Depth Sensing Using Active Coherent Illumination

Boufounos, P.T.

TR2012-020 March 2012

Abstract

We examine the use of active coherent sensingan increasingly available technology for sensing the depth of scenes. A scene is a sparse signal but also exhibits significant structure which cannot be exploited using standard sparse recovery algorithms. Instead, inspired by the model-based compressive sensing literature we develop a scene model that incorporates occlusion constraints in recovering the depth map. Our model is computationally tractable; we develop a variation of the well-known model-based Compressive Sampling Matching Pursuit (CoSaMP) algorithm, and we demonstrate that our approach significantly improves reconstruction performance.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DEPTH SENSING USING ACTIVE COHERENT ILLUMINATION

Petros T. Boufounos

Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, *petrosb@merl.com*

ABSTRACT

We examine the use of active coherent sensing—an increasingly available technology—for sensing the depth of scenes. A scene is a sparse signal but also exhibits significant structure which cannot be exploited using standard sparse recovery algorithms. Instead, inspired by the model-based compressive sensing literature we develop a scene model that incorporates occlusion constraints in recovering the depth map. Our model is computationally tractable; we develop a variation of the well-known model-based Compressive Sampling Matching Pursuit (CoSaMP) algorithm, and we demonstrate that our approach significantly improves reconstruction performance.

Index Terms— Depth sensing, array processing, compressive sensing, signal model

1. INTRODUCTION

Over the years, advances in signal acquisition hardware, theory, and algorithms have significantly improved our ability to coherently capture and process a variety of signal modalities. This enables increasingly accurate distance measurements using technologies such as ultrasonic sensing [1] and millimeter wave (mmWave) radar [2]. Using non-penetrating coherent waves it is thus possible to measure the depth-map of a scene and accurately reconstruct it.

This paper develops a scene model and an accompanying algorithm to reconstruct the depth of a scene using measurements obtained from an active coherent sensing system. Our approach is inspired by recent work on model-based compressive sensing [3], which enables reconstruction using application-specific models. In contrast to simple sparsity models commonly used in such applications our model explicitly enforces occlusion constraints to produce valid depth maps and provide robustness.

Our work is related to recent advances in depth-sensing combining LIDAR and compressive sensing [4, 5]. These approaches use digital micromirror devices (DMDs) to spatially modulate light in time-of-flight sensing systems and exploit the sparsity of the depth-map in the reconstruction. In contrast, we exploit the coherency and the bandwidth of the sensing system and use a more elaborate scene model to sense and reconstruct the depth of the scene.

The next section provides a brief background on active coherent sensing and model-based compressive sensing, establishing the notation used in the remainder of this paper. Section 3 formulates our signal acquisition and scene models and describes our model-based reconstruction algorithm. Section 4 experimentally verifies our approach and Section 5 provides some discussion and concludes.

2. BACKGROUND

2.1. Active Sensor Arrays

Typical coherent active arrays consist of transmitting and receiving components. These, depending on the sensing modality and the available hardware, can be separate physical devices or the same transducer. Each transmitter transmits a pulse, which is reflected from the objects in the scene of interest and received by the receivers. A coherent receiver receives the complete waveform of the reflection, which is processed to recover the desired information from the scene of interest. This is in contrast to incoherent receivers, such as visible-light sensors, which lack the ability to capture the complete waveform of the reflection, only its time-averaged energy.

The ability to capture the reflected waveform allows coherent arrays to measure the time-of-flight of the transmitted pulse from the instance it is transmitted until the reflection is received. It is thus possible to estimate the distance and the position of the reflectors in the scene. Specifically, a transmitter s transmits a pulse $p_s(t)$ to the scene. The pulse is reflected by a reflector at distance d_s from the transmitter and received by a receiver at distance d_r from the reflector, delayed by $\tau_{sr} = (d_s + d_r)/c$, where c is the speed of the transmitted wave. Assuming the transmitter and the receiver are omnidirectional, the received signal is $y_r(t) = xp_s(t - \tau_{sr})$, where x is the reflectivity of the reflector. Often, it is more convenient to express this delay in the frequency domain, i.e., $Y_r(\omega) = xe^{-j\omega\tau_{sr}}P_s(\omega)$, where uppercase denotes the Fourier transform. The propagation equation is linear, i.e., the principle of superposition can be used to describe the received signal from the transmission of multiple pulses and the reflection from multiple reflectors. Sensor directionality is straightforward to incorporate. We do not do so because it complicates the model without providing further intuition.

