High spatio-temporal resolution video with compressed sensing

Roman Koller,[1] Lukas Schmid,[1] Nathan Matsuda,[2] Thomas Niederberger,[1] Leonidas Spinoulas,[2] Oliver Cossairt,[2,*] Guido Schuster[1] and Aggelos K. Katsaggelos[2]

[1] *University of Applied Sciences of Eastern Switzerland, Oberseestrasse 10, CH-8640, Rapperswil, Switzerland*
[2] *Northwestern University, 2145 Sheridan Road, Evanston, IL, 60208, USA*
[*] *olivercossairt@gmail.com*

Abstract: We present a prototype compressive video camera that encodes scene movement using a translated binary photomask in the optical path. The encoded recording can then be used to reconstruct multiple output frames from each captured image, effectively synthesizing high speed video. The use of a printed binary mask allows reconstruction at higher spatial resolutions than has been previously demonstrated. In addition, we improve upon previous work by investigating tradeoffs in mask design and reconstruction algorithm selection. We identify a mask design that consistently provides the best performance across multiple reconstruction strategies in simulation, and verify it with our prototype hardware. Finally, we compare reconstruction algorithms and identify the best choice in terms of balancing reconstruction quality and speed.

OCIS codes: (110.1758) Computational imaging; (100.3010) Image reconstruction techniques; (110.6915) Time imaging; Compressive sampling.

References and links

1. P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," Opt. Express **21**, 10526–10545 (2013).
2. M. Ben-Ezra and S. K. Nayar, "Motion-based Motion Deblurring," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 689–698 (2004).
3. M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Flexible voxels for motion-aware videography," in "Proceedings of European Conference on Computer Vision," (2010), pp. 100–114.
4. G. Bub, M. Tecza, M. Helmes, P. Lee, and P. Kohl, "Temporal pixel multiplexing for simultaneous high-speed, high-resolution imaging," Nature Methods **7**, 209–U66 (2010).
5. B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," ACM Trans. Graph. **24**, 765–776 (2005).
6. A. Agrawal, M. Gupta, A. Veeraraghavan, and S. G. Narasimhan, "Optimal coded sampling for temporal super-resolution," in "Proceedings of IEEE Conference on Computer Vision and Pattern Recognition," (2010), pp. 599–606.
7. R. Pournaghi and X. Wu, "Coded acquisition of high frame rate video," IEEE Trans. Image Process. **23**, 5670–5682 (2014).
8. M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," IEEE Signal Process. Mag. **25**, 83–91 (2008).
9. A. C. Sankaranarayanan, C. Studer, and R. G. Baraniuk, "CS-MUVI: Video compressive sensing for spatial-multiplexing cameras," in "Proceedings of IEEE International Conference on Computational Photography," (2012), pp. 1–10.
10. A. C. Sankaranarayanan, C. Studer, and R. G. Baraniuk, "Video compressive sensing for spatial-multiplexing cameras using motion-flow models," SIAM J. Imag. Sci. (under review) .

