VIDEO RECONSTRUCTION USING COMPRESSED SENSING MEASUREMENTS AND 3D TOTAL VARIATION REGULARIZATION FOR BIO-IMAGING APPLICATIONS

Yoann Le Montagner[1,2]*, Elsa Angelini*[2]*, Jean-Christophe Olivo-Marin*[1]

[1]Institut Pasteur, Unité d'Analyse d'Images Quantitative CNRS URA 2582, F-75015 Paris
[2]Institut Télécom, Télécom ParisTech CNRS LTCI, F-75013 Paris

ABSTRACT

The theory of compressed sensing (CS) predicts that random (or pseudo-random) linear measurements together with non-linear reconstruction can be used to sample and recover structured signals in a compressive manner. Lots of previous results demonstrated the efficiency of CS in recovering 2D images acquired using dedicated CS devices (single-pixel camera, accelerated MRI, etc...).

In this paper, we investigate how this framework can be extended to perform an efficient joint reconstruction of a sequence of time-correlated 2D images, using 3D total variation regularization. We also evaluate the performances of this framework on test sequences issued from the bio-imaging field.

Index Terms— Compressed sensing, total variation, video.

1. INTRODUCTION

1.1. CS background and notations

The inverse problem tackled by CS can be formulated as follows: given a signal of interest $x \in \mathbb{R}^N$ measured through a *random* linear operator Φ that outputs a vector $y \in \mathbb{R}^M$ of observations with $M \ll N$, can x be recovered from y? The randomness of the measurement operator Φ should not be understood in strict meaning, but rather as the fact that Φ should spread the information contained in x over the whole vector y. Examples of such operators include random Gaussian or Bernouilli matrices [5], randomly subsampled Fourier or Hadamard transforms [3], or dedicated unitary matrices [7].

Previous results (see [4, 8, 2]) establishes that x can be recovered from y if it has a sparse representation in some known dictionary Ψ, i.e. there exists a sparse vector $\gamma \in \mathbb{R}^D$ such that $x = \Psi\gamma$, and if Φ behaves *like an isometry* for sparse linear combinations of columns of Ψ; this idea is quantified using the notion of *restricted isometry property* (see [2] for more details). Up to some technical hypothesis, an estimator $\hat{x}$ of x can be defined as a solution of the following convex problem (known as l_1*-analysis problem*):

$$\hat{x} = \underset{x \in \mathbb{R}^N}{\arg\min} \left\|\Psi^* x\right\|_1 \text{ s.t. } \left\|\Phi x - y\right\|_2 \leq \epsilon \tag{1}$$

where ϵ is a parameter tuned according to the level of noise that corrupts the observations y. As an alternative to (1), x can also be estimated through (2) (known as l_1*-synthesis problem*):

$$\hat{x} = \Psi\hat{\gamma} \quad \text{where} \quad \hat{\gamma} = \underset{\gamma \in \mathbb{R}^D}{\arg\min} \left\|\gamma\right\|_1 \text{ s.t. } \left\|\Phi\Psi\gamma - y\right\|_2 \leq \epsilon \tag{2}$$

Although (1) and (2) are equivalent in the case when Ψ is an orthonormal basis, it is not the case for a general dictionary. Empirical studies show that l_1-synthesis can be effective in some situations involving overcomplete dictionaries, but it also leads to practical difficulties, such as working in higher dimensional spaces (see [9]).

1.2. Video reconstruction problem formulation

We focus on the following problem: a signal of interest $\boldsymbol{x} \in \mathbb{R}^{N \times T}$ composed of T successive frames $x_t \in \mathbb{R}^N$ $(1 \leq t \leq T)$[1] is measured through a linear memoryless operator $\boldsymbol{\Phi}$, resulting in a vector $\boldsymbol{y} \in \mathbb{R}^M$ of observations. Formally:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_T \end{bmatrix}}_{\boldsymbol{\Phi}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix}}_{\boldsymbol{x}} \tag{3}$$

The *memoryless* notion means that $\boldsymbol{y}$ is accumulated from T sub-vectors $y_t \in \mathbb{R}^{M_t}$ of observations, with each y_t depending only on a given frame x_t; this measurement model corresponds to a 2D CS measurement device that would image objects in motion, stacking the consecutive 2D measurements. In such acquisition mode, temporal redundancies between 2D frames enable to decrease the sampling rate compared to what is necessary to reconstruct individual 2D images; depending on the actual acquisition device, the saved measurements could then be re-allocated in order to improve the time resolution of the system.

