Compressive Coded Aperture Imaging

Roummel F. Marcia, Zachary T. Harmany, and Rebecca M. Willett

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

ABSTRACT

Nonlinear image reconstruction based upon sparse representations of images has recently received widespread attention with the emerging framework of compressed sensing (CS). This theory indicates that, when feasible, judicious selection of the type of distortion induced by measurement systems may dramatically improve our ability to perform image reconstruction. However, applying compressed sensing theory to practical imaging systems poses a key challenge: physical constraints typically make it infeasible to actually measure many of the random projections described in the literature, and therefore, innovative and sophisticated imaging systems must be carefully designed to effectively exploit CS theory. In video settings, the performance of an imaging system is characterized by both pixel resolution and field of view. In this work, we propose compressive imaging techniques for improving the performance of video imaging systems in the presence of constraints on the focal plane array size. In particular, we describe a novel yet practical approach that combines coded aperture imaging to enhance pixel resolution with superimposing subframes of a scene onto a single focal plane array to increase field of view. Specifically, the proposed method superimposes coded observations and uses wavelet-based sparsity recovery algorithms to reconstruct the original subframes. We demonstrate the effectiveness of this approach by reconstructing with high resolution the constituent images of a video sequence.

Keywords: Compressed sensing, coded aperture, sparse recovery, wavelets

1. INTRODUCTION

In a wide variety of video applications, keeping the focal plan array (FPA) small is useful or even critical. For example, in low light settings, where sensitive detectors are costly, smaller FPAs translate directly to less expensive systems. Smaller FPAs also make systems lighter weight and thus more portable. Finally, smaller cameras can fit into tighter spaces for unobtrusive surveillance. A key goal of many video systems, then, is to extract as much information as possible from a small number of detector array measurements.

Recent work in compressed sensing (CS)[1–3] indicates that it is possible to extract high-resolution images from small numbers of noisy, indirect, projection measurements when the scene is sparse or compressible in some basis. The high-resolution image is inferred from the measurements using a nonlinear, iterative reconstruction method which searches for the collection of basis coefficients which is (a) a good match to the observed data, and (b) sparse. Sparsity has long been recognized as a highly useful metric in a variety of inverse problems, but much of the underlying theoretical support was lacking. However, more recent theoretical studies have provided strong justification for the use of sparsity constraints and quantified the accuracy of sparse solutions to these underdetermined systems.[1,4]

Despite the theoretical promise of CS specifically and sparsity constrained iterative reconstruction in general, very little is known about its application to practical video systems. In particular, imaging systems implicitly place hard constraints on the nature of the measurements which can be collected, such as non-negativity of both the projection vectors and the measurements, which are not considered in the existing compressed sensing literature. Furthermore, it typically is not possible to wait hours or even minutes for an iterative reconstruction routine to produce a single frame of a video; rather, algorithms must be able to operate effectively under stringent time constraints.

In this paper, we explore the potential of several different practical systems for measuring a temporally varying scene. While holding both the total intensity of the scene and the FPA size constant, we simulate collecting measurements using a pinhole camera, a coded aperture camera using conventional Modified Uniformly Redundant Arrays (MURAs),[5] a coded

Further author information: (Send correspondence to Roummel F. Marcia.)
Roummel F. Marcia: E-mail: roummel@ee.duke.edu, Telephone: 1 (919) 613 9123
Zachary T. Harmany: E-mail: zth@duke.edu, Telephone: 1 (919) 613 9123
Rebecca M. Willett: Email: willett@duke.edu, Telephone: 1 (919) 660 5544

aperture camera specifically designed for CS applications,[6] a recently proposed camera which superimposes image halves,[7] and a superimposition camera with CS coded apertures. We will see that CS coded aperture constructions in general yield performance gains over pinhole cameras and conventional coded aperture cameras in video settings. In other words, given a fixed size FPAs, compressive measurements combined with sophisticated optimization algorithms such as the one described in this document can significantly increase the quality and/or resolution of the video.

1.1 Observation Model

Throughout this paper, we use the following observation model:

$$y_t = A_t f_t^\star + w_t, \tag{1}$$

where $f_t^\star \in I\!R^n$ is the image of interest at time t, $A_t \in I\!R^{k\times n}$ linearly projects the scene onto a k-dimensional set of observations, w_t is a vector of white Gaussian noise, and $y_t \in \mathbb{Z}^k$ is a length-k vector of observations.

The measurement matrix A_t can correspond to a wide variety of imaging system models, several of which will be examined explicitly in the course of this paper. For example, A_t could correspond to a coded aperture point spread function composed with a downsampling operation, resulting in a very low resolution coded aperture observation. Alternatively, A_t could represent superimposing different regions in a wide field of view scene, resulting in an ambiguous representation of the scene which must separated into different intensity functions for the different scene regions. This will be made explicit in the following sections.

The problem addressed is this paper is the estimation of $\{f_t^\star\}$ from $\{y_t\}$ in a compressed sensing context, when (a) the number of unknowns is much larger than the number of observations, (b) $f_t^\star$ is sparse or compressible in some basis W, and (c) the matrix product $R_t \triangleq A_t W$ is sufficiently "incoherent" or satisfies some other probabilistic criterion.[8,9] While the problem (1) can be grossly underdetermined, compressed sensing suggests that selecting the *sparsest* solution to this system of equations will yield a highly accurate solution.

1.2 Organization of the Paper

This paper is organized as follows. In Sec. 2 we describe several practical camera architectures which could be used in to yield practical video cameras which make effective use of a limited size FPA: pinhole cameras, MURA coded apertures, compressive coded apertures, and subframe superimposition. Sec. 3 describes mechanisms used in optical video settings to improve the quality and speed of video frame reconstruction algorithms; these methods are used to exploit correlations between successive frames and compensate for the positivity of optical measurements and point spread functions. In Sec. 4 we report our the results of our numerical experiments. Analysis and conclusions are presented in Sec. 5.

2. ARCHITECTURES FOR SNAPSHOT COMPRESSIVE VIDEO CAMERAS

Examining the observation model in (1), a natural question is the following: what is the "best" A_t? As noted above, the compressed sensing literature describes the theoretical optimality of choosing A_t to be populated by random draws from an appropriate distribution, but practical aspects of optical system design and positivity constraints make this infeasible in our setting. Others have suggested taking a different random projection at each time step and producing the projections using digital micromirrors.[10] This approach only requires one detector element, but requires the scene to remain static for a relatively long period of time. Using an array of digital micromirror chips to collect multiple random projects simultaneously would add significant bulk and power requirements to the system design.