11. Z. T. Harmany, R. F. Marcia, and R. M. Willett, "Dual-Scale masks for spatio-temporal compressive imaging," in "Proceedings of IEEE Global Conference on Signal and Information Processing," (2013), pp. 1045–1048.
12. L. Xu, A. Sankaranarayanan, C. Studer, Y. Li, R. G. Baraniuk, and K. F. Kelly, "Multi-Scale compressive video acquisition," in "Imaging and Appl. Opt.", (Optical Society of America, 2013), p. CW2C.4.
13. L. McMackin, M. A. Herman, B. Chatterjee, and M. Weldon, "A high-resolution SWIR camera via compressed sensing," Proc. SPIE **8353**, 835303 (2012).
14. N. Gopalsami, S. Liao, T. W. Elmer, E. R. Koehl, A. Heifetz, A. C. Raptis, L. Spinoulas, and A. K. Katsaggelos, "Passive millimeter-wave imaging with compressive sensing," Opt. Eng. **51**, 091614–1–091614–9 (2012).
15. L. Spinoulas, J. Qi, A. K. Katsaggelos, T. W. Elmer, N. Gopalsami, and A. C. Raptis, "Optimized compressive sampling for passive millimeter-wave imaging," Appl. Opt. **51**, 6335–6342 (2012).
16. A. Veeraraghavan, D. Reddy, and R. Raskar, "Coded Strobing Photography: Compressive sensing of high speed periodic videos," IEEE Trans. Pattern Anal. Mach. Intell. **33**, 671–686 (2011).
17. D. Reddy, A. Veeraraghavan, and R. Chellappa, "P2C2: Programmable pixel compressive camera for high speed imaging," in "Proceedings of IEEE Conference on Computer Vision and Pattern Recognition," (2011), pp. 329–336.
18. D. Liu, J. Gu, Y. Hitomi, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Efficient space-time sampling with pixel-wise coded exposure for high speed imaging," IEEE Trans. Pattern Anal. Mach. Intell. **99**, 1 (2013).
19. J. Holloway, A. C. Sankaranarayanan, A. Veeraraghavan, and S. Tambe, "Flutter shutter video camera for compressive sensing of videos," in "Proceedings of IEEE International Conference on Computational Photography," (2012), pp. 1–9.
20. X. Lin, J. Suo, G. Wetzstein, Q. Dai, and R. Raskar, "Coded focal stack photography," in "Proceedings of IEEE International Conference on Computational Photography," (2013), pp. 1–9.
21. T. Portz, L. Zhang, and H. Jiang, "Random coded sampling for high-speed HDR video," in "Proceedings of IEEE International Conference on Computational Photography," (2013), pp. 1–8.
22. L. Gao, J. Liang, C. Li, and L. V. Wang, "Single-Shot compressed ultrafast photography at one hundred billion frames per second," Nature **516**, 74–77 (2014).
23. K. Nakagawa, A. Iwasaki, Y. Oishi, R. Horisaki, A. Tsukamoto, N. A., H. K., H. Lia, T. Ushida, K. Goda, F. Kannari, and I. Sakuma, "Sequentially timed all-optical mapping photography (STAMP)," Nat. Photonics **8**, 695–700 (2014).
24. N. Katic, M. H. Kamal, A. Schmid, P. Vandergheynst, and Y. Leblebici, "Compressive image acquisition in modern CMOS IC design," Int. J. Circ. Theor. App. (2013).
25. Y. Oike and A. El Gamal, "CMOS image sensor with per-column ΣΔ; ADC and programmable compressed sensing," IEEE J. Solid-State Circuits **48**, 318–328 (2013).
26. R. Robucci, J. D. Gray, L. K. Chiu, J. Romberg, and P. Hasler, "Compressive sensing on a CMOS separable-transform image sensor," Proc. IEEE **98**, 1089–1101 (2010).
27. G. Orchard, J. Zhang, Y. Suo, M. Dao, D. T. Nguyen, S. Chin, C. Posch, T. D. Tran, and R. Etienne-Cummings, "Real time compressive sensing video reconstruction in hardware," IEEE Trans. Emerg. Sel. Topics Circuits Syst. **2**, 604–615 (2012).
28. C. Fernandez-Cull, B. M. Tyrrell, R. D'Onofrio, A. Bolstad, J. Lin, J. W. Little, M. Blackwell, M. Renzi, and M. Kelly, "Smart pixel imaging with computational-imaging arrays," Proc. SPIE **9070**, 90703D–90703D–13 (2014).
29. X. Yuan, P. Llull, X. Liao, J. Yang, D. J. Brady, G. Sapiro, and L. Carin, "Low-Cost compressive sensing for color video and depth," in "Proceedings of IEEE Conference on Computer Vision and Pattern Recognition," (2014), pp. 3318–3325.
30. M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing* (Springer Publishing Company, Incorporated, 2010), 1st ed.
31. Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in "Proceedings of IEEE International Conference on Computer Vision," (2011).
32. A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imag. Sci. **2**, 183–202 (2009).
33. S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," IEEE J. Sel. Topics in Signal Process. **1**, 606–617 (2007).
34. X. Liao, H. Li, and L. Carin, "Generalized Alternating Projection for weighted-2,1 minimization with applications to model-based compressive sensing," SIAM J. Imag. Sci. **7**, 797–823 (2014).
35. M. Elad, "Optimized projections for compressed sensing," IEEE Trans. Signal Process. **55**, 5695–5702 (2007).
36. E. Tsiligianni, L. Kondi, and A. Katsaggelos, "Construction of incoherent unit norm tight frames with application to compressed sensing," IEEE Trans. Inf. Theory **60**, 2319–2330 (2014).