To describe a radar imaging system, we consider the scene of interest in its entirety. We discretize the scene using a grid of N points and represent the reflectivity of each point using x_n. Using τ_{srn} to denote the propagation delay from transmitter s to receiver r through gridpoint n, the propagation equation becomes

$$Y_r(\omega) = \sum_n \sum_s x_n e^{-j\omega\tau_{srn}} P_s(\omega). \tag{1}$$

In this model the reflectivity is assumed constant as a function of frequency. This assumption is partly necessary for the development in the remainder of the paper. While it is straightforward to model frequency-dependent reflectivity in (1), the approach developed later in this paper needs to be modified to incorporate the model. Section 5 discusses a modification that could accommodate this model.

2.2. Model Based Compressive Sensing

Recent advances in compressive sensing have enabled significant improvements in our ability to capture and reconstruct signals at the rate of their complexity rather than the rate of the ambient space in which the signal lies. This is achieved using a signal model. Standard compressive sensing formulations assume the signal is sparse in some basis. The sparsity model, enforced at the reconstruction algorithm, resolves the ambiguities in the underdetermined system

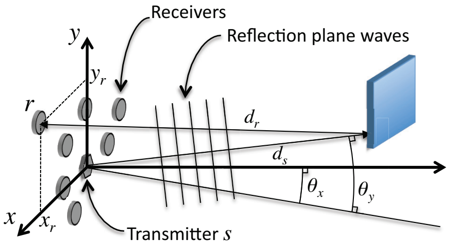

Fig. 1. Array geometry in the far-field approximation

arising from acquiring the signal at a rate lower than the ambient signal dimension.

Signal models other than sparsity can also be used to reduce the sampling requirements and improve the reconstruction performance. Manifolds, group sparsity, joint sparsity, and fusion frame sparsity models are such examples [6–9]. A large number of these models can be described by assuming the signal belongs to a union of subspaces, a more general model with well-studied reconstruction algorithms and recovery conditions [3, 10–12]. Standard signal sparsity is also a special case of a union of subspaces model.

Typically, the union of subspaces model is used in recovering a signal $\mathbf{x}$ from measurements $\mathbf{y}$ acquired using a linear system

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{2}$$

where $\mathbf{A}$ describes the acquisition system, usually under-determined. The signal can be recovered, under certain conditions on $\mathbf{A}$, by determining a vector $\widehat{\mathbf{x}}$ which belongs in the union of subspaces out of the ones that can explain the measurements. Typical greedy algorithms, such as the model based Compressive Sampling Matching Pursuit (CoSaMP) and model-based Iterative Hard Thresholding (IHT) [3] generalize their non-model-based counterparts [13, 14], and attempt to optimize

$$\widehat{\mathbf{x}} = \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2, \text{ s.t. } \mathbf{x} \in \mathcal{S}, \tag{3}$$

where $\mathcal{S}$ is the space of signals admissible to the model.

While there is a large variation, most greedy algorithms iterate between two basic steps. First, they identify a candidate support for the signal of interest and, second, they attempt to invert the system over that support. Their model-based counterparts modify the support identification step, in accordance to the support model for the signal of interest. Sufficient conditions for this modification to work are described in [3].

3. DEPTH SENSING

3.1. Signal Acquisition Model

To measure the depth map of a scene we consider an active coherent sensing array, as described in Sec. 2.1. For simplicity in the exposition, we assume one transmitter illuminating the scene and multiple receivers sensing the reflections. Multiple transmitters can be easily incorporated in a manner similar to [15]. To sense a 2- or 3-dimensional scene we use a linear or planar array, respectively. We are interested in the depth of each reflector in the scene, namely the distance of the reflector from the array. A critical assumption in our model is that the transmitted wave is not penetrating the objects of interest in the scene, only reflected by them. We also assume away secondary reflections. In other words, an object in the scene will obscure or completely hide the objects behind it.

