Sparse representations and dictionary learning: from image fusion to motion estimation

Jean-Yves Tourneret, Adrian Basarab, Nora Leïla Ouzir, Qi Wei

SPARSE REPRESENTATIONS AND DICTIONARY LEARNING: FROM IMAGE FUSION TO MOTION ESTIMATION

J.-Y. Tourneret[(1)]*, A. Basarab*[(2)]*, N. Ouzir*[(3)] *and Q. Wei*[(4)]

[(1)] University of Toulouse/INP-ENSEEIHT/IRIT/TSA, 2 Rue Charles Camichel, Toulouse, France (jean-yves.tourneret@toulouse-inp.fr),

[(2)] University of Toulouse, Université Paul Sabatier Toulouse 3, IRIT, CNRS UMR 5505, France (adrian.basarab@irit.fr),

[(3)] University of Paris-Saclay/Inria/CentraleSupelec/CVN, 9 Rue Joliot Curie, Gif Sur Yvette, France (nora.ouzir@centralesupelec.fr),

[(4)] Machine Learning Center of Excellence, J. P. Morgan, 383 Madison Ave, New York, NY, USA 10017 (bjweiqi@gmail.com).

ABSTRACT

The first part of this paper presents some works conducted with Jose Bioucas Dias for fusing high spectral resolution images (such as hyperspectral images) and high spatial resolution images (such as panchromatic or multispectral images) in order to build images with improved spectral and spatial resolutions. These works are related to Bayesian fusion strategies exploiting prior information about the target image to be recovered constructed by dictionary learning. Interestingly, these Bayesian image fusion methods can be adapted with limited changes to motion estimation in pairs or sequences of images. The second part of this paper explains how the work of Jose Bioucas Dias has been a source of inspiration for developing new Bayesian motion estimation methods for ultrasound images.

***Index Terms*—** Image fusion, motion estimation, sparse representations, dictionary learning.

1. INTRODUCTION

Fusion of multi-sensor images has been explored during recent years and is still a very active research area [1]. 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*. 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 and spectral unmixing. However, while HS sensors provide abundant spectral information, their spatial resolution is generally limited. 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. 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. An example of multi-band image fusion is illustrated in Fig. 1. The problem of fusing images with different spectral and spatial resolutions was studied during the PhD thesis of Qi Wei conducted from Oct. 2012 to Sept. 2015 in the university of Toulouse. Jose Bioucas Bias visited the university of Toulouse in Feb. 2015 and June 2015 as an invited Professor and had the outstanding idea of regularizing this fusion problem with penalizations based on dictionary learning. The resulting fusion algorithms will be presented in Section 2.

Motion estimation from a sequence of temporal medical images is apparently a very different problem, consisting of estimating the displacements between pixels contained in successive frames. In cardiac imaging, motion estimation is the first step towards assessing quantitative parameters such as strain or stiffness, well-known to carry useful clinical information. Among medical imaging modalities, ultrasound is well adapted to cardiac motion estimation given its high frame rate, from tens to hundreds or even thousands of images per second depending on the acquisition scheme, e.g., compared to magnetic resonance or computed tomography. However, estimating accurately the cardiac motion from ultrasound data is still a challenging problem, mainly because of the high amount of speckle noise affecting the images and of the high complexity of myocardium deformation. This motion estimation problem was addressed in the PhD thesis of Nora Ouzir, conducted from Oct. 2015 to Oct. 2018 in the university of Toulouse, in which the multi-sensor image fusion problem was adapted to estimating the motion between several ultrasound images. More precisely, image fusion and motion estimation can be addressed by solving similar inverse problems defined by data fidelity terms and appropriate regularizations based on dictionary learning. Jose Bioucas Dias visited the university of Toulouse as an invited Professor in July 2017 and Feb. 2018 and had very relevant feedback about the motion estimation problem.