In this paper we consider an alternative based on coded apertures. The projections recorded based on coded apertures have somewhat weaker theoretical guarantees than a series of completely different random projects, but this disadvantage is offset by the following considerations: (a) coded apertures are simple to build and incorporate into practical, robust and compact optical system designs, and (b) all the measurements of a single frame of scene can be collected in a single "snapshot", allowing us to reconstruct dynamic scenes with more fidelity than would be possible with a system which only collects one random projection at each time step.

In this section, we review the challenges associated with pinhole cameras which led to the initial development of coded apertures, and then briefly describe the conventional MURA coded aperture design. We then describe how CS leads to a new, optimal coded aperture design, and how these coded apertures can be used in conjunction with subframe superimposition to increase both video resolution and field of view.

2.1 Pinhole cameras

Pinhole cameras are simple imaging systems that use a single opening (aperture) for collecting observations. Very small apertures lead to images with sharp details and crisp edges – aperture size determines the smallest feature a pinhole camera can resolve. However, because they allow relatively little light to pass through, a slow shutter speed is often required to prevent under-exposure. This requirement is problematic for capturing images with non-static components. Also, it places an added constraint that the camera be stationary for prolonged periods of time, which might not be feasible in some settings, such as cameras that are mounted on vehicles. Dim observations resulting from small apertures can be compensated by making the size of the opening larger; but, larger apertures increase light diffraction, leading to increased blurriness in the observations.

For pinhole cameras, the measurement matrix operation can be modeled as

$$A_t^{(\text{pinhole})} f_t^\star = f_t^\star * h^{(\text{pinhole})},$$

where $*$ denotes the convolution operator and h_{pinhole} is a Gaussian blur with full width at half maximum (FWHM) proportional to the size of the pinhole. (Note that in this case the operator A_t is independent of t, but we do not reflect this in the notation to be consistent throughout the paper.) The peak amplitude of the Gaussian is one, so smaller pinholes correspond to Gaussian blurs with smaller integrals which subsequently cause less light to reach the detector.

2.2 Modified Uniformly Redundant Arrays

Coded aperture imaging was developed to allow more light to reach the detector without a loss in resolution. Seminal work in coded aperture imaging includes the development of Modified Uniformly Redundant Arrays (MURAs).[5] These mask patterns are binary, square patterns with prime integer sidelengths (see Fig. 1(a)). The measurement matrix operation can then be modeled as

$$A_t^{(\text{MURA})} f_t^\star = f_t^\star * h^{(\text{MURA})},$$

and $f_t^\star$ can be reconstructed as

$$\widehat{f}_t = y * \widetilde{h}^{(\text{MURA})}$$

for some complementary pattern $\widetilde{h}^{(\text{MURA})}$ (see Fig. 1(b)). In other words, the MURA patterns (and their complements) are specifically designed so that $h^{(\text{MURA})} * \widetilde{h}^{(\text{MURA})}$ approximately equals the Kronecker δ function (Fig. 1(c)) and hence to optimize reconstruction accuracy *subject to the constraint that linear, convolution-based reconstruction methods would be used*. MURA coded apertures are approximately 50% open, and, therefore, a significant improvement over conventional pinhole cameras because they allow in significantly more light.

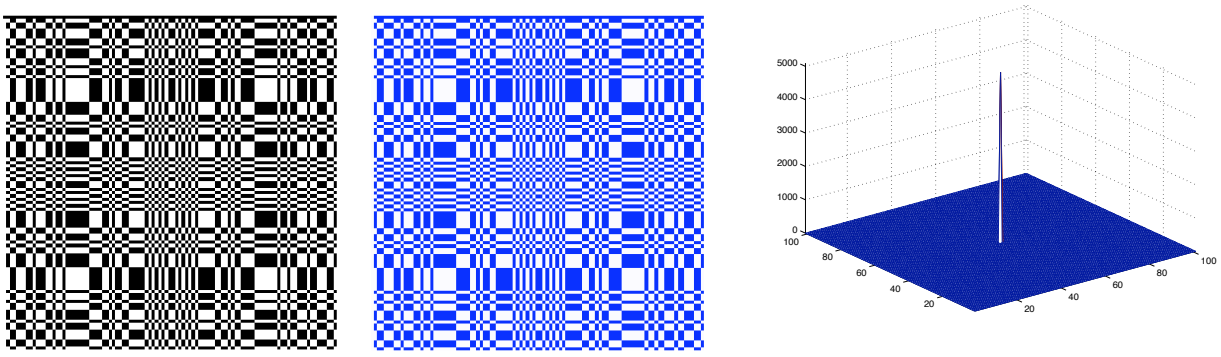
Generally, the size and resolution of the MURA patterns is chosen to match the size of the FPA. Using a higher resolution MURA pattern would not be effective because the FPA would effectively downsample the coded image, making the reconstruction via convolution with a complementary pattern suboptimal. As a result, the resolution of the estimate which can be achieved using MURA patterns is limited by the size of the FPA even when we have prior information about the scene (such as its sparsity in some basis). Furthermore, while MURA coded apertures are successful in the context of linear reconstruction, there exist a wide variety of nonlinear reconstruction methods which can dramatically outperform linear reconstructions when f has a sparse representation in some basis, such as a wavelet basis.

2.3 Compressive coded apertures

2.3.1 Compressed sensing preliminaries

Nonlinear image reconstruction based upon sparse representations of images has received widespread attention recently with the advent of "compressive sensing". This emerging theory indicates that very high dimensional vectors ($f_t^\star \in \mathbb{R}^N$, where $N = n^2$) can be recovered with astounding accuracy from a much smaller dimensional observation (y) when $f_t^\star$ has a "sparse" representation in some basis W (*i.e.,* $f_t^\star = W\theta_t^\star$ where $\theta_t^\star$ has few non-zero coefficients). This result hinges on the *Restricted Isometry Property* (RIP)[8] condition being satisfied on the product of the observation matrix A_t and the basis matrix W, denoted $R_t \triangleq A_t W$. A matrix R is said to satisfy the RIP of order $3m$ if, for $T \subset \{1, 2, \cdots, N\}$ and R_T, a submatrix obtained by retaining the columns of R corresponding to the indices in T, there exists a constant $\delta_{3m} \in (0, 1/3)$ such that for all $z \in \mathbb{R}^{|T|}$,

$$(1 - \delta_{3m})\|z\|_2^2 \leq \|R_T z\|_2^2 \leq (1 + \delta_{3m})\|z\|_2^2 \tag{2}$$