1. Introduction

The subdivision of time by motion picture cameras, the frame-rate, limits the temporal resolution of a camera system. Even though frame-rates above 30 Hz are widely recognized to be imperceptible to human eyes, high speed motion picture capture has long been a goal in scientific imaging and cinematography communities. The ability to resolve motion beyond human perception has great scientific and aesthetic value, as demonstrated by the recent popularity of high frame-rate videos available online. Although video capture rates have increased as hardware prices have fallen, fundamental constraints still limit maximum achievable frame-rates as well as the cost and availability of high speed cameras. Recent advances in compressed sensing have enabled novel methods for achieving ever increasing frame-rates beyond those possible by directly sampling a scene in hardware. In this paper, we compare and optimize the hardware and algorithmic components of compressive video capture. We demonstrate a prototype compressive video camera capable of recovering 2 Mpixel videos at frame-rates in excess of 740 frames-per-second (fps) using commodity CMOS sensors. Using our system, an effective frame-rate increase by a factor of $10\times$ is possible.

We use a high-resolution printed coded aperture mask placed on a fast moving translation stage to create spatio-temporal modulation. We incorporate this into a forward model used to reconstruct multiple frames of video from a single coded, captured image. Our approach has several advantages over previously presented techniques. First, the coded mask is a relatively inexpensive optical component, bringing compressive video capture much closer to commercial viability. Second, our printed masks are fabricated at very high resolutions (*e.g.*, pixel size of $4.5 \times 4.5\ \mu m^2$), leading to a 1-1 mapping between the mask and the sensor. Using these masks in conjunction with high-resolution, off-the-shelf CMOS sensors allows us to achieve compressive video capture with exceptionally high spatial resolution. The contributions of our work are as follows:

1.1. Contributions

- **Large scale system**: We demonstrate compressive video using a 2 Mpixel CMOS sensor, nearly $8\times$ increase in reconstructed image dimensions over previous systems.
- **Analysis of mask and algorithm selection**: We explore various mask designs and algorithm implementations, reporting on their performance through simulations and analyzing their tradeoffs between reconstruction quality and time. We introduce a new mask pattern that proves to perform better than earlier systems (*e.g.*, Llull et al. [1]) while optimizing over its design parameter space.
- **Hardware prototype**: We show results of 2 Mpixel compressive videos at 743 fps, captured using our prototype camera and outfitted with the proposed mask patterns.

1.2. Related work

There is a long history of research in using computational methods to increase camera frame-rates. Ben-Ezra et al. [2] used a hybrid approach that combines a low-speed, high-resolution camera with a high-speed, low-resolution camera. Gupta et al. used a high-speed Digital Light Processing (DLP) projector coupled with a low-speed camera to increase its native frame-rate [3]. Bub et al. employed a similar approach to increase the frame-rate of microscopy systems [4] by using a DLP to modulate a relayed image of the sample. Wilburn et al. [5] and Agarwal et al. [6] utilized camera arrays to capture high speed video. Camera arrays, in conjunction with coded exposure, have also been used in [7] to acquire high frame-rate video. For all the techniques mentioned above, frame-rate speedup is achieved by either sacrificing spatial resolution, or by increasing the number of cameras used.

More recently, a number of researchers have developed systems capable of recovering high frame-rate video using compressive coded measurements. These techniques use a single camera system with resolution limited by the code feature size. At the heart of these techniques is the principle that an underdetermined system of equations can be solved accurately when the underlying signal contains sufficient sparsity. In the context of video capture, this amounts to recovering several frames of video from a small set of measurements, which has been applied using a variety of methods. These methods utilize per-pixel coding which proves superior to per-camera coded exposure due to the sparsity induced on space as well, see [7] for a theoretical analysis.