The algebraic consequence of the memoryless measurement hypothesis is that the operator $\boldsymbol{\Phi}$ is block-diagonal. In [15], the authors demonstrate that restricted isometry inequalities – which ensure that the minimizer of either (1) or (2) is actually a good estimate of the original signal – could be established for block-diagonal $\boldsymbol{\Phi}$ operators when the Φ_t are random matrices. To prove this result, they impose an additional hypothesis on the class of signals $\boldsymbol{x}$ they are working with, namely that the energy $\left\|x_t\right\|_2^2$ of each frame is proportional to the number of measurements M_t allocated to the corresponding sensing operator Φ_t. Based on some empirical observations, we believe that such kind of result could also be established for other types of blocks Φ_t, such as randomly subsampled Fourier transforms.

From a practical viewpoint, assuming that the frame energy is almost constant over time (therefore using the same number of measurements for each frame) leads to satisfactory results (see sec. 3).

1.3. Previous works on designing sparsifying transform adapted to video signals

For 2D natural images, it is well-known that $\left\|\Psi\,\cdot\right\|_1$ with Ψ a 2D wavelet transform, or the total variation $\left\|\cdot\right\|_{\mathrm{TV}}$ can be used as the regularization term in (1). Each energy term depends on the actual image model: total variation is best suited for piecewise constant images (cartoon model), wavelets for piecewise regular images,

[1]In this paper, bold symbols denote 2D+T (or 2D+T related) entities, while regular font is reserved to objects with no temporal dimension.

curvelets for piecewise regular images with discontinuities along smooth edges, etc. In particular, for 2D data issued from biological imaging set-ups, total variation was shown to be suitable for sparsity enforcement and CS reconstruction in [12].

However, few results have been established so far for joint reconstructions of time-correlated 2D images (2D+T data). In [17], the authors propose to use a 3D wavelet basis for $\mathbf{\Psi}$; although it is a natural generalization of the 2D case, this approach does not take into account the fact that the objects appearing in a given 2D+T sequence might have very anisotropic spatio-temporal shape, while wavelets are best suited for isotropic objects.

In [14], the authors introduce a multi-scale video reconstruction framework, which relies on the idea of increasingly refining the spatial scale of the estimated signal: at each step, the algorithm exploits information obtained from coarser estimates to reduce the temporal redundancies and to estimate motion. However, although presenting some promising results, this method requires to adapt the measurement protocol in order to get some information about the coarse versions of the signal. Such modification is possible with the single-pixel camera, but cannot be extended to other CS acquisition device.

In [11], the authors propose to perform a joint reconstruction of sequences of K-consecutive frames (where $K \geq 2$ is a predefined parameter) in the following way: given a basis $\Psi \in \mathbb{R}^N$ in which each frame has sparse or nearly sparse representation, they define the following NK-square matrices:

$$\boldsymbol{B}_K = \begin{bmatrix} \Psi & & & \\ \vdots & \ddots & & \\ \Psi & & \Psi & \\ \Psi & & & \Psi \end{bmatrix} \quad \boldsymbol{C}_K = \begin{bmatrix} \Psi & & & \\ \vdots & \ddots & & \\ \Psi & \dots & \Psi & \\ \Psi & \dots & \Psi & \Psi \end{bmatrix} \tag{4}$$

Then, they propose to use either $\boldsymbol{B}_K$ or $\boldsymbol{C}_K$ as the dictionary in a l_1-synthesis scenario. The underlying idea is to exploit the temporal redundancy in the video sequence by reconstructing the difference between frames instead of the frames themselves; reconstruction artefacts induced by these methods will be discussed in sec. 3.

Other approaches using iterative reconstructions together with motion estimation and motion compensation heuristics (see [13]) exist but were not evaluated here.

2. VIDEO RECONSTRUCTION THROUGH 3D-TV MINIMIZATION

2.1. Three-dimensional total variation

As suggested by [11], considering frame-to-frame differences could be an interesting starting point to exploit temporal redundancies. However, one should notice that the significant coefficients are not randomly distributed in a typical consecutive frame difference. Indeed, if x_t and x_{t+1} are two consecutive frames in a video sequence, then the coefficients of $x_{t+1} - x_t$ with large magnitudes are mostly located close to the edges of x_t and x_{t+1}.