This paper summarizes the works conducted on image fusion and motion estimation with Jose Bioucas Dias. It is organized as follows. Section 2 describes the inverse problem that Jose Bioucas Dias suggested to improve the spatial and spectral resolution of several remote sensing images. It is presented in details for HS and MS images but can be generalized to any pair of remote sensing images. The generalization of this fusion problem to motion estimation is presented in Section 3. Conclusions are reported in Section 4.

2. FUSION OF IMAGES WITH DIFFERENT SPATIAL AND SPECTRAL RESOLUTIONS

2.1. Problem formulation

Mathematically, the fusion of HS and MS images can be formulated as the following minimization problem

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

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 ($\mathbf{Y}_{\mathrm{H}}$ and $\mathbf{Y}_{\mathrm{M}}$ are the observed HS and MS im-

ages, $\boldsymbol{X} = \boldsymbol{HU}$ is the unknown full resolution image that has been projected onto a subspace of lower dimension using the orthogonal matrix $\boldsymbol{H}$, $\boldsymbol{B}$ is a decimation operator acting on the bands, $\boldsymbol{S}$ is a downsampling matrix, $\boldsymbol{R}$ is the spectral response of the MS sensor and $\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{M}}$ are diagonal matrices containing the noise variances, see [2] for more details). Note that recovering the projected high-spectral and high-spatial resolution image $\mathbf{U}$ from the observations $\mathbf{Y}_{\mathrm{H}}$ and $\mathbf{Y}_{\mathrm{M}}$ using (1) is a linear inverse problem (LIP).

Fig. 1: Pavia image: HS (left), MS (middle) and fused image (right).

2.2. Image fusion based on dictionary learning

The strategy proposed in [2] and suggested by Jose Bioucas Dias to solve the LIP (1) was to use the self-similarity property of natural images, and introduce a sparse regularization term to regularize the fusion problem. This regularization relies on a decomposition of the scene on a set of dictionaries (with one example shown in Fig. 2). The goal of this 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 sparse regularization can be defined as

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

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 $i = 1, \ldots, \widetilde{m}_\lambda$ (n is the number of image pixels and $\widetilde{m}_\lambda$ is the dimension of the projection subspace),
- $\mathcal{P}(\cdot) : \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{pat}}} \mapsto \mathbb{R}^{n \times 1}$ is a linear operator that averages the overlapping patches of each band,
- $\bar{\mathbf{D}}_i \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{at}}}$ is an overcomplete dictionary whose columns are basis elements of size n_{p} (patch size),
- $\bar{\mathbf{A}}_i \in \mathbb{R}^{n_{\mathrm{at}} \times n_{\mathrm{pat}}}$ is the ith band code (n_{at} is the number of atoms and n_{pat} is the number of patches for the ith band).

The dictionary atoms and the supports of the corresponding active coding coefficients can be jointly learned from the observed images. Conditionally on these dictionaries and supports, the maximum a posterior (MAP) estimator of the unknown image can be computed using an optimization framework to solve the fusion problem. The resulting fusion method based on sparse representation and dictionary learning provided smaller spatial error and smaller spectral distortion with a reasonable computation complexity compared with

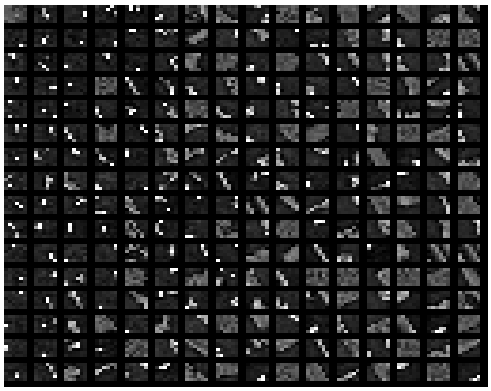

Fig. 2: Example of a learned dictionary.

other state-of-the-art fusion methods as summarized in Table 1 (extracted from [2] where the different performance measures RMSE, UIQI, SAM, ERGAS and DD are defined). This improvement was attributed to the specific sparse prior designed to regularize the resulting inverse problem.