(a) Coding pattern $h^{(\mathrm{MURA})}$ (b) Decoding pattern $\widetilde{h}^{(\mathrm{MURA})}$ (c) Convolution $h^{(\mathrm{MURA})} * \widetilde{h}^{(\mathrm{MURA})}$

Figure 1. The MURA pattern for a 101×101 grid. (a) The white blocks are openings (b) The decoding pattern for (a) is nearly identical. (c) If the white blocks have value 1 and the black blocks have value 0, and the blue blocks in (b) have value -1, then the convolution of the matrices corresponding to this two patterns appear like a Kronecker-δ function.

holds for all subsets T with $|T| \leq 3m$.[8] An observation matrix R satisfying RIP with high probability is often referred to as a compressed sensing (CS) matrix. While the RIP cannot be verified for a given R, it has been shown that matrices with entries drawn independently from some probability distributions satisfy the condition with high probability when $k \geq Cm \log(N/m)$ for some constant C, where $m \equiv \|\theta_t^\star\|_{\ell_0}$ is the number of non-zero elements in the vector $\theta_t^\star$.[8]

We observe $y_t = R_t\theta_t^\star + n_t$, where n_t is white Gaussian noise. The $\ell^2 - \ell^1$ minimization

$$\begin{aligned} \widehat{\theta}_t &= \arg\min_{\theta_t} \frac{1}{2}\|y_t - R_t\theta_t\|_2^2 + \tau\|\theta_t\|_1 \\ \hat{f}_t &= W\widehat{\theta}_t \end{aligned} \tag{3}$$

will yield a highly accurate estimate of $f_t^\star$ with very high probability.[4,11] The regularization parameter $\tau > 0$ helps to overcome the ill-posedness of the problem, and the ℓ^1 penalty term drives small components of θ to zero and helps create sparse solutions.

2.3.2 Coded apertures for compressed sensing

We have previously designed a coded aperture imaging mask, denoted $h^{(\mathrm{CCA})}$, such that the corresponding observation matrix R_t satisfies the RIP.[6] The measurement matrix A_t (recall $R_t = A_tW$) associated with compressive coded apertures (CCA) can be modeled as

$$A_t^{(\mathrm{CCA})} f_t^\star = D(f_t^\star * h^{(\mathrm{CCA})}) \tag{4}$$

where D is a downsampling operator induced by the FPA which consists of partitioning $f_t^\star * h^{(\mathrm{CCA})}$ into k uniformly sized blocks and measuring the total intensity in each block. (This is sometimes referred to as integration downsampling.) In this case the operator A_t is independent of t, but we do not reflect this in the notation to be consistent throughout the paper.

The convolution of $h^{(\mathrm{CCA})}$ with a signal $f_t^\star$ as in (4) can be computed by applying the Fourier transform $\mathscr{F}$ to $f_t^\star$ and $h^{(\mathrm{CCA})}$, then performing element-wise matrix multiplication, and then mapping the product using the Fourier inverse transform. In linear algebra notation, this series of operation can be expressed as

$$h^{(\mathrm{CCA})} * f_t^\star = \mathscr{F}^{-1} C_H \mathscr{F}\, f_t^\star,$$

where $\mathscr{F}$ is the two-dimensional Fourier transform matrix, and C_H is the diagonal matrix whose diagonal elements correspond to the transfer function $H = \mathscr{F}(h^{(\mathrm{CCA})})$. (This is a slight abuse of notation, since on the left hand side $f_t^\star$ is an image, and on the right hand side $f_t^\star$ is a vectorized representation of an image.) The coded aperture masks are designed so that $A_t^{(\mathrm{CCA})}$ satisfies RIP as described in (2). Specifically, the authors developed a method for randomly generating a mask $h^{(\mathrm{CCA})}$ so that the corresponding matrix product $\mathscr{F}^{-1}C_H\mathscr{F}$ is block-circulant and each block was in turn circulant,

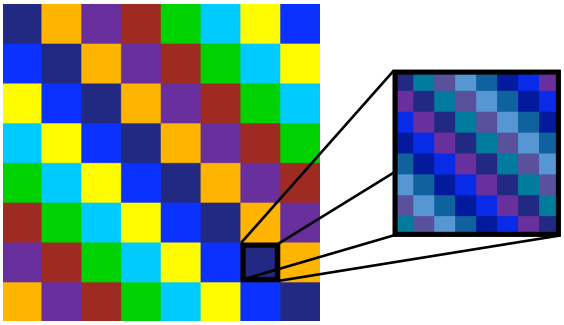


Figure 2. The matrix $\mathscr{F}^{-1}C_H\mathscr{F}$ is block-circulant with circulant blocks.

as illustrated in Fig. 2. Block-circulant matrices are known to be a compressed sensing matrix, and based upon recent theoretical work on Toeplitz-structured matrices for compressive sensing, the proposed masks are fast and memory-efficient to generate.[6,12] The incorporation of the integration downsampling operator D does not prevent the RIP from being satisfied; a key element of the proof that the RIP is satisfied is a bound on the number of rows of $A_t^{(\mathrm{CCA})}$ which are statistically independent. Since the downsampling operator effectively sums rows of a block circulant matrix, downsampling causes the bound on the number of dependent matrix rows to be multiplied by the downsampling factor.

2.4 Subframe superimposition and disambiguation

Recently, we proposed a numerical method by which the FOV of an imaging system can be increased without compromising its resolution.[7,13] In our setup, a dynamic scene to be imaged is partitioned into smaller scenes (called subframes), which are imaged onto a single FPA using beam splitters and mirrors to form a composite image. This is illustrated in Fig. 3. We developed an efficient video processing approach to separate the composite image into its constituent images, thus restoring the complete scene corresponding to the overall FOV. To make this otherwise highly ill-posed problem of disambiguating the image tractable, the superimposed subframes are moved relative to one another between video frames. This approach is easily implemented and is mechanically more robust than a multiple-shutter system that alternately opens and closes shutters to capture the subframes of a scene.

Each frame $f_t^\star$ is partitioned into left and right subframes $f_t^\star = [f_{L,t}^\star; f_{R,t}^\star]$. The measurement matrix operation associated with the superimposition process can be modeled as

$$A_t^{(\mathrm{sup})} f_t^\star = D[h_L * f_{L,t}^\star + h_R * (S_t f_{R,t}^\star)]$$

where h_L and h_R are the point spread function of the imager and can either correspond to a pinhole or to a coded aperture, and S_t is an operator describing the shifting movement of one subframe relative to another. In the superimposition process, the two subimages are merged to form a composite image. For the case where h_L and h_R correspond to pinholes, the intensity of each pixel in the composite image is the simple summation of the intensities of the corresponding pixels in the individual images.

