MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Coded Aperture Compressive 3-D LIDAR

Kadambi, A.; Boufounos, P.T.

TR2015-028 April 2015

Abstract

Continuous improvement in optical sensing components, as well as recent advances in signal acquisition theory provide a great opportunity to reduce the cost and enhance the capabilities of depth sensing systems. In this paper we propose a new depth sensing architecture that exploits a fixed coded aperture to significantly reduce the number of sensors compared to conventional systems. We further develop a modeling and reconstruction framework, based on model-based compressed sensing, which characterizes a large variety of depth sensing systems. Our experiments demonstrate that it is possible to reduce the number of sensors by more than 85%, with negligible reduction on the sensing quality.

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

CODED APERTURE COMPRESSIVE 3-D LIDAR

Achuta Kadambi,[1] *Petros T. Boufounos*[2]

[1] Media Laboratory, Massachusetts Institute of Technology (MIT), Cambridge MA, 02139, *achoo@mit.edu*
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge MA, 02139, *petrosb@merl.com*

ABSTRACT

Continuous improvement in optical sensing components, as well as recent advances in signal acquisition theory provide a great opportunity to reduce the cost and enhance the capabilities of depth sensing systems. In this paper we propose a new depth sensing architecture that exploits a fixed coded aperture to significantly reduce the number of sensors compared to conventional systems. We further develop a modeling and reconstruction framework, based on model-based compressed sensing, which characterizes a large variety of depth sensing systems. Our experiments demonstrate that it is possible to reduce the number of sensors by more than 85%, with negligible reduction on the sensing quality.

Index Terms— 3-D imaging, LIDAR, time of flight, compressed sensing, computational imaging.

1. INTRODUCTION

Recent advances in signal processing theory and acquisition hardware have enabled significant improvements and cost reduction in depth sensing applications. Furthermore, recently emerging applications, such as autonomous navigation, mapping, and home entertainment, have put pressure on this area and increased the demand for inexpensive and high quality depth sensing.

In this paper we fundamentally re-examine the depth sensing problem given recent advances in model-based compressive sensing (CS), the reduced cost of computation, and the availability of off-the shelf hardware. Our main contributions are the following:

- a single-shot compressive hardware architecture for laser-based time-of-flight (TOF) depth sensors,
- a very general system model that characterizes a large number of depth sensing system architectures, and
- a model-based CS reconstruction algorithm that exploits the low total variation (TV) of depth scenes to improve reconstruction.

Conventional high-resolution, high frame-rate LIDAR systems typically use an expensive array of precision TOF sensors and illuminate the whole scene with a single laser pulse [1]. Alternatively, at the expense of reduced frame-rate, the laser scans the scene. A smaller sensor array, which might also scan the scene, acquires the reflection. The resulting system significantly lowers cost, but requires the use of mechanical components, which can be prone to failure.

Instead, compressive approaches can exploit significant gains in computational power, thanks to Moore's law, to reduce the sensing cost. Widespread computation enables elaborate signal models and reconstruction algorithms, which, in turn, allow for reduced sensor complexity. For example, a compressive depth sensing architecture was recently proposed [2–4], using a single sensor combined with a spatial light modulator and multiple pulses illuminating the whole scene. The spatial light modulator is used to implement a variable coded aperture, the code of which is changing with every pulse—which restricts this approach to static scenes.

Our hardware architecture is similar to [2–4], but we use a fixed coded aperture—a very inexpensive component compared to a spatial light modulator—and more than one sensor. Since the code does not change, our architecture only requires a single pulse transmission per frame, which may or may not carry a code. Thus, we can achieve frame-rates equivalent to much more expensive single-pulse systems, but at a significantly lower cost. Specifically, using our approach we are able to reduce the number of time of flight sensors by 85% compared to the full resolution of the acquired image, depending on the complexity of the pulse and the scene.