As a special case we examine a planar array under the far-field approximation: the reflectors are far enough from the array such that the reflected waves are approximately planar as they arrive at the array. Figure 1 describes this approximation. Each point, n in the scene is described by its distance from the array, denoted using d_n, and its 2-dimensional angle with respect to the normal of the array plane, denoted $\boldsymbol{\theta}_n = [\theta_{xn}\ \theta_{yn}]^T$. We also define the transformation $\boldsymbol{\psi}_n = [\sin\theta_{xn}\ \sin\theta_{yn}]^T$. For a linear array sensing a 2-dimensional scene the only modification is that the angle is one dimensional. Henceforth, we refer to the angle $\boldsymbol{\psi}_n$ as the orientation coordinates, in contrast to the depth coordinate d_n.

We denote the location of each receiver r in the coordinate plane of the array using $\mathbf{v}_r = [x_r\ y_r]^T$. Thus, under the far-field approximation, the distance travelled by the signal from the transmitter to scene point n and back to sensor r is equal to $2d_n + x_r \sin\theta_{xn} + y_r \sin\theta_{yn} = 2d_n + \mathbf{v}_r^T\boldsymbol{\psi}_n$ and the corresponding delay is equal to $(2d_n + \mathbf{v}_r^T\boldsymbol{\psi}_n)/c$. Thus, the acquisition system is described by

$$Y_r(\omega) = \sum_n x_n e^{-j\omega(2d_n + \mathbf{v}_r^T\boldsymbol{\psi}_n)/c} P(\omega), \tag{4}$$

where we drop the dependence on s in (1) since we only use one transmitter. We compactly denote the linear system in (4) using

$$\mathbf{Y}(\omega) = \mathbf{A}(\omega)\mathbf{x}. \tag{5}$$

Of course, the system is typically broadband. We discretize the frequency space to $\omega_1, \ldots, \omega_F$ and describe the overall system using

$$\mathbf{Y} = \mathbf{A}\mathbf{x}, \mathbf{Y} = \begin{bmatrix} \mathbf{Y}(\omega_1) \\ \vdots \\ \mathbf{Y}(\omega_F) \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{A}(\omega_1) \\ \vdots \\ \mathbf{A}(\omega_F) \end{bmatrix} \tag{6}$$

We should also note that for implementation purposes, the adjoint of $\mathbf{A}$ is staightforward to compute from the adjoint of $\mathbf{A}(\omega)$:

$$\mathbf{A}^H\mathbf{Y} = \sum_{f=1}^{F} \mathbf{A}^H(\omega_f)\mathbf{Y}(\omega_f). \tag{7}$$

In many cases of uniform or other structured arrays, efficient computation of $\mathbf{A}(\omega)$ and its adjoint is possible. In these cases, (7) is also more efficient than explicit computation of $\mathbf{A}^H$.

3.2. Scene Model

Since we assume the transmitted pulse does not penetrate the reflectors in the scene, all the scene points x_n in front or behind a visible reflector do not reflect and, therefore, their reflectivity is equal to 0. For example, in a planar array, two points m, n with coordinates $\boldsymbol{\psi}_m = \boldsymbol{\psi}_n$ and $d_m \neq d_n$ cannot both have non-zero reflectivity. In other words, if we partition the coefficient space in groups, where all the elements in the group have the same orientation coordinates, then each group is 1-sparse. If the depth map itself is not dense in number of objects, we can also impose further structure, namely that only K of the groups are active, i.e., that there are objects along only K of the orientation directions.