For the case where h_L and h_R correspond to coded apertures, the intensity of each pixel in the composite image is the summation of the corresponding coded aperture images. A schematic illustrating how the coded apertures could be included in the camera design in presented in Fig. 4. These coded observations require fewer measurements than the original subframes to reconstruct them with high accuracy. This fact allows for an even smaller focal plane array while maintaining our ability to reconstruct the original image with high accuracy. The coded aperture masks h_L and h_R are generated similarly as in the compressive coded aperture setup, i.e., they are generated independently such that the resulting projection matrix for each mask satisfies the RIP (2).

The inverse process – the disambiguation of the individual subimages from this composite image – is more challenging. For this, we must determine how the intensity of each pixel in the composite image is distributed over the corresponding pixels in the individual subimages so that the resulting reconstruction accurately approximates the original scene. As before, an estimate can be formed using the $\ell^2 - \ell^1$ minimization described in (3).

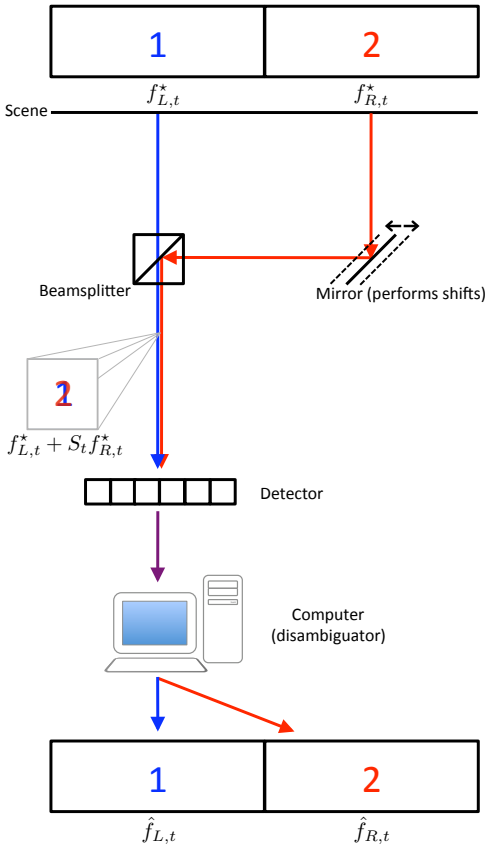


Figure 3. Schematics of the superimposition and disambiguation process.

3. IMPROVING PERFORMANCE OF RECONSTRUCTION ALGORITHMS WITH WARM-STARTING AND MEAN SUBTRACTION

All of the architectures described above result in a system which collects far fewer measurements than there are pixels in the image to be reconstructed. Iterative reconstruction methods are effective for these architectures, as we will demonstrate in Sec. 4. However, we perform two key tasks which significantly improve the performance of iterative solvers for the video reconstruction problem. The first is solving for multiple frames simultaneously combined with warm-starting, which improves the initialization of the reconstruction routine for each frame and significantly reduces the time required to achieve an accurate reconstruction. The second is mean subtraction, which allows us to compensate for R_t not having zero mean in practical camera systems.

3.1 Multi-frame difference reconstruction

Several of the above architectures lead to challenging inverse problems which can be solved using the $\ell_2 - \ell_1$ minimization in (3). While this is an effective approach, the reconstruction results can be significantly improved by solving for multiple video frames simultaneously, instead of solving for each frame separately. To see this, note that we could write two successive observed frames in any of the above architectures as

$$\begin{aligned} \begin{bmatrix} y_t \\ y_{t+1} \end{bmatrix} &\triangleq \begin{bmatrix} R_t & 0 \\ 0 & R_{t+1} \end{bmatrix} \begin{bmatrix} \theta^\star_t \\ \theta^\star_{t+1} \end{bmatrix} + \begin{bmatrix} n_t \\ n_{t+1} \end{bmatrix} \\ &\equiv \begin{bmatrix} R_t & 0 \\ R_{t+1} & R_{t+1} \end{bmatrix} \begin{bmatrix} \theta^\star_t \\ \Delta\theta^\star_t \end{bmatrix} + \begin{bmatrix} n_t \\ n_{t+1} \end{bmatrix}, \end{aligned}$$

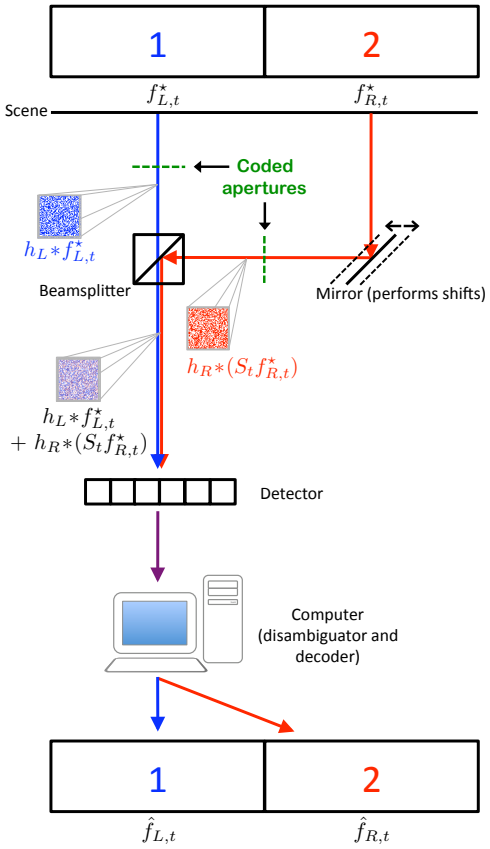


Figure 4. Schematics of the superimposition of coded aperture observation and disambiguation process.

where $\Delta\theta^\star_t \triangleq \theta^\star_{t+1} - \theta^\star_t$. For strongly correlated frames, $\Delta\theta^\star_t$ will be very sparse (even significantly sparser than $\theta^\star_{t+1}$), and, thus, will be highly suitable for sparse recovery algorithms. In fact, the above observation leads to the following optimization problem as an alternative to (3):

$$\begin{bmatrix} \widehat{\theta}_t \\ \widehat{\Delta\theta}_t \end{bmatrix} = \underset{\theta, \Delta\theta_t}{\arg\min} \left\| \begin{bmatrix} y_t \\ y_{t+1} \end{bmatrix} - \begin{bmatrix} R_t & 0 \\ R_{t+1} & R_{t+1} \end{bmatrix} \begin{bmatrix} \theta_t \\ \Delta\theta_t \end{bmatrix} \right\|_2^2 + \tau \left\| \begin{bmatrix} \theta_t \\ \Delta\theta_t \end{bmatrix} \right\|_1 . \tag{5}$$