Algorithmically, our approach is based on recent developments on model-based CS [5–7] and signal models inspired by the ones in [8]. We demonstrate that, the signal model in [8] combined under specific conditions with the constrained earth mover's distance (EMD) model discussed in [6, 7] results to a constrained total variation (TV) model, suitable for depth images. However, this approach can only model two-dimensional (2-D) scenes, which correspond to 1-D depth maps. Thus we exploit a graph cuts formulation of the resulting problem [9], which enables the extension to 3-D scenes, and, correspondingly, to 2-D depth maps.

Our model is very different from the reconstruction framework in [4], termed CoDAC. The latter uses a two-step process, first finding a discrete set of depths present in the scene and then reconstructing the reflectivity of the scene at each depth. The depth-finding step exploits a finite-rate-of innovation (FRI) model, which essentially requires that the scenes have planar reflectors, i.e., that the scene is piece-wise planar. In addition, current FRI models require very specific pulse shapes, which precludes the use of coded pulses [10]. Our model is also more general, characterizing a wider variety of possible compressive LIDAR architectures. For example, while our model can be used with the hardware architecture of [2–4], the converse is not true; the use of the two-stage FRI model would require small architecture modifications in our system.

The next section provides an overview of CS, compressive imaging and model-based CS. Section 3 describes the proposed system and develops the signal and system models. The reconstruction algorithms are is described in Sec. 4. Section 5 provides simulation results validating our approach.

2. BACKGROUND

2.1. Compressive Sensing an Imaging

Compressive Sensing (CS) [11, 12] has emerged as a powerful sensing framework, demonstrating that signals can be acquired using

AK was an intern at MERL when he performed this work. The authors would like to thank Srikumar Ramalingam for helpful discussions and pointers on the state of the art in graph cuts.

much fewer linear measurements than their dimension implies. To reduce the acquisition rate, CS reconstruction algorithms exploit the structure of acquired signals. To capture structure, the most commonly used signal model is sparsity: the signal comprises a linear combination of very few atoms selected from a basis or a dictionary. A few other models, such as signals lying on low-dimensional manifolds and signals with low total variation (TV) have also been shown to be suitable for compressive acquisition [13, 14].

A CS-based acquisition system can be modeled as

$$\mathbf{r} = \mathbf{A}(\mathbf{s}), \tag{1}$$

where $\mathbf{A}(\cdot)$ is a linear function, $\mathbf{s}$ belongs in some appropriate signal space, and $\mathbf{r}$ belongs in the measurement space. The latter space has much lower dimension than the former. A number of possible properties of $\mathbf{A}(\cdot)$—such as low coherence, the Restricted Isometry Property, or others, depending on the model—guarantee that reconstruction is possible using an appropriate algorithm [11, 12].

Compressive Sensing has been proven very successful in imaging systems, in which, typically, the signal $\mathbf{s}$ to be acquired is a 2-D image in $\mathbb{R}^{N_x \times N_y}$. Using compressive approaches, it has been shown that images can be acquired with measurements as few as 10% of the number of pixels $N_x N_y$. These gains are not as relevant in conventional visible-light imaging, where CCD and CMOS sensor technology has made measurements extremely inexpensive. However, this approach has had significant impact in other modalities, such as medical imaging, low-light imaging, hyperspectral imaging, and depth sensing [3, 4, 8, 15–18].

2.2. Model-Based Compressive Sensing

The recently developed model-based CS framework provides a general approach to developing a large number of signal models and characterizing their suitability for CS acquisition [5]. Models under this framework are created by imposing restrictions on the signal support. A fundamental operation is the projection of a general signal to the set of signals that satisfy the model's support restrictions. As long as such a projection can be computed, common greedy CS reconstruction algorithms, such as Compressive Sampling Matching Pursuit (CoSaMP) [19] and Iterative Hard Thresholding (IHT) [20], can be modified to reconstruct signals in the model. Furthermore, it has recently been shown that a pair of approximate projections with different approximation properties is sufficient to guarantee accurate reconstruction, instead of an exact projection [6, 7].