To further reduce the computational complexity of solving the associated optimization problem, following the method developed in [3], we introduced a robust fast multi-band image fusion method based on a generalized Sylvester matrix equation solver. Specifically, this generalized Sylvester matrix equation associated with the multi-band image fusion problem was solved in a more robust and efficient way by exploiting the Woodbury formula, avoiding any permutation operation in the frequency domain as well as the blurring kernel invertibility assumption required in [3]. Thanks to this improvement, the fusion algorithm required fewer computational operations and was also more robust with respect to the blurring kernel of [3]. The resulting algorithm was tested with different priors considered in [3] and the experimental results shown in Table 2 demonstrated the robustness of the fusion algorithm with a Sylvester matrix equation with a reduced computational cost.

Table 1: Performance of different Pansharpening (HS + PAN) methods for the Moffett field image: RMSE (in 10^{-2}), UIQI, SAM (in degree), DD (in 10^{-2}) and Time (in seconds).

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

2.3. Joint fusion and unmixing

To further explore and leverage the intrinsic physical spectral properties of the multi-band images, we proposed to extend the proposed fusion method to solve the spectral unmixing and image fusion problem jointly. Specifically, each spectral vector of an image can be represented by a linear mixture of several spectral signatures, referred to as endmembers with some examples shown in Fig. 3. Mathematically, we have $\boldsymbol{X} = \boldsymbol{MA}$, where $\mathbf{M} \in \mathbb{R}^{m_\lambda \times p}$ is the endmember matrix whose columns are spectral signatures and $\mathbf{A} \in \mathbb{R}^{p \times n}$ is the corresponding abundance matrix whose columns are abundance fractions. This linear mixture model (LMM) has been widely used

Table 2: Performance of HS+MS fusion methods: RSNR (in dB), UIQI, SAM (in degree), ERGAS, DD (in 10^{-3}) and time (in second).

Prior	Methods	RSNR	UIQI	SAM	ERGAS	DD	Time
Gaussian	FUSE	29.243	0.9904	1.513	0.902	6.992	0.27
	R-FUSE	29.243	0.9904	1.513	0.902	6.992	**0.24**
TV	FUSE	29.629	0.9914	1.456	0.853	6.761	133
	R-FUSE	29.629	0.9914	1.456	0.853	6.761	**115**

in HS unmixing. Note that using this LMM is the main difference with the algorithm of [2], in which each spectral vector of an image is decomposed in a subspace as $\mathbf{X} = \mathbf{HU}$. In this setting, instead of learning sparse representation from the observed images, the regularization is implicitly imposed by a low rank representation inherent to the linear spectral mixture model and by non-negativity and sum-to-one constraints resulting from the intrinsic physical properties of the abundances. The endmember signatures and abundances are then jointly estimated from the observed multi-band images, using constrained linear regression problems, which can be solved efficiently by the alternating direction method of multipliers (ADMM).

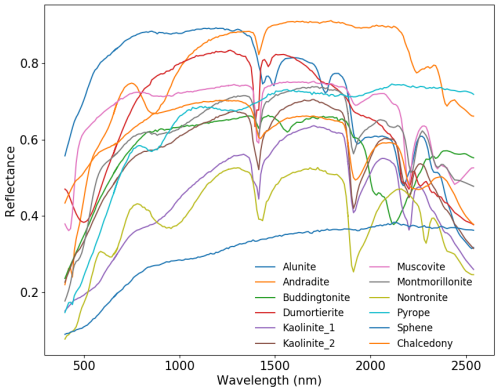

Fig. 3: Examples of endmember signatures for different materials.

