Compressive Imaging for Video Representation and Coding

Michael B. Wakin, Jason N. Laska, Marco F. Duarte, Dror Baron, Shriram Sarvotham
Dharmpal Takhar, Kevin F. Kelly, and Richard G. Baraniuk*

Dept. of Electrical and Computer Engineering
Rice University, Houston, TX, USA

Abstract. *Compressive Sensing* is an emerging field based on the revelation that a small group of non-adaptive linear projections of a compressible signal contains enough information for reconstruction and processing. In this paper, we propose algorithms and hardware to support a new theory of *Compressive Imaging*. Our approach is based on a new digital image/video camera that directly acquires random projections of the light field without first collecting the pixels/voxels. Our camera architecture employs a digital micromirror array to perform optical calculations of linear projections of an image onto pseudorandom binary patterns. Its hallmarks include the ability to obtain an image with a single detection element while measuring the image/video fewer times than the number of pixels/voxels; this can significantly reduce the computation required for video acquisition/encoding. Since our system relies on a single photon detector, it can also be adapted to image at wavelengths that are currently impossible with conventional CCD and CMOS imagers. We are currently testing a prototype design for the camera and include experimental results.

Index Terms: camera, compressive sensing, imaging, incoherent projections, linear programming, random matrices, sparsity, video

1 INTRODUCTION

The large amount of raw data acquired in a conventional digital image or video often necessitates immediate compression in order to store or transmit that data. This compression typically exploits a priori knowledge, such as the fact that an N-pixel image can be well approximated as a sparse linear combination of $K \ll N$ wavelets. These appropriate wavelet coefficients can be efficiently computed from the N pixel values and then easily stored or transmitted along with their locations. Similar procedures are applied to videos containing F frames of P pixels each; we let $N = FP$ denote the number of "voxels".

This process has two major shortcomings. First, acquiring large amounts of raw image or video data (large N) can be expensive, particularly at wavelengths where CMOS or CCD sensing technology is limited. Second, compressing raw data can be computationally demanding, particularly in the case of video. While there may appear to be no way around this procedure of "sample, process, keep the important information, and throw away the rest," a new theory known as Compressive Sensing (CS) has emerged that offers hope for directly acquiring a compressed digital representation of a signal without first sampling that signal [1–3].

In this paper, we propose algorithms and hardware to support a new theory of Compressive Imaging (CI). Our approach is based on a new digital image/video camera that directly acquires random projections without first collecting the N pixels/voxels [4]. Due to this unique measurement approach, it has the ability to obtain an image with a single detection element while measuring the image far fewer times than the number of pixels. Because of this single detector, it can be adapted to image at wavelengths that are currently impossible with conventional CCD and CMOS imagers. Our camera can also be used to take streaming measurements of a video signal, which can then be recovered using CS techniques designed for either 2-dimensional (2D) frame-by-frame reconstruction or joint 3D reconstruction. This allows a significant reduction in the computational complexity of the video encoding process.

This paper is organized as follows. Section 2 provides an overview of CS, the theoretical foundation for our CI approach. Section 3 overviews our CI framework and hardware testbed and Section 4 presents experimental results.

* Supported by NSF, ONR, AFOSR, DARPA, and the Texas Instruments Leadership University Program.
Email: {wakin, laska, duarte, drorb, shri, kaka, kkelly, richb}@rice.edu; Web: dsp.rice.edu/cs.

2 COMPRESSIVE SENSING

2.1 Transform coding

CS builds upon a core tenet of signal processing and information theory: that signals, images, and other data often contain some type of *structure* that enables intelligent representation and processing. Current state-of-the-art compression algorithms employ a decorrelating transform to compact a correlated signal's energy into just a few essential coefficients. Such *transform coders* exploit the fact that many signals have a *sparse* representation in terms of some basis Ψ, meaning that a small number K of adaptively chosen transform coefficients can be transmitted or stored rather than $N \gg K$ signal samples. For example, smooth images are sparse in the Fourier basis, and piecewise smooth images are sparse in a wavelet basis; the commercial coding standards JPEG and JPEG2000 and various video coding methods (c.f. Secker and Taubman [5]) directly exploit this sparsity.