Furthermore, in [6, 7], a novel signal model is developed, motivated by signals in 2-D seismic imaging. A signal $\mathbf{s} \in \mathbb{R}^{N \times T}$ in this model is a matrix with N rows and T columns. Each row of the signal only has S non-zero entries, which should be spatially close to the S non-zero entries of the row above or below. This is enforced by restricting the EMD between the support of subsequent rows of the signal. The projection under this signal model is performed by solving a sequence of network flow problems.

As will become evident in Sec. 4.1, a restricted version of this model, with $S = 1$ is a very good model for LIDAR scenes. However, it requires extension of this model to 3-D volumes, namely signals $\mathbf{s} \in \mathbb{R}^{N_x \times N_y \times T}$ which, unfortunately, is not obvious. Section 4.1, develops such an extension for the restricted case of $S = 1$.

3. COMPRESSIVE LIDAR SENSING

3.1. System Architecture

Fundamentally, the LIDAR architecture we propose relies on classical time-of-flight (TOF) LIDAR principles: a pulse is transmitted to illuminate the scene and the delayed reflections from each object in the scene are acquired by the system and used to determine the depth of each object. Classical architectures, however, rely on carefully separating and sensing the reflections from each direction in order to determine the TOF and assign it to the appropriate place in the depth map. Instead, we exploit CS principles and intentionally mix reflections from all directions, in order to reduce the sampling burden. Thus, we rely on computational methods to separate the mixture and assign the correct depth value to each point in the scene.

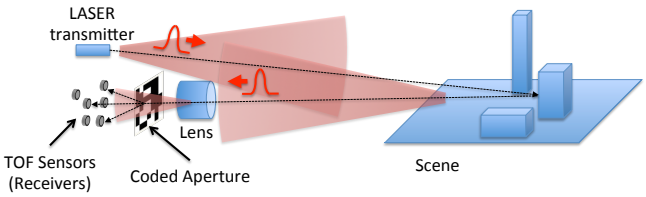

Fig. 1: Schematic of the LIDAR system architecture.

The system architecture is shown in Fig. 1. The scene is illuminated using a laser that transmits a pulse through a lens system. In contrast to many conventional architectures, the laser has a wide beamwidth, which illuminates the whole scene and is not scanned. The pulse may be a simple pulse, such as a Gaussian-shaped one, or a coded sequence of pulses.

The pulse travels through the medium, reflects on objects in the scene, and returns to the sensing system. The reflected delayed pulse is acquired though a lens system, a fixed coded aperture and a set of sensors behind the coded aperture. The coded aperture is typically implemented using a mask that blocks light according to a 0/1 pattern in the code. A coded aperture where the pattern can take continuous values in $[0, 1]$ is also physically possible.

The combination of the lens and the coded aperture mixes pulses received from each direction and projects them to the sensor plane [21]. Each location on the sensor plane receives a different mixture, determined by the aperture code. Thus, the signal recorded by each sensor is a mixture of the reflected pulses, as received at the sensor's location. These mixtures are used to reconstruct the scene.

3.2. Signal Model

To model the system we start with a single reflecting scene point at distance d from the sensing plane, and we assume no coded aperture. For convenience we assume the laser is located at the same position as the sensing plane, but this assumption can be trivially relaxed. Thus, the distance the pulse travels from the laser to the reflecting scene point and back to the sensor is $2d$, corresponding to a pulse delay, i.e., TOF, $\tau = 2d/c$, where c is the speed of light in the medium. Through this correspondence, time is equivalent to distance from the sensor plane, i.e., depth, and we will often use them interchangeably in the remainder of this paper.

We consider a 3-D scene, comprising of two spatial directions, transverse to the sensor plane, and one depth direction, which we also refer to as time or delay, and is perpendicular to the sensor plane. Thus, the scene is a function $s_{x,y,t}$, to be acquired, where s represents the reflectivity of the scene at point (x, y, t). If there is no reflection from a point, the corresponding reflectivity is zero.