More precisely, the joint fusion and unmixing problem can be addressed by solving the following minimization problem

$$\min_{\mathbf{M},\mathbf{A}} L(\mathbf{M},\mathbf{A}) \quad \text{s.t. } \mathbf{A} \geq \mathbf{0},\ \mathbf{1_p^T A} = \mathbf{1_n^T},\ \mathbf{0} \leq \mathbf{M} \leq \mathbf{1}, \tag{3}$$

where

$$L(\mathbf{M},\mathbf{A}) = \frac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{H}}^{-\frac{1}{2}}\left(\mathbf{Y}_{\mathrm{H}} - \mathbf{MABS}\right)\right\|_F^2 + \frac{1}{2}\left\|\mathbf{\Lambda}_{\mathrm{M}}^{-\frac{1}{2}}\left(\mathbf{Y}_{\mathrm{M}} - \mathbf{RMA}\right)\right\|_F^2.$$

Instead of solving this problem by decoupling two data terms as in [2], the algorithm based on the LMM, referred to as FUMI, directly minimizes the associated objective function and updates the endmembers and abundances alternatively, both using an ADMM. Specifically, the updates for abundances consisted of solving a Sylvester matrix equation and projecting onto a simplex. Thanks to the R-FUSE algorithm of [4], this Sylvester equation can be solved analytically thus efficiently, requiring no iterative update. On the other hand, the endmember updating can be divided into two steps: a least square regression and a thresholding, that are both not computationally intensive. This FUMI algorithm gave better fused images and better unmixing results when compared to other state-of-the-art methods with slightly more cost, mainly due to the alternating updates of the endmembers and abundances.

3. MOTION ESTIMATION

3.1. Image pairwise motion estimation

Traditionally, optical flow-based methods are used to estimate the cardiac motion maps from a pair of ultrasound images representing the heart at two different instants of the cardiac cycle. Within such approaches, the motion is estimated by minimizing a cost function composed of two terms, a data fidelity term, generally derived from statistical assumptions on the ultrasound images, and a spatial regularization term mitigating the ill-posedness of the problem. In our previous work, we have shown that the accuracy of the estimation can be greatly improved by incorporating two additional ingredients: (i) a sparsity-based regularization term employing a dedicated dictionary containing typical local cardiac motion patterns [5], and (ii) a robustification of each term of the cost function against outliers (such as image attenuation, parasite wave reflections, or motion boundaries) [6]. Let us denote by $\boldsymbol{I} \in \mathbb{R}^{2n}$ the vectorized pair of input images and by $\boldsymbol{U} \in \mathbb{R}^{2n}$ the motion field to be estimated, n being the number of pixels in each image. In [6], the robust approach (ii) made use of M-estimator-based weights $\boldsymbol{Q}$, $\boldsymbol{S}$, and $\boldsymbol{W}$ rejecting the outliers of the optical flow data fidelity term, spatial, and sparsity-based regularizations, respectively. The resulting minimization problem was

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{\alpha}} \|\boldsymbol{Q}^{1/2}(\partial_t \boldsymbol{I} + \nabla \boldsymbol{I}^T \boldsymbol{U})\|_2^2 + \lambda_s \|\boldsymbol{S}^{1/2}\nabla \boldsymbol{U}\|_2^2 \\ + \lambda_d \sum\nolimits_p \|\boldsymbol{W}_p^{1/2}(\boldsymbol{P}_p \boldsymbol{U} - \boldsymbol{D}\boldsymbol{\alpha}_p)\|_2^2 \\ \text{s.t } \|\alpha_{u,p}\|_0 \leq K, \|\alpha_{u,p}\|_0 \leq K, \forall p,\end{aligned} \tag{4}$$

where ∇ is the spatial gradient operator and the parameters $\lambda_s \in \mathbb{R}^+$ and $\lambda_d \in \mathbb{R}^+$ control the influence of each term on the cost function. The third term in (4) is the sparsity-based regularization, which operates patch-wise on the horizontal and vertical motion components. The patches are extracted using the operator $\boldsymbol{P}_p$, where the subscript p denotes the patch index. The sparse coefficients indicating the elements in the dictionary $\boldsymbol{D}$ to be used for each patch are denoted by $\boldsymbol{\alpha}_{u,p}$ and $\boldsymbol{\alpha}_{v,p}$ for the horizontal and vertical components. Finally, K is a user-defined parameter adjusting the sparsity level.
This robust approach using a sparsity-based regularization showed interesting improvements over the standard optical-flow [6]. Fig. 4(a) shows an example of an estimated motion field illustrating how the method accurately describes the complex inward motions of the systole phase. While this method allowed us to estimate the cardiac motions from pairs of images efficiently, it did not incorporate the temporal aspect of the problem. Specifically, we did not account for the temporal dynamics of motion, that can be captured by processing multiple ultrasound frames at the same time. Fig. 4(b) shows an example of temporal evolution of the displacements for a physical structure in the mid segment of the heart. One can see how the displacements of this point, obtained from image pairwise estimation lack smoothness while the motion itself is continuous in time.