The standard procedure for transform coding of sparse signals is to (*i*) acquire the full N-sample signal x; (*ii*) compute the complete set $\{\theta(n)\}$ of transform coefficients $\theta(n) = \langle \psi_n, x \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product; (*iii*) locate the K largest, significant coefficients and discard the (many) small coefficients; and (*iv*) encode the *values and locations* of the largest coefficients. In cases where N is large and K is small, this procedure is quite inefficient. Much of the output of the analog-to-digital conversion process ends up being discarded (though it is not known a priori which pieces are needed).

This raises a simple question: For a given signal, is it possible to directly estimate the set of large coefficients that will not be discarded by the transform coder? While this seems improbable, the recent theory of *Compressive Sensing* introduced by Candès, Romberg, and Tao [1] and Donoho [2] demonstrates that a signal that is K-sparse in one basis (call it the *sparsity basis*) can be recovered from cK *nonadaptive* linear projections onto a second basis (call it the *measurement basis*) that is incoherent with the first, where where c is a small *overmeasuring* constant. While the measurement process is linear, the reconstruction process is decidedly *nonlinear*.

2.2 Incoherent projections

In CS, we do not measure or encode the K significant $\theta(n)$ directly. Rather, we measure and encode $M < N$ projections $y(m) = \langle x, \phi_m^T \rangle$ of the signal onto a *second set* of basis functions $\{\phi_m\}, m \in \{1, 2, \ldots, M\}$, where ϕ_m^T denotes the transpose of ϕ_m. In matrix notation, we measure

$$y = \Phi x, \tag{1}$$

where y is an $M \times 1$ column vector, and the *measurement basis* matrix Φ is $M \times N$ with each row a basis vector ϕ_m. Since $M < N$, recovery of the signal x from the measurements y is ill-posed in general; however the additional assumption of signal *sparsity* makes recovery possible and practical.

The CS theory tells us that when certain conditions hold, namely that the basis $\{\phi_m\}$ cannot sparsely represent the elements of the sparsity-inducing basis $\{\psi_n\}$ (a condition known as *incoherence* of the two bases [1, 2]) and the number of measurements M is large enough, then it is indeed possible to recover the set of large $\{\theta(n)\}$ (and thus the signal x) from a similarly sized set of measurements $\{y(m)\}$. This incoherence property holds for many pairs of bases, including for example, delta spikes and the sine waves of the Fourier basis, or the Fourier basis and wavelets. Significantly, this incoherence also holds with high probability between an arbitrary fixed basis and a randomly generated one (consisting of i.i.d. Gaussian or Bernoulli/Rademacher ± 1 vectors). Signals that are sparsely represented in frames or unions of bases can be recovered from incoherent measurements in the same fashion.

2.3 Signal recovery

The recovery of the sparse set of significant coefficients $\{\theta(n)\}$ can be achieved using *optimization* by searching for the signal with ℓ_0-sparsest[1] coefficients $\{\theta(n)\}$ that agrees with the M observed measurements in y (recall that $M < N$). Unfortunately, solving this ℓ_0 optimization problem is prohibitively complex and is believed to be NP-hard [6]. The practical revelation that supports the new CS theory is that it is not necessary to solve the ℓ_0-minimization problem to recover the set of significant $\{\theta(n)\}$. In fact, a much easier problem yields an equivalent solution (thanks again to the incoherency of the bases); we need only solve for the ℓ_1-sparsest coefficients θ that agree with the measurements y [1, 2]

$$\widehat{\theta} = \arg\min \|\theta\|_1 \quad \text{s.t. } y = \Phi\Psi\theta. \tag{2}$$

[1] The ℓ_0 "norm" $\|\theta\|_0$ merely counts the number of nonzero entries in the vector θ.

This optimization problem, also known as *Basis Pursuit* [7], is significantly more approachable and can be solved with traditional linear programming techniques whose computational complexities are polynomial in N. Although only $K + 1$ measurements are required to recover sparse signals via ℓ_0 optimization [8], one typically requires $M \geq cK$ measurements for Basis Pursuit with an overmeasuring factor $c > 1$.

Unfortunately, linear programming techniques are still somewhat slow. At the expense of slightly more measurements, fast iterative greedy algorithms have also been developed to recover the signal x from the measurements y. Examples include the iterative Orthogonal Matching Pursuit (OMP) [9], matching pursuit (MP), and tree matching pursuit (TMP) [10] algorithms. Group testing [11] has been shown to yield even faster reconstruction algorithms. All of these methods have also been shown to perform well on *compressible signals*, which are not exactly K-sparse but are well approximated by a K-term representation. Such a model is more realistic in practice.