We assume Lambertian surfaces with no transparency, which implies that, for any (x, y) pair, there is only one depth t that has non-zero reflectivity. In other words, if there is a reflection from (x, y, t), then there is no reflector in-between that location and the sensing plane. Furthermore, the light does not reach any reflectors

behind that point for the same (x, y) pair, and, therefore, their reflectivity is also zero. We refer to this constraint, first introduced in [8], in the context of coherent sensing systems, as the *depth constraint* on s. A valid scene signal s should satisfy this depth constraint.

Given a depth map $d_{x,y}$, representing the depth of the scene at coordinates (x, y), and a reflectivity (albedo) map $a_{x,y}$ for the same coordinates, the scene is equal to

$$s_{x,y,t} = a_{x,y}\delta_{t-2d_{x,y}/c}, \tag{2}$$

where δ_t is the Dirac impulse function. From any scene satisfying the depth constraint, it is trivial to extract depth and albedo maps.

3.3. Acquisition Model

We use p_t to denote the transmitted pulse, which gets reflected by the scene. Assuming that a pinhole aperture is present, then the received reflection at location (x, y) is equal to

$$\widehat{r}_{x,y,t} = a_{x,y}p_{t-2d_{x,y}/c} = p_t \circledast_t s_{x,y,t}, \tag{3}$$

where $\circledast_t$ denotes linear convolution along the time direction. The addition of a coded aperture, with spatial code $m_{x,y}$ introduces a mixing of the received signal, which can be shown to be the spatial convolution of $\widehat{r}_{x,y,t}$ with the mask [21]. Thus, the received light flux at the sensor plane is equal to

$$\widetilde{r}_{x,y,t} = m_{x,y} \circledast_{x,y} \widehat{r}_{x,y,t} = m_{x,y} \circledast_{x,y} p_t \circledast_t s_{x,y,t}. \tag{4}$$

This signal is sensed by M sensors, indexed by $m = 1, \ldots, M$, each positioned at location (x_m, y_m). Each sensor samples the light flux at the sensor plane and samples $r_{m,t} = \widetilde{r}_{x_m,y_m,t}$ in time.

We discretize the scene to $\mathbf{s} \in \mathbb{R}^{N_x,N_y,N_t}$, where N_x, N_y is the number of spatial pixels to be acquired—specified as the desired resolution of the system—and N_t is the number of time samples. The discretization is such that each reflector can be assumed approximately flat and parallel to the sensor plane over the area of a pixel, such that the depth constraint is preserved in the discrete representation. Furthermore, the time is sampled at a rate higher than the pulse Nyquist rate. Similarly, we discretize the pulse, the coded aperture mask, and the received signals, such that the convolution with the pulse and the mask shape can be expressed in discrete-time.

The discretized received signal at sensor m, $\mathbf{r} \in \mathbb{R}^{M,N'_t}$ can then be computed as a sequence of linear transformations

$$\mathbf{r} = \mathbf{S}(\mathbf{M}(\mathbf{P}(\mathbf{s}))) = \mathbf{A}(\mathbf{s}), \tag{5}$$

where $\mathbf{P}$, $\mathbf{M}$ and $\mathbf{S}$ denote, respectively, the linear transformations due to the pulse, the mask of the coded aperture and the sampling of the optical field by the sensors. Their composition, $\mathbf{A}$, is the forward linear transformation mapping the scene $\mathbf{s}$ to the received signal $\mathbf{r}$.