Note that in this formulation, the coefficients for the current frame and the *difference* between two subsequent frames are solved. The coefficients for the subsequent frame will be given by $\widehat{\theta}_{t+1} = \widehat{\theta}_t + \widehat{\Delta\theta}_t$. The following optimization problem for frame $(t+1)$ is initialized using

$$\begin{bmatrix} \theta^{(0)}_{t+1} \\ \Delta\theta^{(0)}_{t+1} \end{bmatrix} \equiv \begin{bmatrix} \widehat{\theta}_t + \widehat{\Delta\theta}_t \\ \widehat{\Delta\theta}_t \end{bmatrix} .$$

This approach can be extended for solving arbitrarily many frames simultaneously but limited by memory constraints and by the amount of time a solver is allowed per optimization problem.[7,13] In our simulations, we solved for four frames simultaneously in order to balance computation time per frame versus reconstruction accuracy. This approach is related to recent work in high-dimensional joint support recovery, in which the fact that successive frames share a partially common support is used to significantly improve reconstruction accuracy.[14]

3.2 Mean subtraction

Generative models for random projection matrices used in CS involve drawing matrix elements independently from a zero-mean probability distribution,[1,2,4,8,12,15] and likewise a zero-mean distribution was used to generate the coded aperture

masks described in Sec. 2.3. However, a coded aperture mask with zero mean is not physically realizable in optical systems. As described in,[6] we generate our physically realizable mask by taking our randomly generated, zero mean mask pattern and scaling it so that all mask elements are in the range $[0, 1/n]$. This rescaling ensures that the coded aperture corresponds to a valid probability transition matrix which describes the distribution of photon propagation through the optical system.

This rescaling, while necessary to accurately model real-world optical systems, negatively impacts the performance of the proposed $\ell_2 - \ell_1$ reconstruction algorithm. More specifically, if we generate a non-realizable zero-mean mask ($\widetilde{h}^{(\mathrm{CCA})}$) with elements in the range $[-\frac{1}{2n}, \frac{1}{2n}]$ and simulate measurements of the form

$$\widetilde{y} = D(f^\star * \widetilde{h}^{(\mathrm{CCA})}) \equiv \widetilde{A}^{(\mathrm{CCA})} f^\star \tag{6}$$

(we omit the t subscripts in this section for simplicity of presentation), then the corresponding observation matrix $\widetilde{A}^{(\mathrm{CCA})}$ will satisfy the RIP with high probability and $f^\star$ can be accurately estimated from $\widetilde{y}$ using $\ell_2 - \ell_1$ minimization. In contrast, if we rescale $\widetilde{h}^{(\mathrm{CCA})}$ to be in the range $[0, 1/n]$ and denote this $h^{(\mathrm{CCA})}$, then we have a practical and realizable coded aperture mask, but observations of the form

$$y = D(f^\star * h^{(\mathrm{CCA})}) \equiv A^{(\mathrm{CCA})} f^\star$$

cannot be directly used with $\ell_2 - \ell_1$ minimization to yield as accurate an estimate. To address this problem, we note that $h^{(\mathrm{CCA})} = \widetilde{h}^{(\mathrm{CCA})} + \frac{1}{2n}$, yielding

$$y = A^{(\mathrm{CCA})} f^\star = \left(\widetilde{A}^{(\mathrm{CCA})} + \frac{1}{2n} \mathbb{1}_{k \times n} \right) f^\star = \widetilde{y} + \frac{\|f^\star\|_1}{2n} \mathbb{1}_{k \times 1},$$

where $\mathbb{1}_{k \times n}$ is a $k \times n$ matrix of ones and we exploit the known positivity of $f^\star$. Furthermore, since y is also positive we note that

$$\mathrm{E}\left[\|y\|_1\right] = \sum_{i=1}^{k} \sum_{j=1}^{n} A_{i,j}^{(\mathrm{CCA})} f_j^\star = \sum_{j=1}^{n} \left(\sum_{i=1}^{k} A_{i,j}^{(\mathrm{CCA})} \right) f_j^\star = C_A \|f^\star\|_1,$$

where C_A is the sum of each column of $A^{(\mathrm{CCA})}$ and is known by construction. Putting this all together we can estimate

$$\widetilde{y} \approx y - \frac{\|y\|_1}{2nC_A} \mathbb{1}_{k \times 1},$$

and use this estimate to solve for $f^\star$ in (6). It can readily be seen that solving for $f^\star$ in (6) will produce a solution with zero mean, and so we add $\|y\|_1 / C_A$ to this result to achieve our final, accurate estimate.

4. NUMERICAL EXPERIMENTS

Previously, we conducted experiments demonstrating that our compressive coded aperture reconstruction methods are able to preserve edges better and capture more details than using uncoded observations.[6,16] Our goal here is to more precisely assess the resolution gains and FOV increase associated with the compressive coded apertures and superimposition architectures for a fixed focal plane array size.

4.1 Video

For our numerical experiments, we designed a video that tests whether the proposed subframe superimposition and disambiguation approach with compressive coded apertures described in Sec. 2.4 will be effective in accomplishing this goal. Each video frame is in 256×512 gray-scale pixel resolution and consists of vertical bars of various widths spaced unevenly in a row (see Fig. 5), moving slowly from the middle of the frame to the top. For this video, we created 50 frames. The frames are sparse in the Haar wavelet domain and are, therefore, suitable for the compressed sensing framework. Although this video seems simple, it is particularly challenging because several of these vertical bars are very narrow, and in several instances, are one pixel in width. In addition, there are spacings between vertical bars that are also one pixel or very narrow. Zero-mean additive white Gaussian noise is added to the observations.