3 COMPRESSIVE IMAGING

In this paper, we develop a new system to support what can be called *Compressive Imaging* (CI). Our system incorporates a microcontrolled mirror array driven by pseudorandom and other measurement bases and a single or multiple photodiode optical sensor. This hardware optically computes incoherent image measurements as dictated by the CS theory; we then apply CS reconstruction algorithms to obtain the acquired images. Our camera can also be used to take streaming measurements of a video signal, which can then be recovered using CS techniques designed for either 2D frame-by-frame reconstruction or joint 3D reconstruction.

Other desirable features of our system include the use of a single detector (potentially enabling imaging at new wavelengths that are currently impossible with CCD and CMOS technology), universal measurement bases (incoherent with arbitrary sparse bases), encrypted measurements (tied to a random seed that can be kept secure), and scalable progressive reconstruction (yielding improved quality with more measurements) [4].

3.1 Camera hardware

Our hardware realization of the CI concept is a *single pixel camera*; it combines a microcontrolled mirror array displaying a time sequence of M pseudorandom basis images ϕ_m with a single optical sensor to compute incoherent image measurements y as in (1) (see Figure 1). By adaptively selecting how many measurements to compute, we can trade off the amount of compression versus acquisition time; in contrast, conventional cameras trade off resolution versus the number of pixel sensors.

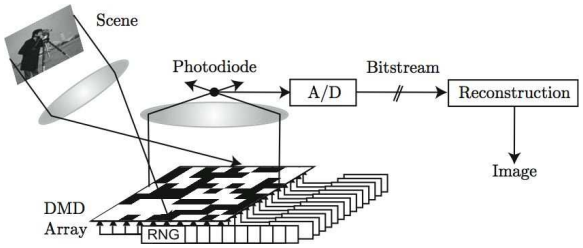


Fig. 1. *Compressive Imaging (CI) camera. Incident lightfield (corresponding to the desired image x) is reflected off a digital micromirror device (DMD) array whose mirror orientations are modulated in the pseudorandom pattern ϕ_m supplied by the random number generators (RNG). Each different mirror pattern produces a voltage at the single photodiode that corresponds to one measurement $y(m)$.*

We employ a Texas Instruments digital micromirror device (DMD) for generating the random basis patterns. The DMD consists of a 1024×768 array of electrostatically actuated micromirrors where each mirror of the array is suspended above an individual SRAM cell. Each mirror rotates about a hinge and can be positioned in one of two states ($+12$ degrees and -12 degrees from horizontal); thus light falling on the DMD may be reflected in two directions depending on the orientation of the mirrors.

With the help of a biconvex lens, the desired image is formed on the DMD plane; this image acts as an object for the second biconvex lens, which focuses the image onto the photodiode. The light is collected from one of the two directions in which it is reflected (e.g., the light reflected by mirrors in the $+12$ degree state). The light from a given configuration of the DMD mirrors is summed at the photodiode to yield an absolute voltage that yields a coefficient $y(m)$ for that configuration. The output of the photodiode is amplified through an op-amp circuit and then digitized by a 12-bit analog-to-digital converter. These photodiode measurements can be interpreted as the inner product of the desired image x with a measurement basis vector ϕ_m. In particular, letting $\rho(m)$ denote the mirror positions of the m-th measurement pattern, the voltage reading

from the photodiode v can be written as

$$v(m) \propto \langle x, \phi_m \rangle + \text{DC offset}, \tag{3}$$

where

$$\phi_m = \mathbf{1}_{\{\rho(m)=+12 \text{ degrees}\}} \tag{4}$$

and $\mathbf{1}$ is the indicator function. (The DC offset can be measured by setting all mirrors to -12 degrees; it can then subtracted off.)