To enforce this property, we propose to use the three dimensional total variation as a regularization term in the reconstruction problem (1). 3D total variation (3D-TV) is defined as:

$$\|\boldsymbol{x}\|_{\mathbf{TV}} = \sum_P \sqrt{(\boldsymbol{D_h x})(P)^2 + (\boldsymbol{D_v x})(P)^2 + (\boldsymbol{D_t x})(P)^2} \tag{5}$$

where P visits every pixels of every frames, and $\boldsymbol{D_h}$, $\boldsymbol{D_v}$ and $\boldsymbol{D_t}$ stand for the discrete derivative operators respectively in the horizontal, vertical and temporal directions. The reason why $\|\cdot\|_{\mathbf{TV}}$ favors that the non-zero coefficients of the temporal difference map cluster around the edges can be explained thanks to *block sparsity*.

Block sparsity is a notion introduced by [16, 10], to refine the prior made on the sparse representation γ of a signal of interest x. The idea proposed by the authors is to favor sparse representations with clustered structures; more precisely, given a partition $(\Omega_g)_{g\in G}$ of the set of index values of γ, the authors introduce the following mixed $l_{1,2}$-norm:

$$\|\gamma\|_{1,2} = \sum_{g\in G} \sqrt{\sum_{k\in\Omega_g} |\gamma(k)|^2} \tag{6}$$

Then, they demonstrate that replacing $\|\cdot\|_1$ with $\|\cdot\|_{1,2}$ in the CS reconstruction problems (1) or (2) results in block-sparse estimators, meaning that the non-zero coefficients of the estimator $\hat{\gamma}$ are clustered within a few subsets Ω_g.

Obviously, $\|\boldsymbol{x}\|_{\mathbf{TV}} = \|\boldsymbol{\nabla x}\|_{1,2}$ where $\boldsymbol{\nabla}$ is the discrete gradient operator, constructed by stacking $\boldsymbol{D_h}$, $\boldsymbol{D_v}$ and $\boldsymbol{D_t}$, and where each underlying subsets Ω_P contains the three directional derivatives computed at a given 2D+T coordinate P: $\boldsymbol{D_t}$ computes the temporal difference map, while $\boldsymbol{D_h}$ and $\boldsymbol{D_v}$ act as edge detectors; then, minimization of $\|\boldsymbol{x}\|_{\mathbf{TV}}$ favors the clustering of non-zero coefficients present in these three maps.

2.2. Mean background correction

There are some situations where the difference $x_{t+1} - x_t$ between two consecutive frames is not sparse at all, even if x_t and x_{t+1} are well-correlated. This includes the case when the global illumination of the observed scene changes over time.

To make 3D-TV regularization more robust with respect to this problem, we reformulate the reconstruction problem as follows:

$$\hat{x} = \underset{x\in\mathbb{R}^N}{\arg\min} \|\boldsymbol{x} - \boldsymbol{b}\|_{\mathbf{TV}} \text{ s.t. } \|\boldsymbol{\Phi x} - \boldsymbol{y}\|_2 \leq \epsilon \tag{7}$$

where $\boldsymbol{b}$ is a sequence in which every pixel of a given frame t is set to the same value m_t, equal to the mean value of the corresponding frame x_t in the original signal.

The sequence $\boldsymbol{b}$, or equivalently the mean value m_t of each frame, has to be estimated prior to the resolution of (7) from the vector of observations $\boldsymbol{y}$. Obviously, $m_t = \frac{1}{N}\langle a, x_t\rangle$ where a is the constant vector $\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}$. Then, if the measurement operators Φ_t contains a row proportional to a, the values m_t can directly be read from the vector of observations $\boldsymbol{y}$; this is for example the case when the Φ_t are randomly subsampled Fourier transforms for which the sampling pattern is designed such that the DC component is always sampled. If Φ_t does not contain a row proportional to a, m_t can still be estimated using the framework developed in [6]; according the results presented in this paper, m_t, being a linear function of the signal of interest, can be estimated by:

$$\hat{m}_t = \frac{1}{N}\langle \Phi_t a, y_t\rangle \approx \frac{1}{N}\langle a, x_t\rangle \tag{8}$$

Finally, one should observe that the optimization scheme (7) is actually equivalent to the usual CS optimization problem with TV regularization, up to the variable change $\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{b}$. Therefore, it can be solved in practice with the usual dedicated CS solvers.