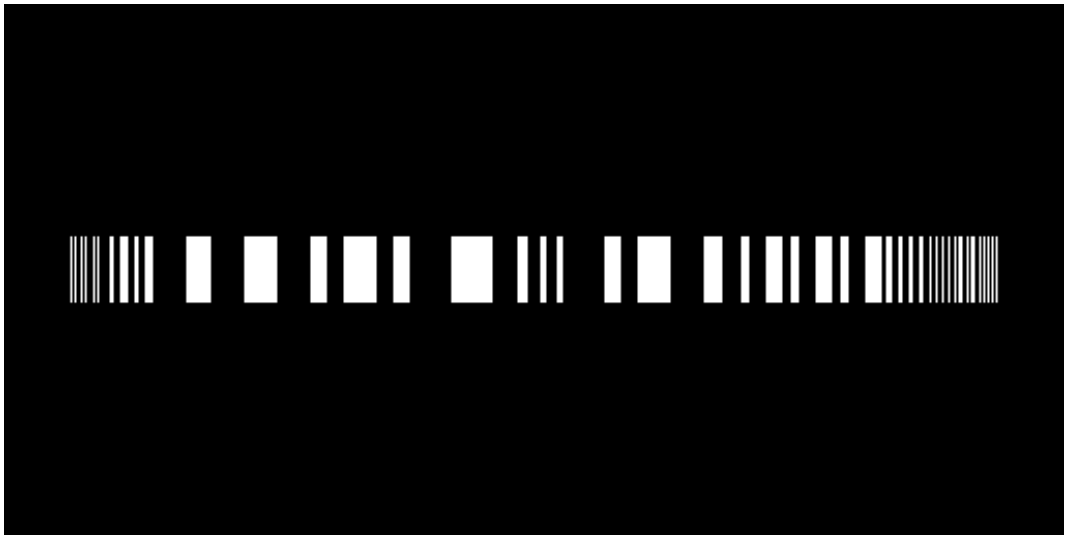


Figure 5. A single frame of the video to be reconstructed. The video depicts the simultaneous movements of the vertical bars from the middle of the frame to the top. Low resolution cameras blur the fine lines, narrow spaces, and edges of the vertical bars. Compressive coded apertures help mitigate these effects.

Our experiments compare the reconstruction results from using (1) a pinhole camera (which we label as Pinhole), (2) a camera with a MURA coded aperture (MURA), (3) a camera with a compressive coded aperture (CCA), (4) a pinhole camera with the superimposition and disambiguation setup (Pinhole Dis), and (5) a camera with the superimposition and disambiguation setup with compressive coded apertures (CCA Dis). We do not include a superimposition and disambiguation setup with MURA coded apertures because MURA is designed for linear reconstruction using the MURA decoding pattern, where as the disambiguation problem formulation is nonlinear. Without prior knowledge about the scene, it is impossible to use a linear reconstruction method to disambiguate the two halves of the scene.

We require that each camera setup have a 128×128 focal plane array size. Thus, in all setups, a downsampling by a factor of two in each dimension must be applied because the original video is twice the resolution of the FPA in the vertical direction and four times in the horizontal direction. In addition, to take measurements on the 128×128 FPA, half of the original scene be cropped out in Setups (1)-(3). Setups (4) and (5) superimpose the scene halves and disambiguate the subframes in the reconstruction process. These experiments explore the tradeoffs between increasing the video field-of-view and potentially decreasing spatial resolution, and how using compressive coded apertures can mitigate some of the loss in resolution.

In these experiments, we solve the optimization problem (5) using the Gradient Projection for Sparse Reconstruction (GPSR) algorithm,[17] a derivative-based optimization algorithm that uses a projected gradients. It is very fast, accurate, and efficient. In addition, GPSR has a *debiasing* phase, where upon solving the $\ell^2 - \ell^1$ minimization problem, it fixes the non-zero pattern of the optimal θ_t and minimizes the ℓ^2 term of the objective function, resulting in a minimal least-squares error in the reconstruction while keeping the number of non-zeros in the wavelet coefficients at a minimum. Published results have shown that it outperforms many of the state-of-the-art codes for solving the $\ell^2 - \ell^1$ minimization problem or its equivalent formulations. In our numerical experiments, we restricted the number of GPSR iterations due to time constraints imposed on reconstructing the entire duration of the video sequence. Higher quality reconstructions are possible with more computation time, but not realistic in video settings with a steady stream of new observations.

4.2 Results and analysis

Here we compare the performance and assess the relative merits of the proposed methods so that one may choose which method to implement for a given task. (The image is cropped to focus on the resolution bar pattern for visual clarity.) In describing the performance of image reconstruction methods, the normalized mean squared error, MSE $= \|\widehat{f} - f\|_2 / \|f\|_2$, allows us to quantify the fidelity of the reconstruction to the original scene. However, because the MSE is a global metric over the entire scene, it fails to capture how well small high-resolution features are preserved in the estimates. For this

Figure 6. A comparison of the ground truth with the reconstructions using the five different setups: pinhole camera, camera with MURA coded aperture, camera with compressive coded aperture (CCA), disambiguation with a pinhole camera, and disambiguation with a camera using CCA. All five setups use the same FPA size. The pinhole, MURA, and CCA must be cropped, whereas the pinhole and CCA disambiguation setups superimpose the two halves onto the same FPA.

reason, we also examine one-dimensional normalized intensity profiles of the reconstructions to examine the peak-to-valley differences as an alternative metric for image quality.

Figure 6 shows, for a single frame, a comparison between the reconstructions obtained in each method. Notice immediately that the subframe superimposition and disambiguation approaches offer twice the field of view compared to the other methods that are focused only on the right half of the scene. We first compare these setups that can only capture half of the scene. All three setups reconstruct the general shape of the blocks on the right hand side of the scene. Pinhole and MURA require upsampling to make their reconstructions comparable to the original scene. We used bicubic interpolation, which blurred the edges but produced lower MSE values. In contrast, CCA incorporates the downsampling operator in its problem formulation (4) and, hence, does not require an upsampling scheme. For the disambiguation methods that are able to reconstruct the entire scene, we notice that using the CCA we are able to reconstruct details (especially on the left half of the image) that the pinhole is unable to capture. In addition, both of the disambiguation methods incorporate the downsampling operator in their formulations, and therefore do not require upsampling.

When we compare the MSE of the reconstructions to the wide FOV scene (Fig. 7(a)), there is a large gap between the disambiguation approaches versus the other methods simply because they make observations of and reconstruct both halves of the scene. We first focus on the disambiguation approaches (Pinhole Dis and CCA Dis). Because we use relatively few iterations per frame to solve the optimization problem, the reconstruction for the first few frames are often not very accurate. However, because we take full advantage of the warm-starting strategy for subsequent frames, the MSE quickly improves, settling to a steady-state value, with the CCA disambiguation method consistently outperforming the pinhole disambiguation method after 20 video frames.