3.4. Model Implementation

Efficient computation using the linear model in (5) is paramount in modern iterative reconstruction methods. To this end, the operators $\mathbf{P}$ and $\mathbf{M}$ can be efficiently implemented in discrete time using FFT-based convolution algorithms. Furthermore, $\mathbf{S}$ is trivial computationally since it simply selects the appropriate signals. The adjoint $\mathbf{A}^*$, necessary for most reconstruction algorithms, is also trivial to implement by composing the adjoint of each operator in the reverse order, i.e., using $\mathbf{A}^*(\mathbf{r}) = \mathbf{P}^*(\mathbf{M}^*(\mathbf{S}^*(\mathbf{r})))$.

The implementation can also exploit the separability of space and time operators in (5), and rearrange them to reduce complexity: $\mathbf{S}(\mathbf{M}(\mathbf{P}(\mathbf{s}))) = \mathbf{P}(\mathbf{S}(\mathbf{M}(\mathbf{s})))$. Thus, the pulse convolution $\mathbf{P}$ is applied to an $M \times N_t$-dimensional object, rather than a $N_x \times N_y \times N_t$-one, while the complexity of $\mathbf{S}$ and $\mathbf{M}$ does not change.

Finally, an efficient implementation can use a depth map/albedo representation, i.e., $\mathbf{d}$ and $\mathbf{a}$, for storage, thus significantly reducing memory requirements. The forward operator is straightforward to compute efficiently from such a representation, although the adjoint would require a temporary memory expansion to store the full $\mathbf{s}$.

4. DEPTH RECONSTRUCTION

4.1. Depth Scenes, Total Variation and Network Flows

To reconstruct a subsampled signal and provide robustness to noise, CS exploits the structure of the acquired signal. In particular, depth maps have been shown to have low TV norm [22, 23]. Signals with low TV norms are generally flat, with very few discontinuities and very few areas with small gradients.

The (ℓ_1) TV norm of a discrete map $\mathbf{d} \in \mathbb{R}^{N_x \times N_y}$, is defined as

$$\|\mathbf{d}\|_{\mathrm{TV}} = \|\nabla_x \mathbf{d}\|_1 + \|\nabla_y \mathbf{d}\|_1, \tag{6}$$

where $\nabla_x \mathbf{d}$ is the discrete gradient along direction x and $\|\cdot\|_1$ is the element-wise ℓ_1 norm of a matrix, i.e.,

$$\|\nabla_x \mathbf{d}\|_1 = \sum_{n_y=1}^{N} \sum_{n_x=1}^{N-1} \left| d_{n_x,n_y} - d_{n_x+1,n_y} \right|, \tag{7}$$

with the y direction similarly defined.

Given a scene satisfying the depth constraint described in Sec. 3.2, an additional constraint on the TV norm of the depth map is exactly equivalent to a constraint on the support of the non-zeros in $\mathbf{s}$. Specifically, two spatially adjacent non-zero coefficients of $\mathbf{s}$—i.e., with respect to the n_x and n_y coordinates—should also have similar depth—i.e., n_t coordinate—except for very few discontinuities. However, this one-to-one correspondence requires that $\mathbf{s}$ satisfies the depth constraint. A general, dense $\mathbf{s}$ does not; a projection to both the depth and the TV constraints is, thus, required for model-based CS algorithms, such as CoSaMP and IHT.

For a one-dimensional depth map problem, i.e. for $\mathbf{s} \in \mathbb{R}^{N_x,N_t}$ and $\mathbf{d} \in \mathbb{R}^{N_x}$, the projection in [6, 7] provides a solution. In this work, 2-D signals are considered, represented as a matrix. Each row of the matrix has only S non-zeros, and the support of those non-zeros from row to row changes very little, according to a predetermined constraint on the total EMD between supports. It is straightforward to show that with $S = 1$, the EMD constraint applied to $\mathbf{s}$ becomes a TV constraint on its support, i.e., on the depth map. The projection onto the constraint can be computed using a sequence of simple dynamic programs solving a network flow. Unfortunately this approach does not generalize to 2-D depth maps. For those we need a different formulation, described in the next section.