This is a union of subspaces model. Assuming the coordinate space is discritized to $N_x \times 4N_y$ orientation points and N_d depth points, the number of subspaces comprising the model is equal to

$$\binom{N_x N_y}{K} N_d^K \leq \left(\frac{eN_xN_yN_d}{K}\right)^K, \tag{8}$$

Algorithm 1 The modified model-based CoSaMP algorithm to enforce the model described in this paper.

```
1:  **initialize** Iteration count $l = 0$, Initial estimate $\widehat{\mathbf{x}}^l = \mathbf{0}$.
2:  **while** not converged **do**
3:      Increase iteration count: $l \leftarrow l + 1$
4:      Compute residual:
            $\mathbf{R}(\omega) = \mathbf{Y}(\omega) - \mathbf{A}(\omega)\widehat{\mathbf{x}}^{l-1}$, for all $\omega$
5:      Compute proxy (gradient): $\mathbf{p}^l = \sum_{\omega=\omega_1}^{\omega_F} \mathbf{A}^H(\omega)\mathbf{R}(\omega)$
6:      Identify support candidate:
            $\mathcal{T}^l = \text{supp}\left(\text{trunc}(\mathbf{p}^l, K)\right) \cup \text{supp}\left(\widehat{\mathbf{x}}^{l-1}\right)$
        {$\text{supp}(\cdot)$ determines the support set of a vector; $\text{trunc}(\mathbf{x}, K)$
        selects the $K$ coefficients of $\mathbf{x}$ according to our model and
        sets the remaining to zero, as described in Algorithm 2}
7:      Line search: Find $\tau$ to minimize
            $\sum_{\omega} \left\| \mathbf{Y}(\omega) - \mathbf{A}(\omega)\left(\widehat{\mathbf{x}}^{l-1} - \tau\mathbf{p}^l|_{\mathcal{T}}\right)\right\|_2^2$, s.t. $\mathbf{x}|_{\mathcal{T}^c} = 0$
8:      Form temporary estimate:
            $\mathbf{b}^l = \left(\widehat{\mathbf{x}}^{l-1} - \tau\mathbf{p}^l\right)\Big|_{\mathcal{T}}$
9:      Compute final support: $\mathcal{S}^l = \text{trunc}(\mathbf{b}^l, K)$
10:     Truncate and update estimate: $\widehat{\mathbf{x}}^l = \mathbf{b}^l\big|_{\mathcal{S}^l}$
11: **end while**
12: **return** Signal estimate $\widehat{\mathbf{x}}^l$
```

where $K = N_x N_y$ if the depth map is dense. We should note that the upper bound on the number of subspaces on the left hand side of (8) is the same as an unrestricted K-sparse scene model. This implies (a) that the upper bound is loose and (b) that certain theoretical guarantees that depend on the number of subspaces cannot be improved using that bound (e.g., see [3,11]). This, however does not mean that the model is not an improvement over simple sparsity. As we show in simulations, enforcing the model significantly improves the results over an unrestricted sparsity model.

3.3. Reconstruction Algorithm

To enforce the acquisition model during reconstruction we use a variation of model-based CoSaMP [3], described in Algorithm 1.

The first difference is the use of line-search [16] to reduce the cost in step 7 instead of a full least-squares minimization using the pseudoinverse, in a manner similar to [15]. This makes the algorithm significantly more efficient if the propagation matrix $\mathbf{A}(\omega)$ is only available in functional form and not explicitly. The modified algorithm requires the applications of $\mathbf{A}(\omega)$ and its adjoint, which is usually easily computable in array applications.

The second difference is the modification of the support identification steps 6 and 9, according to the requirements of the model. Specifically, the truncation operation $\text{trunc}(\mathbf{x})$ is described in Algorithm 2. The truncation first selects the single largest coefficient in each orientation and then selects the K largest of those. Thus, it ensures that only one coefficient in each orientation is ever selected, enforcing the occlusion constraint.