Equation (3) holds the key for implementing a CI system. For a given image x, we take M measurements $\{y(1), y(2), \ldots, y(M)\}$ corresponding to mirror configurations $\{\rho(1), \rho(2), \ldots, \rho(M)\}$. Since the patterns $\rho(m)$ are programmable, we can select them to be incoherent with the sparsity-inducing basis (e.g., wavelets or curvelets). As mentioned previously, random or pseudorandom measurement patterns enjoy a useful universal incoherence property with any fixed basis, and so we employ pseudorandom ± 12 degree patterns on the mirrors. These correspond to pseudorandom 0/1 Bernoulli measurement vectors $\phi_m = \mathbf{1}_{\{\rho(m)=+12 \text{ degrees}\}}$. (The measurements may easily be converted to ± 1 Rademacher patterns by setting all mirrors in $\rho(1)$ to $+12$ degrees and then letting $y(m) \leftarrow 2y(m) - y(1)$ for $m > 1$.) Other options for incoherent CI mirror patterns include $-1/0/1$ group-testing patterns [11]. Mirrors can also be duty-cycled to give the elements of ϕ finer precision, for example to approximate Gaussian measurement vectors [2, 3].

This system directly acquires a reduced set of M incoherent projections of an N-pixel image x *without* first acquiring the N pixel values. Since the camera is "progressive," better quality images (larger K) can be obtained by taking a larger number of measurements M. Also, since the data measured by the camera is "future-proof," new reconstruction algorithms based on better sparsifying image transforms can be applied at a later date to obtain even better quality images.

3.2 Streaming video acquisition

Our CI system is immediately applicable to video acquisition. The key is that, as described above, the measurements $\{\phi_m\}$ are taken sequentially in time. Hence, one can view each measurement as a linear projection against a snapshot of the scene at that instant. Viewing the video as a 3D signal (in which the 2D snapshots are stacked), the measurements vectors $\{\phi(m)\}$ themselves are each localized onto a different 2D snapshot for each m.

In order to recover a video sequence from these measurements, we make some simplifying assumptions. Specifically, traditional CS considers an ensemble of measurements taken from a single signal; in our streaming setting, however, each measurement will act on a different snapshot. We can overcome this problem by assuming that the image changes slowly across a group of snapshots, which we can then equate to a single *video frame*. The number of snapshots assigned to a frame will be determined by the speed of our acquisition system and the desired temporal resolution of the reconstructed video. Under this assumption, we represent the video acquired as a sequence of F frames, each one measured using M/F measurement vectors that we can group as rows of a matrix Φ_i, $i = 1, \ldots, F$.

We have several options for reconstructing the video from these measurements. First, we could reconstruct each frame using 2D wavelets, performing a total of F CI reconstructions. Each reconstruction would use the same 2D wavelet sparsity basis Ψ but with a different measurement matrix Φ_i. We refer to this process as *frame-by-frame* reconstruction.

Alternative methods more fully exploit the correlation between frames. One solution is to use 3D wavelets as a sparse representation for the video sequence; i.e., to define the joint measurement matrix

$$\Phi = \begin{bmatrix} \Phi_1 & 0 & \ldots & 0 \\ 0 & \Phi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Phi_F \end{bmatrix}$$

for the video sequence and then perform *joint reconstruction* of the entire video sequence using a 3D wavelet sparse basis Ψ for the frame ensemble. As we see in Section 4, despite its block diagonal structure, the 3D measurement matrix Φ enjoys sufficient incoherence with the 3D sparsity matrix Ψ.

Future work may consider extending our imaging architecture to acquire full 3D measurements of a video sequence (that is, where each ϕ_m has 3D support). Under this setting, we reconstruct the entire video sequence using a single measurement matrix Φ that operates on all of the frames and a suitable 3D sparse basis Ψ such as wavelets. In Section 4 we demonstrate that such a scheme would enjoy better incoherence with the video structure. However, it also increases the complexity of both the measurement and reconstruction processes. Possible solutions to this increased complexity include partitioning the video into blocks, which are then reconstructed separately.

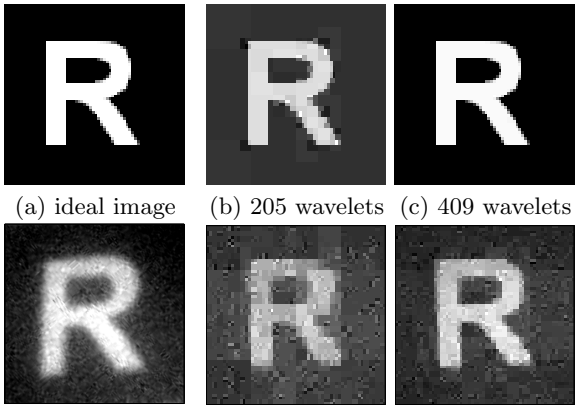
3.3 Related work

Other efforts on CI include [12, 13], which employ optical elements to perform transform coding of multispectral images. The elegant hardware designed for these purposes uses concepts that include optical projections, group testing [11], and signal inference. Two notable previous DMD-driven applications involve confocal microscopy [14] and micro-optoelectromechanical (MOEM) systems [15]. For more about related work, see [4].