4.2. 2-D Total Variation and Graph Cuts

To generalize the projection to 3-D objects, i.e., 2-D depth maps, we use a graph cuts formulation. For an undirected, weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we consider the general graph cuts problem. That is, given a set of observations $\mathcal{X}$, the task is to assign each vertex $v \in \mathcal{V}$ a label $\mathbf{l}_v \in \mathcal{L}$ such that the joint labeling of all vertices, $\mathbf{l}$, minimizes an energy function between labels and observations $E(\mathbf{l}, \mathcal{X})$.

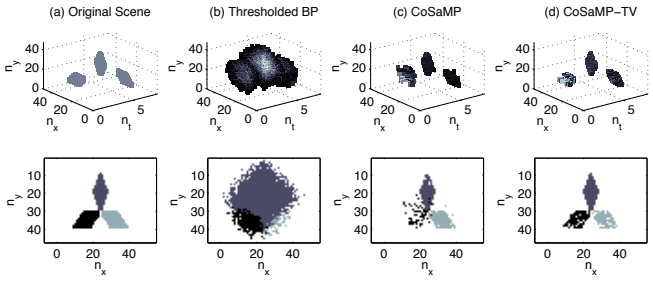

Fig. 2: Example of a (top) 3-D scene and (bottom) the corresponding depth map. From left to right, (a) original scene and the reconstruction results using (b) thresholded backprojection, (c) conventional sparsity models (CoSaMP) and (d) bounded TV reconstruction (CoSaMP-TV). Darker colors represent, respectively, lower amplitudes in the 3-D scene and closer objects in the depth map. Absence of color indicates no reflection.

In our depth sensing problem, we map each vertex to represent a spatial location $v = (n_x, n_y)$ of the scene, and each label to represents a discrete depth value $\mathbf{l}_v = \mathbf{d}_{n_x,n_y}$. Hence, the cardinality of sets $\mathcal{V}$ and $\mathcal{L}$ is $N_x N_y$ and N_t, respectively. We also map the set of observations $\mathcal{X}$ to the scene $\mathbf{s}$. Finally, we express the energy function as a sum of unary and pairwise terms

$$E(\mathbf{l}, \mathbf{s}) = \underbrace{-\sum_{v \in \mathcal{V}} \mathbf{s}^2_{v,\mathbf{l}_v}}_{\text{Unary}} + \underbrace{\sum_{\substack{v \in \mathcal{V} \\ u \in \mathcal{N}_v}} \lambda \left| \mathbf{l}_v - \mathbf{l}_u \right|}_{\text{Pairwise}} \tag{8}$$

$$= - \sum_{(n_x,n_y) \in \mathcal{V}} \mathbf{s}^2_{n_x,n_y,\mathbf{d}_{n_x,n_y}} + \sum_{\substack{(n_x,n_y) \in \mathcal{V} \\ (n'_x,n'_y) \in \mathcal{N}_{n_x,n_y}}} \lambda \left| \mathbf{d}_{n_x,n_y} - \mathbf{d}_{n'_x,n'_y} \right|, \tag{9}$$

where $\mathcal{N}_{n_x,n_y} = \{(n_{x+1}, n_y), (n_{x-1}, n_y), (n_x, n_{y+1}), (n_x, n_{y-1})\}$ is the neighborhood of $\mathcal{N}_{n_x,n_y}$ (i.e. $\mathcal{N}_v$ contains all vertices that are directly adjacent to $v = (n_x, n_y)$ in the graph).

The unary term is a fidelity term which uses the label—i.e., the depth value—to select the appropriate data point from the scene $\mathbf{s}$ and impose an ℓ_2 data penalty. The pairwise term imposes a smoothness constraint between the label of v and the label of vertices in the neighborhood set $\mathcal{N}_v$. Thus, the pairwise term from equation 8 is the ℓ_1 norm of the gradient of depth values, i.e., the TV norm of $\mathbf{d}$. Analogous to Rudin-Osher-Fatemi total-variation, the parameter λ weights the tradeoff between data fidelity and smoothness [9].