Algorithm 2 The truncation operator $\text{trunc}(\mathbf{x}, K)$.

```
1: **Input:** Data $\mathbf{x}$ and desired sparsity $K$.
2: Find maximum value along each orientation $\boldsymbol{\psi}$:
       $(\mathbf{u})_{\boldsymbol{\psi}} = \max_d(\mathbf{x})_{d,\boldsymbol{\psi}}$, $(\mathbf{l})_{\boldsymbol{\psi}} = \arg\max_d(\mathbf{x})_{d,\boldsymbol{\psi}}$
   {Vectors $\mathbf{u}$ and $\mathbf{l}$, indexed in $\boldsymbol{\psi}$ store the maximum values along
   each orientation and the corresponding distance, respectively}
3: Find the $K$ largest of the maximum values $\mathcal{S} = \text{supp}(\mathbf{u}|_K)$
   {$\mathbf{x}|_K$ keeps the $K$ coefficients of $\mathbf{x}$ with the largest magnitude
   and sets the remaining to zero.}
4: Compute index set $\mathcal{T}$, containing the support of the $K$ largest
   values $\mathcal{S}$ and the distance of the maximum along the corresponding
   direction, $\mathbf{l}|_{\mathcal{S}}$.
5: **return** Truncated vector $\mathbf{x}|_{\mathcal{T}}$
```

4. EXPERIMENTAL RESULTS

To validate our approach, we performed experiments on a simulated mmWave system operating at a 3GHz frequency bandwidth, centered at 76.5GHz. To better illustrate the issues with classical approaches, our experiments are performed on a 2-D scene using a linear array of 15 and 21 elements with 2m aperture size. The angle space ψ is discretized at a resolution of 0.01 (in the dimensions of the sine of the angle) for 201 total gridpoints. The scene has 1m maximum depth with 2cm grid resolution. An example of our simulations is shown in Figure 2.

The figure shows from left to right the sensed scene, the reconstructed scene using standard backprojection, the reconstructed scene using the standard CoSaMP algorithm and the reconstructed scene using our model. The top row demonstrates the results for a 21-element array and the bottom for a 15-element one. The results were consistent in a variety of experiments.

The top row shows a typical failure mode of standard sparsity-based compressive sensing that we attempt to fix using our model. As we can see from the backprojection results, there is significant ambiguity around the reflectors, due to the relatively high coherence of nearby pixels. These can confuse standard sparsity-based CS reconstruction algorithms. For example, Fig. 2(c) demonstrates how CoSaMP only picked up one of the reflectors, and the region around it. Even by reducing the desired sparsity of the reconstructed signal, performance does not improve. Figure 2(d) shows that the model can resolve these ambiguities and accurately reconstruct the signal.

Even in cases of significant ambiguities, as the example shown in the bottom row, the model significantly improves reconstruction performance. Specifically, the reduced number of array elements increases array ambiguities, as demonstrated by the significant horizontal blur of the backprojection reconstruction. As expected, simple sparse reconstruction fails to recover the scene. The model-based approach significantly improves the identification of the support of the depth-map, although it does not recover the signal perfectly. Subsequent processing such as total-variation based smoothing can also be used to further smooth the produced depth map.

5. DISCUSSION

Increasingly available coherent sensing technology can be used in depth sensing applications. For such applications, sparsity and scene models play a key role in improving the system performance. The model and corresponding algorithm we present in this paper is explicitly designed to take into account the particular structure of depth-maps.

Although we present in detail and in simulations only the far-field approximation formulation, the same model can be applied in