3.2. Generalization to image sequences

Image pairwise motion estimation is sub-optimal because it does not allow for incorporating and reinforcing the temporal coherence of

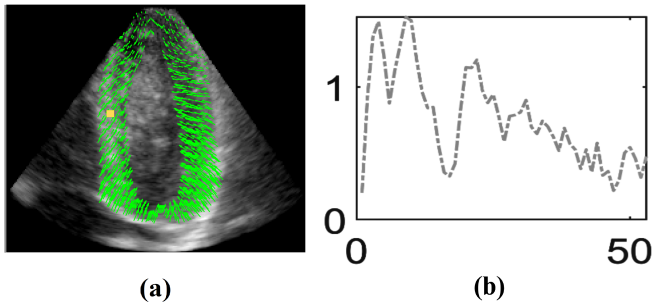

Fig. 4: (a) Example of an estimated motion field using the pair-wise method, and (b) Displacements for a specific pixel (yellow square) over the entire cardiac sequence.

motion fields. Generalizing the previous algorithm to image sequences was thus an appealing idea, albeit not straightforward to apply. The main difficulties were related to the much larger size of the problem, the non-linearity associated with incorporating temporal coherence, as well as the non-convexity of the original ℓ_0-based sparsity constraint. These challenges were addressed in close collaboration with Jose Bioucas-Dias during his one month visit as an invited professor at Toulouse University in 2018. The contributions of Prof. Bioucas-Dias in medical imaging, and particularly in ultrasound imaging, date back to 1996, with his inspiring work on contour estimation in echocardiography [7]. His contributions to the field of ultrasound imaging continued more recently in the context of 3D reconstruction [8]. It was also with great enthusiasm that Prof. José Bioucas-Dias accepted to collaborate on the topic of cardiac motion estimation. He valued the easy access to ultrasound data using existing platforms that were not available a few years ago.

Our collaboration with Prof. Bioucas-Dias focused on two key aspects of cardiac motion estimation from image sequences. First, a reformulation of (4) using the entire sequence of images and different regularizations was proposed. This aspect of our collaboration involved 1) a careful reformulation of the optical-flow data fidelity term, 2) a relaxation of the sparsity constraints, and 3) the design of a novel linear temporal regularization ensuring a smooth evolution of the motions. The resulting minimization problem was

$$\min_{\boldsymbol{U},\boldsymbol{\Omega}} \frac{1}{2}\|\boldsymbol{Y}+\boldsymbol{A}\boldsymbol{U}\|_F^2+\lambda_s(\|\boldsymbol{L}_h\boldsymbol{U}\|_F^2+\|\boldsymbol{L}_v\boldsymbol{U}\|_F^2) + \lambda_d\|\boldsymbol{P}\boldsymbol{U}-\boldsymbol{D}\boldsymbol{\Omega}\|_F^2+\lambda\|\boldsymbol{\Omega}\|_1+\lambda_{\mathrm{T}}E_{\mathrm{T}}(\boldsymbol{U}), \quad (5)$$