Solvers for minimizing (8) are, by now, widely available. Algorithmic techniques include alpha-expansion and alpha-beta swap [24, 25] as well as Boolean approaches [26, 27]. A survey on such algorithms can be found in [28]. In our experiments, we use the alpha-expansion technique from Boykov et. al. [24].

The truncation step of our model-based algorithm incorporates a K-term truncation of $\mathbf{s}$ by first optimizing (8) to obtain a candidate depth map $\mathbf{d}$, which corresponds to a candidate support set $(n_x, n_y, \mathbf{d}_{n_x,n_y})$. From this candidate support set the largest K components of $\mathbf{s}$ are selected to be used by the appropriate step of model-based CoSaMP or IHT. In contrast, conventional truncation in these algorithms just selects the K largest components of the data $\mathbf{s}$, not constrained by the graph cuts solution. Compared to conventional sparsity, our model-based thresholding produces a scene $\mathbf{s}$ that (a) satisfies the depth constraint and (b) has low depth TV. In the interest of space, we refer to [5–7] for more details on model-based algorithms and the role of truncation in their operation.

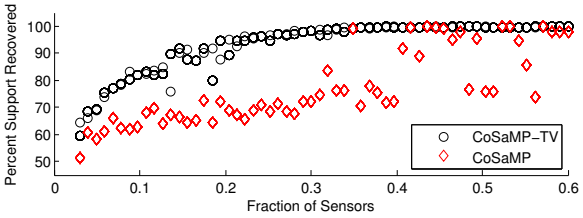

Fig. 3: Proportion of measurements (i.e. number of sensors) versus percentage of the ground truth support that is recovered for the simple sparsity model and the model-based approach.

5. EXPERIMENTS

To validate our model we performed simulations on a variety of scenes. In our experiments we illuminated, acquired and reconstructed scenes using our proposed system architecture and model-based CoSaMP reconstruction algorithm—referred to as CoSaMP-TV in the remainder. For comparison, we also reconstructed the scenes using thresholded backprojection and the CoSaMP algorithm with a standard sparsity model. Our experiments use a randomly generated binary mask and random sensor placement for the acquisition. We performed our experiments for a variety of conditions, scene sizes and number of sensors, and the findings were consistent.

An example result is shown in Fig. 2. The figure, from left to right, illustrates (a) the scene and the three reconstruction results: (b) thresholded backprojection, (c) standard CoSaMP, and (d) CoSaMP-TV. The top row of the figure plots the whole 3-D scene, $\mathbf{s}$, either original or reconstructed, while the bottom row illustrates the corresponding depth map $\mathbf{d}$. Amplitude values in the former represent albedo, and amplitude values in the latter represent depths. The measurements were collected at 35 dB SNR, using 15% of measurements, an 85% reduction, with an impulse pulse shape.

As evident in the figure, and expected, thresholded backprojection, with the threshold chosen in hindsight for best performance, results to very poor reconstruction, even though it identifies three depth planes at the correct depths. The standard sparsity model, shown in column (c), significantly improves the result. However, the reconstruction of the closest plane contains errors. Within this region, we observe a high total variation in the reconstructed depth image. The reconstructed scene using our model in CoSaMP-TV, shown in column (d), leads to significantly improved reconstruction. The model based approach recovers 85.1% of the original support, while standard CoSaMP only recovers 58.8%.

Figure 3 further demonstrates quantitatively the performance improvements using our model. In particular, the figure plots the reconstruction performance of CoSaMP and CoSaMP-TV in terms of the detection rate, i.e., the percentage of the support recovered. This equals the size of the intersection of the recovered support with the true support of the signal, normalized by the size of the true support. The performance is plotted with respect to the number of sensors, as a function of the scene size in spatial pixels. Use of the TV model, as appropriate for such scenes, increases detection rate by more than 30% for the same number of sensors.