4 EXPERIMENTAL RESULTS

4.1 Still image acquisition

For our imaging experiment, we displayed a printout of the letter "R" in front of the camera; Figure 2(a) shows the printout. For acquisition and reconstruction, we use an imaging resolution of $N = 64 \times 64 = 4096$. Since our test image is piecewise constant (with sharp edges) it can be sparsely represented in the wavelet domain. Figures 2(b) and 2(c) show the best K-term Haar wavelet approximation of the idealized image in Figure 2(a) with $K = 205$ and 409, respectively. Using $M = 819$ and 1,638 measurements (roughly $4\times$ the K used in (b) and (c)), we reconstructed the images shown in Figures 2(e) and 2(f) using the Dantzig Selector [16], a robust scheme for CS reconstruction. This preliminary experiment confirms the feasibility of the CI approach; we are currently working to resolve minor calibration and noise issues to improve the reconstruction quality.



(a) ideal image (b) 205 wavelets (c) 409 wavelets

(d) image on DMD (e) 819 meas. (f) 1,638 meas.

Fig. 2. *CI DMD imaging of a* 64×64 *(*$N = 4096$ *pixel) image. Ideal image (a) of full resolution and approximated by its (b) largest 205 wavelet coefficients and (c) largest 409 wavelet coefficients. (d) Conventional* 320×240 *camera image acquired at the DMD plane. CS reconstruction using Dantzig Selector from (e) 819 random measurements and (f) 1,638 random measurements. In all cases Haar wavelets were used for approximation or reconstruction.*

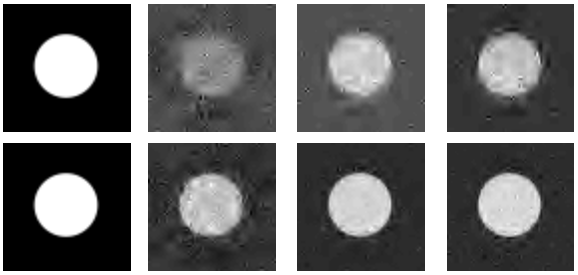
4.2 Video simulation

To demonstrate the potential for applications in video encoding, we present a series of simulations for video measurement/reconstruction. Figure 3(a) shows a single frame taken from our $F = 64$ frame video sequence that consists of $P = 64 \times 64$ images; in total the video contains $N = FP = 262{,}144$ 3D voxels. The video shows a disk moving from top to bottom and growing from small to large. We measure this video sequence using a total of M measurements, either 2D random measurements (with M/F measurements/frame) or 3D random measurements. (For the 2D measurements, we make the simplifying assumption that the image remains constant across all snapshots within a given frame.) To reconstruct the video from these measurements we compare two approaches: 2D frame-by-frame reconstruction using 2D wavelets as a sparsity-inducing basis and 3D joint reconstruction using 3D wavelets as a sparsity-inducing basis.

Figure 3 shows Matching Pursuit reconstruction results using $M = 20{,}000$ (top row) and $M = 50{,}000$ (bottom row). Comparing columns (b) and (c), we observe that 3D wavelets offer a significant improvement in reconstruction quality over 2D wavelets; we attribute this improvement to the ability of 3D wavelets to capture correlations between frames. Comparing columns (c) and (d), we also observe that full 3D measurements allow better reconstruction than frame-by-frame 2D measurements; we believe this improvement is due to the better incoherency between the measurement basis and the wavelet basis. Fortunately, this improvement is somewhat moderate, which indicates that 2D frame-by-frame measurements (easily obtained from our hardware) may contain sufficient information for high-quality video reconstruction, presuming that a *joint* 3D technique is used for reconstruction. Current work focuses on developing better joint reconstruction techniques, perhaps by extending our algorithms for Distributed CS [8].