The single pixel camera from Rice has been demonstrated at video rates, and compression has been achieved in both space [8] and more recently, time [9]. Several extensions of these approaches have been proposed, ranging from improved algorithmic design [10] and mask optimization [11] to non-visible spectrum sensing [12]. The single pixel camera is most useful when imaging with particularly expensive detectors (*e.g.*, Infrared [12, 13] or Terahertz [14, 15]), but does not take advantage of the massively parallel sensing capabilities of silicon imager arrays. Several researchers have developed compressive video systems that incorporate high-speed spatio-temporal optical modulation with high resolution CMOS and CCD arrays. Successful reconstruction of high-speed periodic videos has been demonstrated in [16] using strobing photography. Reddy et al. [17] and Liu et al. [18] used fast-switching Liquid Crystal on Silicon (LCoS) Spatial Light Modulators (SLMs) to provide spatio-temporal modulation at speeds much greater than typical frame-rates of CMOS/CCD sensors. These techniques recover a set of high-speed video frames from a single coded photograph using compressive sensing reconstruction techniques. However, the achieved spatial resolution is limited by the resolution of the optical modulator ($\ll$ 1Mpixel in these experiments). Furthermore, inclusion of an SLM can dramatically increase system cost and power consumption, presenting a large barrier to adoption outside academic settings.

A few techniques have been introduced that omit the need for an SLM. Holloway et al. used temporal modulation only (*i.e.*, a flutter shutter) to recover compressive video using only fast-speed switching modes on a commodity sensor [19]. Such switching modes have also proven successful in capturing focal stacks [20] or high-speed high dynamic range video [21]. In a recent Nature publication, a static mask pattern displayed on an SLM to reconstruct 10 picosecond resolution video of non-periodic events was demonstrated [22]. A comparable temporal resolution was also recently achieved using temporal and spatial dispersive elements [23]. The aforementioned techniques, however, require complicated optical elements and expensive cameras (*e.g.*, streak camera in [22]) which are prohibitively costly for all but a small number of applications.

Finally, a few approaches have been proposed in order to directly apply compressive sampling on the sensor circuitry without the use of additional optical elements. Such techniques have mainly been presented for imaging applications, implementing spatial multiplexing, where multiplication of the single image with pseudorandom sequences is applied in the analog domain [24, 25] or by directly applying an on-chip image sparsifying transformation [26]. For video applications, a compressive CMOS design, based on an Asynchronous Time Image Sensor has been proposed by [27] while a CMOS prototype with custom per-pixel read-out modes has recently been showcased by [28].

The work most similar to ours is by Llull et al., who first used a printed coded aperture mask placed on a translation stage to create spatio-temporal modulation in lieu of an SLM [1] and later extended the same architecture for extracting color video and depth [29]. Our work introduces several improvements over their system, including increased spatial dimensions of the reconstructed frames and optimized mask design.

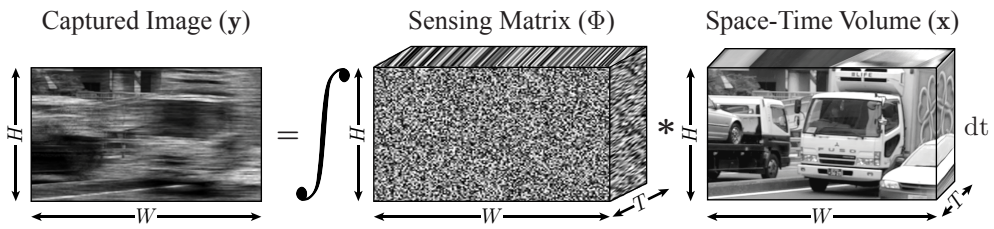


Fig. 1: **The forward sensing model used in this paper**. A space-time volume $\mathbf{x}$ consisting of a set of T frames, with $H \times W$ pixels each, is multiplied by a set of mask patterns embedded within the sensing matrix Φ. The sensor integrates over time, producing a single coded captured image $\mathbf{y}$ consisting of $H \times W$ pixels.

2. Forward model

Our forward model maps a coded spatio-temporal volume onto a single image using the linear projective model (see Fig. 1):

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

where $\mathbf{y} \in \mathbb{R}_+^{(H\cdot W)\times 1}$ is a lexicographically ordered version of the sensed image consisting of $H \cdot W$ pixels, $\Phi \in \mathbb{R}_+^{(H\cdot W)\times(H\cdot W\cdot T)}$ is the coding matrix that maps a set of T video frames of $H \cdot W$ pixels onto a single $H \times W$ pixel image, and $\mathbf{x} \in \mathbb{R}_+^{(H\cdot W\cdot T)\times 1}$ is the unknown space-time volume that we desire to recover. The elements of matrix Φ are determined by the optics of the system, whose design constraints could be optimized to improve performance. For the compressive video systems considered in this paper, the observations are modulated by a set of 2D mask patterns with the same resolution as the sensor (*i.e.*, the feature size of the coded mask and the sensor pixel size are identical).