Our experiments demonstrate the validity of our model, as well as the feasibility of the proposed hardware architecture. Of course, further study is necessary, with more complex scenes and varying noise and acquisition conditions. We reserve this study, as well as an in-depth theoretical analysis, for an extended version of this paper.

6. REFERENCES

[1] Shahram Izadi, Ayush Bhandari, Achuta Kadambi, and Ramesh Raskar, "3D imaging with time of flight cameras: theory, algorithms and applications," in *ACM SIGGRAPH 2014 Courses*. ACM, 2014, p. 5.

[2] Gregory Howland, Petros Zerom, Robert W Boyd, and John C Howell, "Compressive Sensing LIDAR for 3D imaging," in *CLEO: Science and Innovations*. Optical Society of America, 2011, p. CMG3.

[3] Ahmed Kirmani, Andrea Colaço, Franco N. C. Wong, and Vivek K. Goyal, "Exploiting sparsity in time-of-flight range acquisition using a single time-resolved sensor," *Opt. Express*, vol. 19, no. 22, pp. 21485–21507, Oct 2011.

[4] Ahmed Kirmani, Andrea Colaço, Franco NC Wong, and Vivek K Goyal, "CoDAC: a compressive depth acquisition camera framework," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5425–5428.

[5] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[6] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt, "Approximation-tolerant model-based compressive sensing.," in *Proc. ACM Symposium on Discrete Algorithms (SODA)*. SIAM, January 2014, pp. 1544–1561.

[7] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt, "Approximation algorithms for model-based compressive sensing," *arXiv preprint arXiv:1406.1579*, 2014.

[8] Petros T. Boufounos, "Depth sensing using active coherent illumination," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, March 25-30 2012.

[9] Antonin Chambolle, "Total variation minimization and a class of binary mrf models," in *Energy minimization methods in computer vision and pattern recognition*. Springer, 2005, pp. 136–152.

[10] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar, "Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 167, 2013.

[11] Emmanuel J Candes, Justin K Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[12] David L Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[13] Richard G Baraniuk and Michael B Wakin, "Random projections of smooth manifolds," *Foundations of computational mathematics*, vol. 9, no. 1, pp. 51–77, 2009.

[14] Deanna Needell and Rachel Ward, "Stable image reconstruction using total variation minimization," *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, pp. 1035–1058, 2013.

[15] Marco F. Duarte, Mark A. Davenport, Dharmpal Takhar, Jason .N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, March 2008.

[16] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly, "Compressed sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[17] ME Gehm, R John, DJ Brady, RM Willett, and TJ Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics Express*, vol. 15, no. 21, pp. 14013–14027, 2007.

[18] Petros T. Boufounos, "Compressive sensing for over-the-air ultrasound," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 22-27 2011.

[19] Deanna Needell and Joel A Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

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

[21] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 69, 2007.

[22] Frank Lenzen, Henrik Schäfer, and Christoph Garbe, "Denoising time-of-flight data with adaptive total variation," in *Advances in Visual Computing*, pp. 337–346. Springer, 2011.

[23] Achuta Kadambi, Ayush Bhandari, and Ramesh Raskar, "3D depth cameras in vision: Benefits and limitations of the hardware," in *Computer Vision and Machine Learning with RGB-D Sensors*, pp. 3–26. Springer, 2014.

[24] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[25] Andrew Delong, Anton Osokin, Hossam N Isack, and Yuri Boykov, "Fast approximate energy minimization with label costs," *International journal of computer vision*, vol. 96, no. 1, pp. 1–27, 2012.

[26] Srikumar Ramalingam, Pushmeet Kohli, Karteek Alahari, and Philip HS Torr, "Exact inference in multi-label CRFs with higher order cliques," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[27] Hiroshi Ishikawa, "Exact optimization for markov random fields with convex priors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1333–1336, 2003.

[28] Yuri Boykov and Vladimir Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.