5 DISCUSSION AND CONCLUSIONS

In this paper, we have presented a prototype imaging system that successfully employs compressive



(a) frame 32 (b) 2D meas 2D recon (c) 2D meas 3D recon (d) 3D meas 3D recon

Fig. 3. *Frame 32 from reconstructed video sequence using (top row) $M = 20{,}000$ and (bottom row) $M = 50{,}000$ measurements. (a) Original frame. (b) Frame-by-frame 2D measurements; frame-by-frame 2D reconstruction; MSE = 3.63 and 0.82. (c) Frame-by-frame 2D measurements; joint 3D reconstruction; MSE = 0.99 and 0.24. (d) Joint 3D measurements; joint 3D reconstruction; MSE = 0.76 and 0.18. The results in (d) are comparable to the MSE obtained by wavelet thresholding with $K = 655$ and 4000 coefficients, respectively.*

sensing principles. The camera has many attractive features, including simplicity, universality, robustness, and scalability, that should enable it to impact a variety of different applications. An interesting and potentially useful practical feature of our system is that it off-loads processing from data collection into data reconstruction. Not only will this lower the complexity and power consumption of the sensing device, but it will also enable new adaptive measurement schemes. Another intriguing feature of the system is that, since it relies on a single photon detector, it can be adapted to image at wavelengths that are currently impossible with conventional CCD and CMOS imagers. Finally, our imaging system is immediately extensible to video acquisition, providing streaming measurements that can then be processed using frame-by-frame 2D or joint 3D CS techniques.[2]

References

1. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52** (2006) 489–509
2. Donoho, D.: Compressed sensing. (2004) Preprint.
3. Candès, E., Tao, T.: Near optimal signal recovery from random projections and universal encoding strategies. (2004) Preprint.
4. Takhar, D., Laska, J.N., Wakin, M., Duarte, M., Baron, D., Sarvotham, S., Kelly, K.K., Baraniuk, R.G.: A new camera architecture based on optical-domain compression. In: Proc. IS&T/SPIE Symposium on Electronic Imaging: Computational Imaging. Volume 6065. (2006)
5. Secker, A., Taubman, D.S.: Highly scalable video compression with scalable motion coding. IEEE Trans. Image Processing **13** (2004) 1029–1041
6. Candès, E., Tao, T.: Error correction via linear programming. (2005) Preprint.
7. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. SIAM J. on Sci. Comp. **20** (1998) 33–61
8. Baron, D., Wakin, M.B., Duarte, M.F., Sarvotham, S., Baraniuk, R.G.: Distributed compressed sensing. (2005) Available at http://www.dsp.rice.edu/cs.
9. Tropp, J., Gilbert, A.C.: Signal recovery from partial information via orthogonal matching pursuit. (2005) Preprint.
10. Duarte, M.F., Wakin, M.B., Baraniuk, R.G.: Fast reconstruction of piecewise smooth signals from random projections. In: Proc. SPARS05, Rennes, France (2005)
11. Cormode, G., Muthukrishnan, S.: Towards an algorithmic theory of compressed sensing. DIMACS Tech. Report 2005-40 (2005)
12. Pitsianis, N.P., Brady, D.J., Sun, X.: Sensor-layer image compression based on the quantized cosine transform. In: SPIE Visual Information Processing XIV. (2005)
13. Brady, D.J., Feldman, M., Pitsianis, N., Guo, J.P., Portnoy, A., Fiddy, M.: Compressive optical MONTAGE photography. In: SPIE Photonic Devices and Algorithms for Computing VII. (2005)
14. Lane, P.M., Elliott, R.P., MacAulay, C.E.: Confocal microendoscopy with chromatic sectioning. In: Proc. SPIE. Volume 4959. (2003) 23–26
15. DeVerse, R.A., Coifman, R.R., Coppi, A.C., Fateley, W.G., Geshwind, F., Hammaker, R.M., Valenti, S., Warner, F.J.: Application of spatial light modulators for new modalities in spectrometry and imaging. In: Proc. SPIE. Volume 4959. (2003)
16. Candès, E., Tao, T.: The Dantzig selector: Statistical estimation when p is much larger than n. (2005) Preprint.

[2] Thanks to Texas Instruments for providing the TI DMD developer's kit and accessory light modulator package (ALP). Thanks also to Dave Brady and Dennis Healy for enlightening discussions.