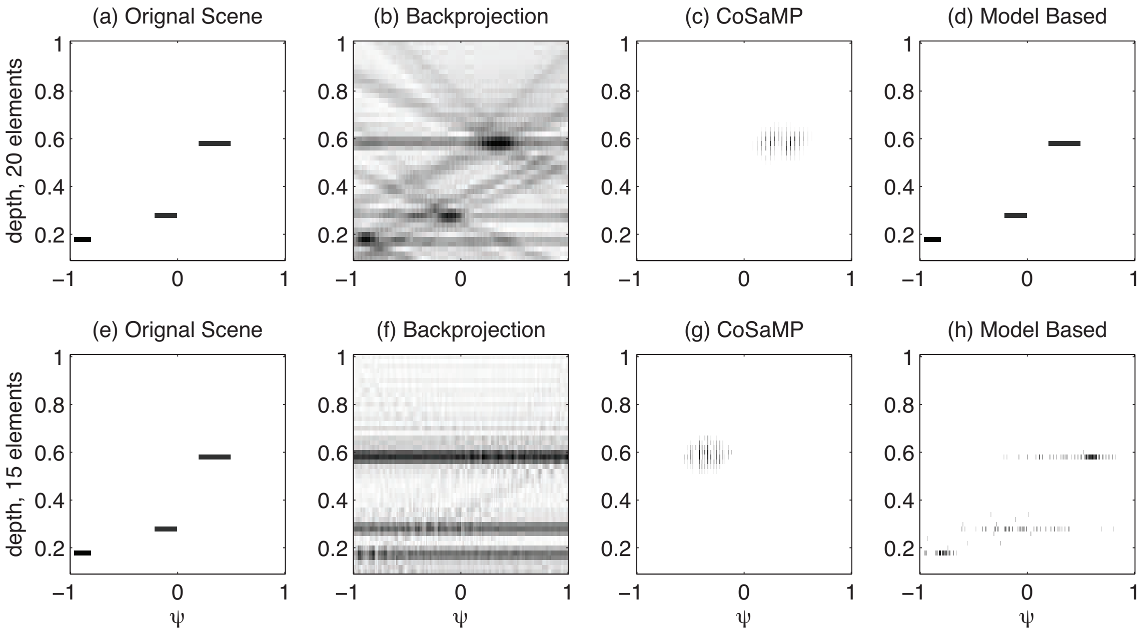

Fig. 2. Examples of simulation results. From left to right, the original scene, backprojection reconstruction, standard sparsity-based reconstruction using CoSaMP, and model-based reconstruction using model-based CoSaMP. Top row demonstrates results for a 21-element array and bottom row for a 15-element array.

near-field formulations. One complication in this case, not appropriately handled by the model, is partial occlusions, i.e., cases where an object in the background is visible only to a few of the sensors. In the far-field approximation this can never occur. Managing partial occlusions is not straightforward and requires further research.

Another case we do not examine is the case of multiple transmitters. There are several ways to incorporate multiple transmitters in our sensing model, for example by having each transmitter transmit a pulse separately from the others, recording the reflections, and forming an enlarged reconstruction problem. Alternatively, all transmitters can pulse simultaneously using different pulses, either orthogonal or randomly generated, similar to [1].

Finally, we should note that we can further improve reconstruction performance by additional modeling of the smoothness of the objects in the scene, or by further grouping the orientation directions. Such modeling could further improve results, but could make the model overly sensitive to scene characteristics. Furthermore, we can incorporate frequency-dependent scene reflectivity using joint-sparsity models, in a manner similar to [15].

6. REFERENCES

[1] P. Boufounos, "Compressive sensing for over-the-air ultrasound," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP),*, may 2011, pp. 5972 –5975.

[2] A. Accardi, *Generating Pictures from Waves: Aspects of Image Formation*, Doctoral thesis, MIT, Cambridge, MA, 2010.

[3] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[4] G.A. Howland, P.B. Dixon, and J.C. Howell, "Photon-counting compressive sensing laser radar for 3d imaging," *Applied Optics*, vol. 50, no. 31, pp. 5917–5920, 2011.

[5] A. Kirmani, A. Colaço, F.N.C. Wong, and V.K. Goyal, "Exploiting sparsity in time-of-flight range acquisition using a single time-resolved sensor," *Optics Express*, vol. 19, no. 22, pp. 21485–21507, 2011.

[6] R.G. Baraniuk and M.B. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, no. 1, pp. 51–77, 2009.

[7] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075 –3085, aug. 2009.

[8] Y.C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3042 –3054, 2010.

[9] P. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," *IEEE Trans. Info. Theory*, vol. 57, no. 6, pp. 3864–3876, 2011.

[10] Y.M. Lu and M.N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2334–2345, 2008.

[11] T. Blumensath and M.E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[12] Y.C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[13] D. Needell and J.A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[14] T. Blumensath and M.E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[15] P. Boufounos, P. Smaragdis, and R. Bhiksha, "Joint sparsity models for wideband array processing," in *Proc. SPIE, Wavelets and Sparsity XIV*, San Diego, CA, 21–25 August 2011.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.