Let the 2D mask patterns be denoted by the matrices $M_i \in \mathbb{R}_+^{H\times W}, i \in \{1, \cdots T\}$. The mask patterns are then used to populate a new matrix $\mathbb{M}_i = diag(vec(M_i)) \in \mathbb{R}_+^{H\cdot W\times H\cdot W}$, where function $vec(M)$ lexicographically reorders the $H \times W$ matrix M into a $H \cdot W \times 1$ vector, and $diag(\mathbf{x})$ creates an $H \cdot W \times H \cdot W$ matrix by placing the $H \cdot W$ entries of the vector $\mathbf{x}$ along the main diagonal and zeros everywhere else. The coding matrix can then be expressed as a concatenation of these new matrices as,

$$\Phi = \begin{bmatrix}\mathbb{M}_0 & \mathbb{M}_1 & \cdots & \mathbb{M}_T\end{bmatrix}. \tag{2}$$

For the approaches of Reddy et al. [17] and Liu et al. [18], each of the mask patterns M_i may be chosen independently by programming an SLM to switch between different patterns at high speeds. For the work of Llull et al. [1] and Giu et al. [22], as well as the prototype camera described in this paper, a translating static mask pattern is used. One of the disadvantages of the translating mask-based technique is that the displayed patterns are more restrictive compared to the ones possible using an SLM. As a result, the mask patterns M_i are not necessarily independent, and can be described by $M_i = s_i(M)$, where the operator $s_i(M)$ shifts the single native mask pattern M by i columns to the right.

3. Reconstruction algorithms

In an attempt to achieve an implementation agnostic characterization of mask designs, we tested a number of different reconstruction algorithms. Candidate reconstruction algorithms were selected for their utility in solving the under-determined system of Eq. (1). Most approaches rely on enforcing fidelity to the data using a quadratic term as well as employing a regularization functional $F(\cdot)$ which promotes sparsity of the unknown signal $\mathbf{x}$ on some chosen transform

domain. The selection of $F(\cdot)$ aims at encouraging consistency with natural image priors, *i.e.*, the following representation is sought after,

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{y} - \Phi\Psi^{-1}\mathbf{a}\|_2^2 + \lambda F(\mathbf{a}). \tag{3}$$

In Eq. (3), Ψ represents the transform to the chosen domain resulting in a sparse representation $\mathbf{a}$, such that $\mathbf{x}$ is given by $\mathbf{x} = \Psi^{-1}\mathbf{a}$, and $\lambda > 0$ is a tunable regularization parameter. We use standard natural image representations (*e.g.*, discrete cosine transform (DCT), learned dictionaries or image gradients). We evaluated the following optimization algorithms as potential solvers:

ℓ_0 **regularization**: In ℓ_0 regularization algorithms the functional is selected as $F = \|\mathbf{a}\|_0$ and the problem in Eq. (3) can be reformulated as,

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{a}\|_0 \qquad \text{subject to} \qquad \|\mathbf{y} - \Phi\Psi^{-1}\mathbf{a}\|_2^2 \leq \varepsilon, \tag{4}$$

where ε is an error tolerance that defines the trade-off between quality and runtime requirements and is related to the regularization parameter λ of Eq. (3). This is a NP-hard problem [30] and can be solved using Greedy methods such as Orthogonal Matching Pursuit (OMP). OMP in combination with learned dictionaries has been introduced by [31] specifically for this type of problem. We used the implementation OMP-Box v10 from http://www.cs.technion.ac.il/ ronrubin/software.html. Reconstruction is performed block-wise on overlapping sets of 7×7 patches of pixels, using a learned overcomplete dictionary consisting of $10,000$ atoms introduced in [31], provided courtesy of the authors. In this case, Ψ^{-1} in Eq. (4), denotes the dictionary.