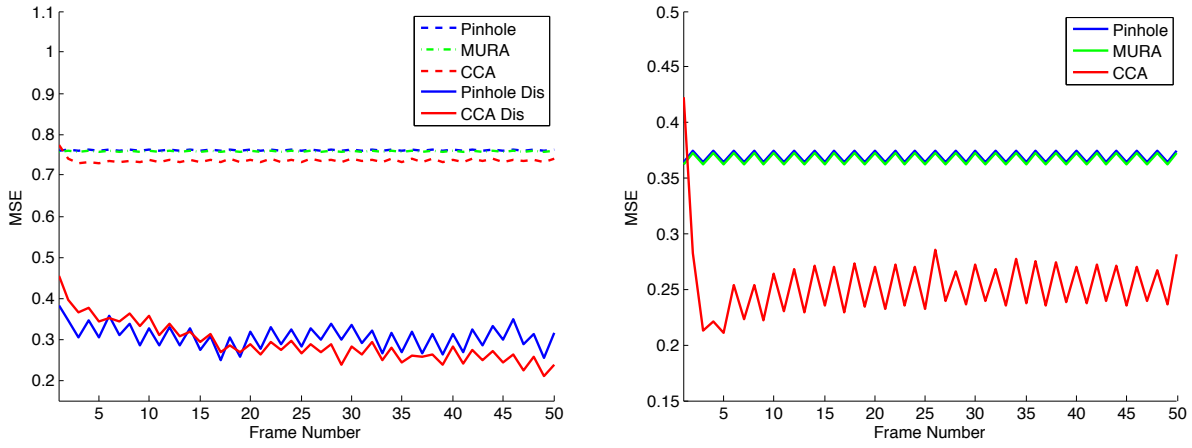
If we focus only on the ability of the other methods to reconstruct the right half of the scene (Fig. 7(b)), we see that the CCA method yields much higher performance than the MURA or pinhole methods. These MSE values were computed only over the pixels in the right half of the scene. The MSE values for the pinhole and MURA are very steady, because although moving, the shape of the objects in the video are not changing. Due to the low noise conditions of our simulation, the MURA and pinhole images are virtually indistinguishable. CCA is an iterative method similar to CCA Dis and therefore exhibits the same MSE convergence behavior, i.e., dramatic improvements in MSE values are obtained after very few frames. In this case, CCA outperforms pinhole and MURA after only two frames. These results validate the use of compressive coded apertures offers an effective means to improve reconstruction accuracy.

The plots in Fig. 7 exhibit a zig-zag pattern which we observe consistently across all considered architectures and reconstruction algorithms. In the video, the vertical bars move one pixel upwards at each frame. The downsampling in the

observations are obtained by averaging four adjacent pixels. Thus, in every even-numbered frame, there is a blurring of the edges along the top and bottom of each bar (the bars are of even length), while in the odd-numbered frames this blurring doesn't occur and the MSE is smaller. This issue is present in the pinhole and MURA simulations, and exacerbated by our use of Haar wavelets for the reconstructions from CCA observations. For the Haar wavelet basis, sharp edges are encoded in fewer coefficients than smooth transitions, and so the slight blurring in even-numbered frames at the top and bottoms of the vertical bars introduces additional significant wavelet coefficients that systematically change the sparsity pattern of the wavelet coefficients over alternating frames of the reconstruction. This alternating between sharp discontinuities and smooth transitions yields the triangular wave shape of the MSE curves in the reconstructions.

Figure 8(a) shows in detail a cross section of ground truth (in black) and of the reconstruction from the Pinhole (blue), MURA (green), and CCA (green) observations, focused on the fourteen rightmost bars. The peaks correspond to the locations of the vertical bars, and close alignment of the peaks of the different methods to the peaks of the ground truth imply more accurate reconstructions of the vertical bars. Both the MURA and the pinhole camera obscure features that are smaller a few pixels wide where in most cases the compressive coded aperture reconstruction is able to achieve higher resolution, failing only when single-pixel wide features are separated by single-pixel wide spaces. Specifically, CCA correctly identifies the peaks near pixel locations 450 and 455. Pinhole and MURA incorrectly identify a peak at 452-453. Also, CCA identifies the peaks at 463 and 469 and the valleys (corresponding to spaces in the original scene) near 465 and 470. In contrast, Pinhole and MURA are unable to differentiate the peaks and valleys from locations 460 to 475 and produces only the average pixel intensities.

The results for the second group of setups (Pinhole Dis and CCA Dis) are comparable to each other (Fig. 9(b)). Perhaps an interesting result of this experiment is that Pinhole Dis, in this case, is able to identify the four peaks around pixel locations 450 and 455, which Pinhole is unable to. This result is due to the movement of the mirror in the disambiguation camera setup (see Fig. 3). The position of the mirror can be changed very slightly, thus, inducing a sub-pixel change in the observation. In our disambiguation setup, this change is modeled by letting the shifting operator S_t act before the downsampling operator D. Thus, in the measurements, a shift of one pixel before downsampling translates to a sub-pixel movement. In comparison, the Pinhole Dis reconstruction on the leftmost vertical bars (see Fig. 9(a)) averages the first four peaks (near pixel locations 35 and 40), whereas CCA Dis accurately captures the first two. Also, CCA Dis is better able to approximate the peaks between locations 50 and 55 and near 60. The increase in resolution in Pinhole Disambiguation on the right half of the scene is not present on the left half of the scene because the left subimage remains stationary, and, therefore, remains unaltered.



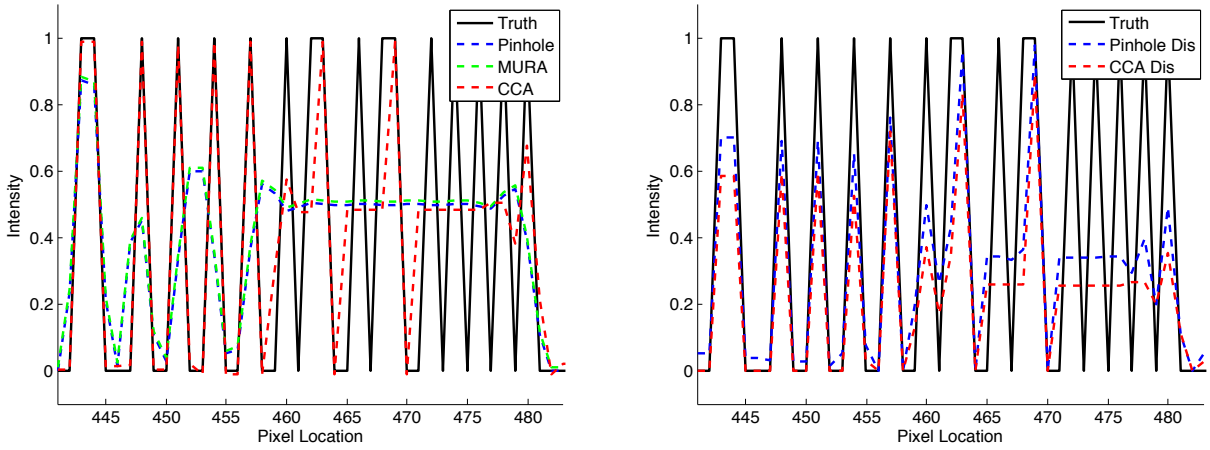
(a) (b)

Figure 7. MSE plots of the reconstruction for each frame. (a) MSE values for all five setups to reconstruct the full scene. (b) MSE values restricted to the right subframe for the non-disambiguating setups (Pinhole, MURA, and CCA).

