).

Methods	RMSE	UIQI	SAM	ERGAS	DD	Time
MAP [14]	1.111	0.9879	1.921	1.012	8.359	**24**
Wavelet MAP [17]	1.037	0.9895	1.717	0.942	7.719	624
CNMF [18]	1.174	0.9864	1.903	1.061	8.457	102
HMC [16]	1.010	0.9899	1.668	0.928	7.513	27412
Rough $\tilde{U}$	1.103	0.9881	1.904	1.005	8.301	\
Proposed	**0.944**	**0.9911**	**1.491**	**0.864**	**6.887**	641

TABLE IV

PERFORMANCE OF DIFFERENT MS + HS FUSION METHODS (MOFFETT FIELD): RMSE (IN 10^{-2}), UIQI, SAM (IN DEGREE), ERGAS, DD (IN 10^{-2}) AND TIME (IN SECOND).

Methods	RMSE	UIQI	SAM	ERGAS	DD	Time
MAP [14]	2.583	0.9580	4.586	3.106	1.814	**3**
Wavelet MAP [17]	2.150	0.9695	3.815	2.584	1.509	68
CNMF [18]	2.218	0.9669	3.885	2.663	1.529	17
HMC [16]	1.813	0.9772	3.201	2.221	1.239	12708
Rough $\tilde{U}$	2.585	0.9579	4.575	3.107	1.814	\
Proposed	**1.745**	**0.9790**	**3.182**	**2.117**	**1.222**	280

TABLE V

PERFORMANCE OF DIFFERENT PANSHARPENING (HS + PAN) METHODS (MOFFETT FIELD): RMSE (IN 10^{-2}), UIQI, SAM (IN DEGREE), DD (IN 10^{-2}) AND TIME (IN SECOND).

Methods	RMSE	UIQI	SAM	ERGAS	DD	Time
MAP [14]	1.857	0.9690	4.162	2.380	1.356	**2**
Wavelet MAP [17]	1.848	0.9697	4.191	2.354	1.360	55
CNMF [18]	1.964	0.9669	4.569	2.467	1.450	5
HMC [16]	1.748	0.9730	3.996	2.234	1.288	7828
Rough $\tilde{U}$	1.853	0.9691	4.158	2.375	1.355	\
Proposed	**1.745**	**0.9731**	**3.948**	**2.231**	**1.281**	252