ℓ_1 **regularization**: Using $F = \|\mathbf{a}\|_1$, Eq. (3) can be reformulated into the LASSO (Least-Absolute Shrinkage and Selection Operator) problem as,

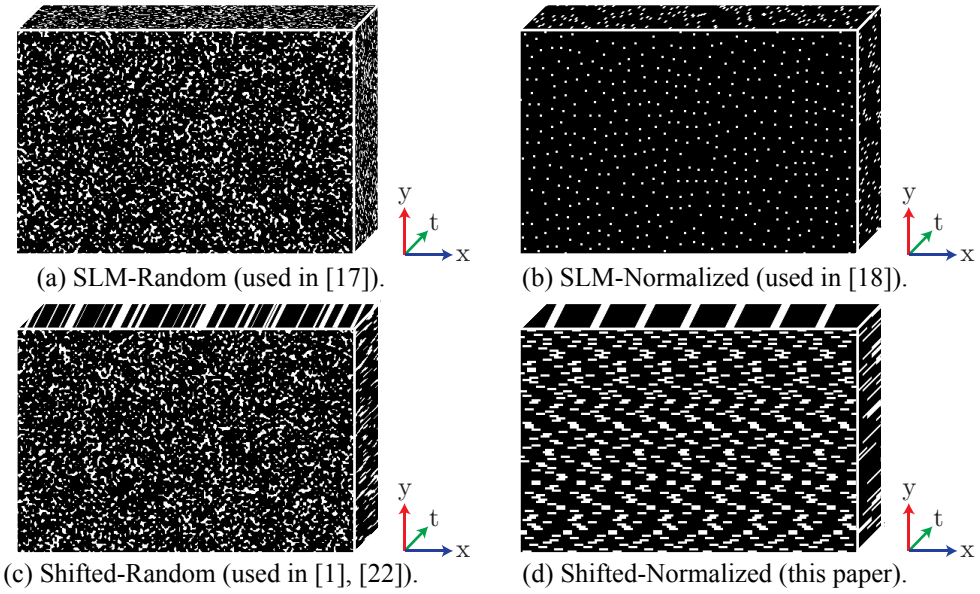
$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{y} - \Phi\Psi^{-1}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1. \tag{5}$$

There is a variety of algorithms for solving the problem in Eq. (5). In our experiments we used the FISTA (Fast Iterative Shrinkage Thresholding Algorithm) of [32] and the interior point method (L1_LS) of [33] employing the DCT as a sparsifying basis. Additionally, we investigate the patch-wise implementation of ℓ_1 optimization schemes by employing the GAP (Generalized Alternating Projection) algorithm. GAP was introduced by Liao et al. [34] and employed for this particular problem by Llull et al. [1]. It incorporates the idea of an ℓ_1 minimization by a weighted ℓ_2 problem. We use the implementation from Llull et al. [1], provided courtesy of the authors.

ℓ_2 **regularization**: Using $F = \|\Theta\mathbf{x}\|_2^2$, Eq. (3) can be solved directly in the signal domain as,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda \|\Theta\mathbf{x}\|_2^2. \tag{6}$$

Using Θ as a high-pass filter operator (HP) we implicitly enforce the gradient of the signal, or else its energy at high frequencies, to be small. This formulation is referred to as CLS (Constrained Least Squares). Strictly speaking, the CLS formulation cannot be categorized as a sparse optimization method since it enforces a Gaussian regularizer instead of a Laplacian one. It is included here for comparison because the optimization problem can be computed extremely efficiently, which is of critical importance since our system captures images with much higher resolution than previously demonstrated.



(a) SLM-Random (used in [17]). (b) SLM-Normalized (used in [18]).

(c) Shifted-Random (used in [1], [22]). (d) Shifted-Normalized (this paper).

Fig. 2: **The four different mask patterns considered in this paper**. (a) and (b) show a set of T different masks to be displayed on an SLM; (a) shows a set of thresholded Gaussian masks; (b) shows a set of masks that, when summed across the t direction, result to the same number of samples per pixel. The forefront masks presented in (c) and (d) depict a mask pattern which will be placed on a translating stage in order to produce the datacube shown, when translated horizontally on the x direction; (c) corresponds to a thresholded Gaussian mask; (d) corresponds to the optimized mask proposed in this paper. The proposed mask, when translated, produces an average value that is identical across all pixels of the sensor.