3. SIMULATION RESULTS

3.1. Comparison with other methods

We compared the proposed 3D-TV based methods with other existing algorithms, including:

Reconstruction method	PSNR (dB)		
	Amiba	*Foreman*	*Disks*
Frame-by-frame, TV regul.	42.5	16.1	26.6
Frame-by-frame, DB4 regul.	38.3	11.8	15.5
3D total variation	46.8	27.9	22.0
3D TV with background estim.	46.8	27.3	38.9
3D Haar regul. [17]	45.2	23.0	18.6
3D DB4 regul. [17]	45.3	21.3	15.4
$\boldsymbol{B}_4$ with DB4 regul. [11]	30.7	14.4	17.2
$\boldsymbol{C}_4$ with DB4 regul. [11]	43.8	17.7	18.1
$\boldsymbol{B}_{20}$ with DB4 regul. [11]	43.0	20.9	17.9
$\boldsymbol{C}_{20}$ with DB4 regul. [11]	45.8	23.6	18.2

Fig. 1. Mean square error (expressed as PSNR) between original sequences *amiba*, *foreman* and *disks*, and their estimators.

- l_1-analysis using a 3D-wavelet transform (see [17]), using the Haar wavelet (as suggested by the authors) and the Dauchechies-4 orthogonal wavelet;
- l_1-synthesis using the $\boldsymbol{B}_K$ and $\boldsymbol{C}_K$ dictionaries (see [11]), using a block size of $K = 4$ or $K = 20$ frames, and a Dauchechies-4 wavelet transform as the 2D dictionary.

To assess the improvement offered by 3D reconstruction methods thanks to temporal redundancies, we also provide the results obtained with frame-by-frame reconstruction, using either TV or Dauchechies-4 wavelet regularization.

For each video sequence, a vector $\boldsymbol{y}$ of observations was measured in the Fourier domain, using a random uniform sampling strategy with Hermitian symmetry; the DC component was also always measured. The noise parameter ϵ was tuned by hand, with the same value for every reconstruction method. Simulations were run using Matlab® and the NESTA optimization toolbox [1]. We extended this toolbox to deal with 3D total variation; such a modification is quite straightforward as NESTA already supports 2D total variation minimization.

We assessed the reconstruction fidelity of the algorithms for each test sequence by measuring the peak signal-to-noise ratio (PSNR) between the input and the reconstructions (fig. 1); visual qualitative evaluation of the artefacts was also performed (figs. 2 to 4). We present the results obtained for three test video sequences:

- *Amiba*, sized $256 \times 256 \times 80$[2], which is a microscopy sequence of moving and deforming amiba cells;
- *Foreman*, sized $288 \times 352 \times 80$, which represents a talking person;
- *Disks*, sized $256 \times 256 \times 80$, which is a synthetic sequence representing moving disks with random gray levels, sizes and speeds. We designed this synthetic sequence so that it breaks the underlying model corresponding to 3D total variation regularization; in particular, the gray level of the background oscillates quickly between two values, simulating rapid variations of the global illumination.

For the simulation results presented here, we used a sampling rate of 10% for both *Amiba* and *Disks*, and 20% for *Foreman*.

3.2. Data fidelity and reconstruction artefacts

In terms of PSNR, the proposed methods obtain the best reconstruction results, although the improvement over the other best performing methods ($\boldsymbol{C}_{20}$ [11] or 3D-wavelet [17] regularizations) is not

[2]height $\times$ width $\times$ number of frames

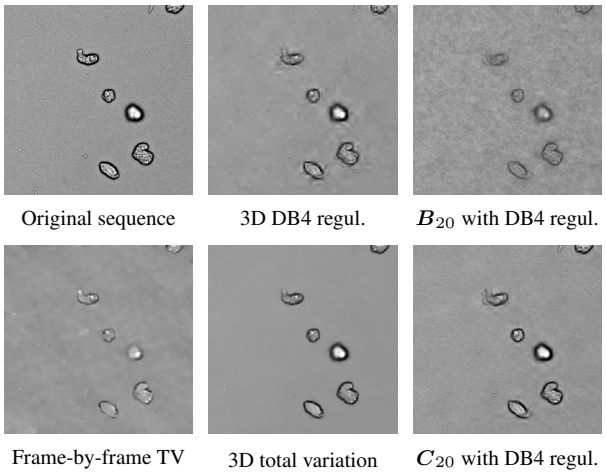

Fig. 2. Frame $t = 50$ in various estimators of the sequence *amiba*.