5. CONCLUSIONS

High-resolution video measurement can be challenging in many settings where small cameras (and hence small focal plane arrays) are essential. Conventionally, small focal plane arrays translated directly into low-resolution data. Recent theoretical work in Compressed Sensing suggests that this is not a fundamental limitation, and that high-resolution video can be estimated from a relatively small number of random projection measurements. However, it is very difficult to build a practical camera which computes random projections of scene for several reasons. First, projection matrices consistent with realizable optical systems cannot have negative elements, which is contrary to most generative models for random projection matrices. Second, having all the projections (i.e. rows of R_t in the notation of this paper) completely independent would force us to have a very large, expensive and complex optical system, which would not be usable in many applications. Third, computing a single different random projection at each time step (cf.[10]) severely limits temporal resolution of the video camera.

In this paper, we showed that coded aperture designs based on the principles of Compressed Sensing lead to very simple and robust camera architectures which are particularly effective for low light video settings: (a) a simple coded aperture



(a)

Figure 8. A cross section of the rightmost vertical bars from Fig. 6 of (a) the ground truth (black) and the reconstruction using Pinhole (blue), MURA (green), CCA (red); and (b) the Pinhole Disambiguation (dashed blue) and CCA Disambiguation (dashed red).

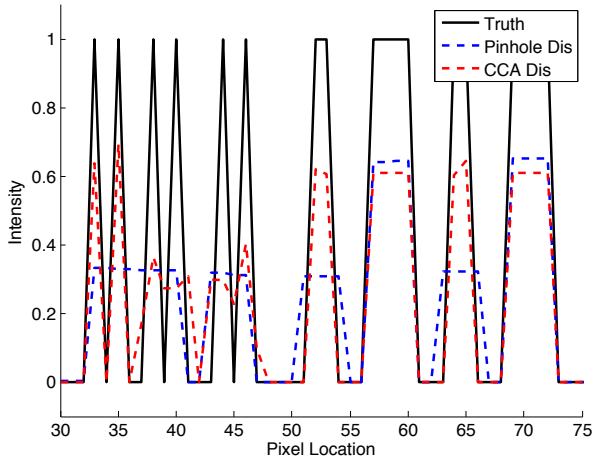


Figure 9. A cross section of the leftmost vertical bars of Fig. 6 of the Pinhole Disambiguation (dashed blue) and CCA Disambiguation (dashed red).

camera and (b) a superimposition coded aperture camera. Coded apertures have added benefit of allowing more light to hit detector than a pinhole would, and so yield less noisy observations. Conventional coded apertures, such as MURAs, are optimal only under the assumption of linear reconstruction (based on convolution). In contrast, we have proved that the proposed coded apertures are optimal when making compressive measurements and allowing for nonlinear reconstruction. We showed through simulation that this allows us to achieve higher spatial resolution than MURA codes when the focal plane array size is kept constant.

ACKNOWLEDGMENTS

The authors have been supported by DARPA Contract No. HR0011-04-C-0111, ONR Grant No. N00014-06-1-0610, DARPA Contract No. HR0011-06-C-0109, and NSF-DMS-08-11062.

REFERENCES

[1] Candès, E., Romberg, J., and Tao, T., "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory* **52**(2), 489 – 509 (2006).

[2] Candès, E. and Tao, T., "Near optimal signal recovery from random projections: Universal encoding strategies," To be published in IEEE Transactions on Information Theory. http://www.acm.caltech.edu/ emmanuel/papers/OptimalRecovery.pdf (2006).

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

[4] Haupt, J. and Nowak, R., "Signal reconstruction from noisy random projections," *IEEE Trans. on Information Theory* **52**(9), 4036–4048 (2006).

[5] Gottesman, S. R. and Fenimore, E. E., "New family of binary arrays for coded aperture imaging," *Appl. Opt.* **28** (1989).

[6] Marcia, R. F. and Willett, R. M., "Compressive coded aperture superresolution image reconstruction," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)* (2008).

[7] Marcia, R. F., Kim, C., Eldeniz, C., Kim, J., Brady, D. J., and Willett, R. M., "Superimposed video disambiguation for increased field of view," *Opt. Express* **16**(21), 16352–16363 (2008).

[8] Candès, E. J. and Tao, T., "Decoding by linear programming," *IEEE Trans. Inform. Theory* **15**(12), 4203–4215 (2005).

[9] Tropp, J. A., "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory* **52**(3), 1030–1051 (2006).

[10] Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G., "Single pixel imaging via compressive sampling," *IEEE Signal Processing Magazine* **25**(2), 83–91 (2008).

[11] Candès, E. J. and Tao, T., "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics* (2005). To appear.

[12] Bajwa, W., Haupt, J., Raz, G., Wright, S., and Nowak, R., "Toeplitz-structured compressed sensing matrices," in [*Proc. of Stat. Sig. Proc. Workshop*], (2007).

[13] Marcia, R. F., Kim, C., Kim, J., Brady, D. J., and Willett, R. M., "Fast disambiguation of superimposed images for increased field of view," in [*Proceedings of the IEEE International Conference on Image Processing*], 2620–2623 (October 2008).

[14] Negahban, S. and Wainwright, M., "Phase transitions for high-dimensional joint support recovery," in [*NIPS*], (2008).

[15] Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M., "A simple proof of the restricted isometry property for random matrices," To appear in *Contructive Approximation* (2007).

[16] Marcia, R. F. and Willett, R. M., "Compressive coded aperture video reconstruction," in [*Proceedings of the 16th European Signal Processing Conference, EUSIPCO 2008*], (August 2008).

[17] Figueiredo, M. A. T., Nowak, R. D., and Wright, S. J., "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing* **1**(4), 586–597 (2007).