4. Mask design

The choice of the matrix Φ plays a key role in the reconstructed image quality irrespectively of the selected reconstruction algorithm. Various popular matrices in the literature are known to perform particularly well (*e.g.*, Gaussian, Bernoulli) for signals that can be represented sparsely in some basis. In addition, several researchers have analyzed the problem of optimizing the Φ matrix [35, 36, 15]. Unfortunately, optimization approaches usually rely on minimizing coherence between the sampling matrix and the sparsifying basis, which mostly applies to spatial compressive sensing where dense matrices are used. Instead, the masks used for temporal compressive sensing systems, as the one described herein, result in a sparse matrix with entries across diagonals, as presented by Eq. (2), with only $T \cdot (H \cdot W)$ non-zero values. For SLM-based systems, each of these $T \cdot (H \cdot W)$ entries can be chosen independently by programming the projected patterns. For mask-translation systems, however, only $H \cdot W + T \cdot H$ values can be chosen independently (assuming horizontal translation at T discrete locations).

Two questions then naturally arise 1) "How does performance of SLM and translating mask-based compressive video systems compare?", and 2) "Which mask patterns will produce the optimal sensing matrix Φ?". With regards to 1), we intuitively expect SLM-based approaches to provide superior performance since they enable greater flexibility on pattern selection, giving additional degrees of freedom in sensing matrix design. This is confirmed by our experiments, to be described later. With regards to 2), Llull et al. [1] and Reddy et al. [17] use a thresholded Gaussian mask, presumably because it results in a sensing matrix that most closely resembles

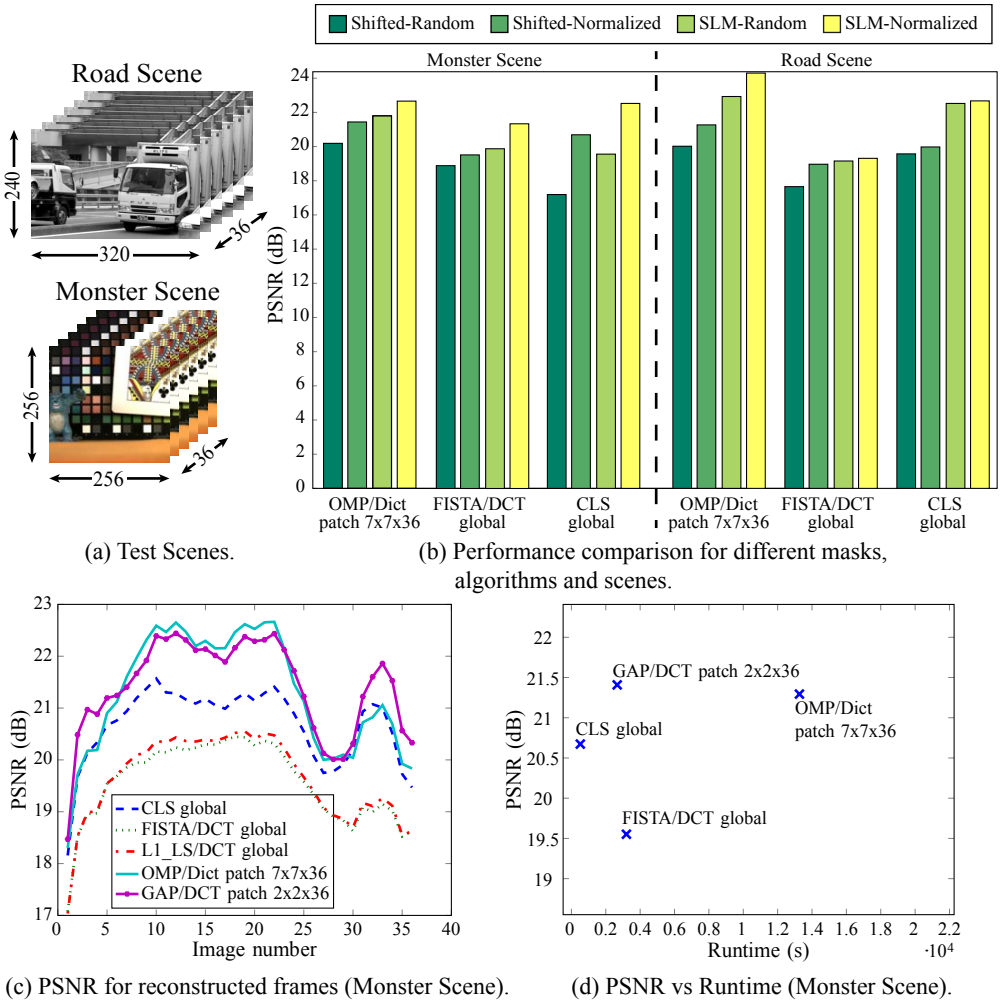
a dense Gaussian matrix. In contrast, the mask patterns of Liu et al. [18] are normalized so that the total amount of light collected at each pixel is constrained to be constant (a condition not met by the random mask patterns). These normalized patterns have the advantage that stationary regions in the scene are photographed exactly without the need for any decoding. Our analysis shows that such masks tend to produce improved performance across a wide variety of reconstruction algorithms.