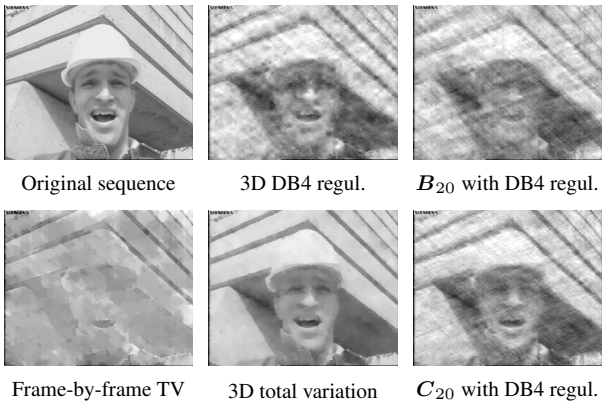

Fig. 3. Frame $t = 23$ in various estimators of the sequence *foreman*.

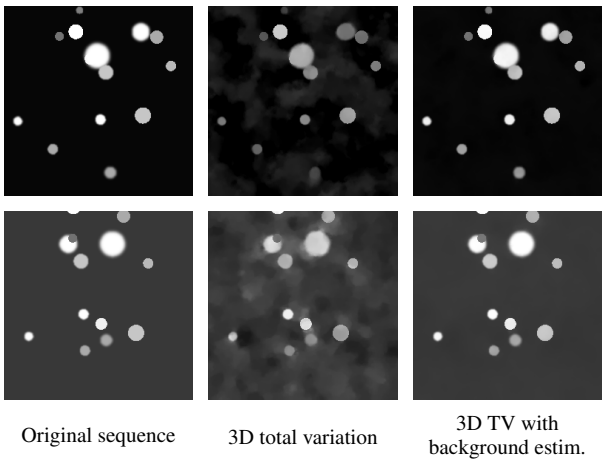

Fig. 4. Frame $t = 17$ (top row) and $t = 44$ (bottom row) obtained with the two proposed methods for the sequence *disks*.

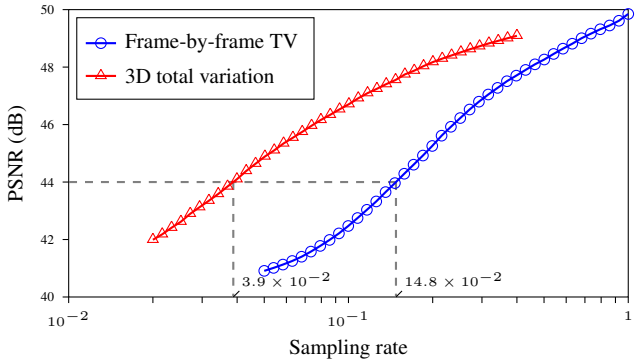

Fig. 5. Trade-off curves between sampling rate and reconstruction error for the *amiba* sequence, depending on the reconstruction method. To achieve reconstruction with a given error bound, the 3D TV regularization method needs three to four times less measurements than its 2D frame-by-frame counterpart.

dramatic in most cases (1.5 dB for *amiba*, about 4 dB for *foreman*). However, this measurement does not reflect the gain in terms of visual perception brought by the two 3D-TV based methods.