where the optical-flow related temporal and spatial derivatives of all the m images in the sequence have now been included in $\boldsymbol{Y}\in\mathbb{R}^{nm}$ and $\boldsymbol{A}\in\mathbb{R}^{nm\times 2nm}$. Similarly, Ω contains the sparse coefficients of all the patches in the sequence ($\boldsymbol{P}$ now operates on m images), and the motions to be estimated are $\boldsymbol{U}\in\mathbb{R}^{2nm}$. Note that we used a convex relaxation of the sparsity constraint using the ℓ_1-norm and an additional regularization term enforcing temporal coherence E_{T}. For conciseness, we will not develop this temporal regularization term introduced in the conference paper [9]. Note also that the spatial regularization has been split to horizontal and vertical terms using the gradient operators L_h and L_v, respectively. The parameters λ and λ_{T} again control the trade-off between the different terms of the objective function.

A second crucial contribution focused on developing an efficient optimization strategy for (5). Prof. Bioucas-Dias had conducted leading works on solving challenging inverse problems, such as the dictionary learning-based fusion method described in the previous section. In our joint work, he proposed a new C-SALSA-based method (with reduced computational cost) for solving the minimization problem (5) (see [9] for more details). Fig. 5 shows the displacements corresponding to the same pixel of Fig. 4 obtained by this new temporally coherent method. One can see the obvious improvement in the smoothness of the estimated displacements.

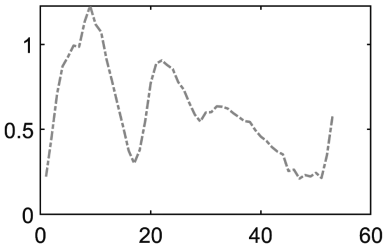

Fig. 5: Displacements of the pixel (yellow square) of Fig. 4 (a) obtained by minimizing the new cost function in (5) incorporating temporal coherence.

4. CONCLUSIONS

The authors of this paper would like to dedicate this work to their friend Jose Bioucas Dias who sadly and unexpectedly passed away in September 2020. They will never forget his enthousiasm that motivated these works and his personal kindness.

5. REFERENCES

[1] L. Loncan, L. B. Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. Licciardi, M. Simoes, J-Y. Tourneret, M. Veganzones, G. Vivone, Q. Wei, and N. Yokoya, "Hyperspectral pansharpening: a review," *IEEE Geosci. and Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sept. 2015.

[2] Q. Wei, J. M. Bioucas-Dias, N. Dobigeon, and J-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. and Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, July 2015.

[3] Q. Wei, N. Dobigeon, and J-Y. Tourneret, "Fast multi-band image fusion based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.

[4] Q. Wei, N. Dobigeon, J-Y. Tourneret, J. M. Bioucas-Dias, and S. Godsil, "R-fuse: Robust fast fusion of multi-band images based on solving a sylvester equation," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1632–1636, Nov. 2016.

[5] N. Ouzir, A. Basarab, H. Liebgott, B. Harbaoui, and J.-Y. Tourneret, "Motion estimation in echocardiography using sparse representation and dictionary learning," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 64–77, Jan. 2018.

[6] N. Ouzir, A. Basarab, O. Lairez, and J.-Y. Tourneret, "Robust optical flow estimation in cardiac ultrasound images using a sparse representation," *IEEE Trans. on Medical Imaging.*, vol. 38, no. 3, pp. 741–752, March 2019.

[7] J. M. Bioucas-Dias and J. M. N. Leitao, "Wall position and thickness estimation from sequences of echocardiographic images," *IEEE Trans. on Medical Imaging.*, vol. 15, no. 1, pp. 25–38, Jan. 1996.

[8] J. Sanches, J. M. Bioucas-Dias, and J. S. Marques, "Minimum total variation in 3D ultrasound reconstruction," in *Proc. International Conference on Image Processing (ICIP)*, Genoa, Italy, Sept. 11-14 2005.

[9] N. Ouzir, J. Bioucas-Dias, A. Basarab, and J.-Y. Tourneret, "Robust cardiac motion estimation with dictionary learning and temporal regularization for ultrasound imaging," in *Proc. IEEE Int. Ultrasonics Symposium (IUS)*, Glasgow, Scotland, Oct. 6-9 2019.