Indeed, compared to the wavelet-based regularization methods, 3D-TV tends to produce sequences with very sharp edges, without the oscillatory patterns typically present close to the edges in 3D wavelet reconstructed sequences. 3D-TV reconstructions also do not have the typical problems encountered with $\boldsymbol{B}_K$ and $\boldsymbol{C}_K$ estimators:

- $\boldsymbol{B}_K$ dictionaries tend to produce estimators where all the K frames belonging to a given block are very similar from one to each other (the gray level of a given pixel is almost piecewise constant over time), resulting in a jerky effect.
- $\boldsymbol{C}_K$ reconstructed sequences display precognition and trailing artefacts, meaning that the reconstructed frame corresponding to time t contains some piece of data belonging to the original frames $t+1$, $t-1$, $t+2$, $t-2$, etc. This is particularly noticeable close to the moving objects.

Finally, for most sequences, the simple 3D total variation estimator is very similar to its 3D-TV with background estimation counterpart, both in terms of PSNR and visual quality. The only exception is the *disks* sequence, which was designed on purpose to challenge the 3D-TV reconstruction: since the difference between two consecutive frames is non-zero at almost every pixel, the hypothesis on which the 3D-TV estimator[3] relies does not hold. Using the 3D-TV regularization term with background estimation tackles this issue, leading to a result almost identical to the original in the case of the *disks* sequence (cf. fig. 4).

3.3. Gain over frame-by-frame reconstruction

To quantify the gain provided by the 3D-TV reconstruction methods over simple frame-by-frame reconstructions, we measured the evolution of the reconstruction error as a fonction of the sampling rate on our test sequences (see fig. 5). Then, we compared the sampling rates corresponding to a given level of fidelity of the estimator to the original data, measured as a mean square error (expressed as PSNR).

In the case of our test sequences, we observed that the sampling rate corresponding to frame-by-frame TV reconstruction is generally 3 to 4 times bigger than the one corresponding 3D-TV reconstruction for a given value of the PSNR; this ratio tends to decrease when PSNR increases. One should mention also that this result does not depend on whether 3D-TV reconstruction with background correction or 3D-TV alone is considered, except in the case of the *disks* sequence, for which 3D-TV reconstruction alone completely fails.

[3] as well as many other estimators, especially those using the $\boldsymbol{B}_K$ and $\boldsymbol{C}_K$ dictionaries

4. CONCLUSION

In this paper, we presented a new framework for video reconstruction from frame-by-frame 2D CS measurements, based on the use of 3D total variation as the regularization function in the l_1-reconstruction problem. More precisely, we proposed two reconstruction schemes: one based on 3D-TV alone, which succeeds in reconstructing most signals, and one combining 3D-TV with estimation of the mean background value in each frame, which produces better reconstructions in the case of difficult sequences.

We compared these schemes to existing reconstruction methods, and showed that 3D-TV regularization outputs estimators with better qualitative properties, especially sharper edges and fewer motion artefacts. Finally, we demonstrated empirically that the number of measurements needed to reach a given reconstruction fidelity with our methods is 3 to 4 times smaller than what is required in the case of frame-by-frame reconstruction.

5. REFERENCES

[1] S. Becker, J. Bobin, and E. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[2] E. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2010.

[3] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, June 2007.

[4] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, Dec. 2005.

[5] E. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, Dec. 2006.

[6] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk. Signal processing with compressive measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, Apr. 2010.

[7] T. T. Do, L. Gan, N. H. Nguyen, and T. D. Tran. Fast and Efficient Compressive Sensing Using Structurally Random Matrices. *IEEE Transactions on Signal Processing*, 60(1):139–154, 2012.

[8] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, Apr. 2006.

[9] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23:947–968, 2007.

[10] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[11] R. F. Marcia and R. M. Willett. Compressive coded aperture video reconstruction. In *European Signal Processing Conference*, 2008.

[12] M. Marim, E. Angelini, and J.-C. Olivo-Marin. A compressed sensing approach for biological microscopic image processing. In *International Symposium on Biomedical Imaging*, pages 1374–1377. IEEE, 2009.

[13] S. Mun and J. E. Fowler. Residual reconstruction for block-based compressed sensing of video. In *Data Compression Conference*, pages 183–192, 2011.

[14] J. Y. Park and M. B. Wakin. A multiscale framework for compressive sensing of video. In *Picture Coding Symposium*, 2009.

[15] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin. Concentration of measure for block diagonal matrices with applications to compressive signal processing. *IEEE Transactions on Signal Processing*, 59(12):5859–5875, 2011.

[16] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.

[17] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk. Compressive imaging for video representation and coding. In *Picture Coding Symposium*, 2006.