In the following, we analyze how the choice of mask pattern affects the performance of compressive reconstruction, for both SLM and translating mask-based systems. The four mask patterns we compared are depicted in Fig. 2. The SLM-Random mask patterns from Fig. 2(a) are similar to the ones used by Reddy et al. [17]. They consist of a set of T $H \times W$ pixel masks produced from thresholded random Gaussian matrices. The SLM-Normalized mask patterns from Fig. 2(b) are similar to the ones used by Liu et al. [18]. These patterns are pseudo-random, but when summed, over time, produce an average intensity that is identical for each pixel. The Shifted-Random mask pattern from Fig. 2(c) is similar to the ones used by Llull et al. [1] and Gao et al. [22]. It consists of a single $H \times W$ pixel mask produced by thresholding a random Gaussian matrix. The mask is placed on a translating stage to produce a sensing matrix Φ as described by Eq. (2). Figure 2(d) shows the translating mask used in our prototype. Our system is a hybridization of the implementations from Liu et al. [18] and Llull et al. [1]. We use a shifted mask-based system that avoids the need for a costly SLM, while producing normalized mask patterns that result in superior reconstruction quality.

As observed in Fig. 2(d), our proposed mask consists of randomized lines of a repetitive pattern allowing light to go through for P out of T frames for each pixel, hence ensuring that all pixels are sampled the same amount of times, as in [18]. The apparent white streaks are a consequence of the fact that we are using a translated mask instead of a SLM where random selection of patterns per frame would be possible. The ratio P/T controls the level of spatio-temporal mixing and can significantly affect reconstruction performance. For $P = 1$ the reconstruction problem reduces to a per-frame inpainting problem, while for $P = T$ the captured scene would be corrupted by motion blur not allowing recovery of multiple frames.

4.1. Simulated performance

We simulated noiseless compressive video capture for two different scenes (Monster Scene from [9], and Road Scene from [31], see Fig. 3(a)). We tested three types of algorithms (ℓ_0, ℓ_1 and ℓ_2 regularization) whose details were presented in section 3. For each algorithm, we determined the optimal regularization parameters using cross-validation. We tested the four mask types shown in Fig. 2, Random and Normalized masks for both SLM-based and translating mask-based systems. For each mask, there is a parameter that can be tuned to optimize performance: The binarizing threshold for Random masks, and the average transmitted intensity (value of P) for Normalized masks. After conducting an extensive search along the parameter space, we report on the performance of the masks with the optimum parameter selection for T being fixed at 36 frames. Figure 3(b) shows reconstruction quality for each mask and algorithm combination, using the PSNR (Peak Signal to Noise Ratio) metric. Each group of 4 bars describes the relative performance gain due to a certain mask selection under a given reconstruction algorithm, while the different groups of bars compare the performance gain resulting by each reconstruction approach. First, we note that the performance loss of a translating mask over an SLM-based implementation can be quantified at around 1-3 dB, in terms of PSNR. Second, we observe that the Normalized masks perform consistently better compared to their Random counterparts, across all algorithms. We also note that, various levels of performance gain can be achieved by using Shifted-Normalized masks, compared to the Shifted-Random ones.



(a) Test Scenes.

(b) Performance comparison for different masks, algorithms and scenes.

(c) PSNR for reconstructed frames (Monster Scene).

(d) PSNR vs Runtime (Monster Scene).

Fig. 3: **Simulated reconstruction results**: (a) The Monster and Road test scenes used for simulation; (b) Average reconstruction PSNRs for three different algorithms discussed in section 3 and the four masks discussed in section 4. The SLM-based approach performs 1-3 dB better than the translating mask approach and the Normalized masks provide an increase of 1-3 dB in reconstruction quality, compared to their Random counterparts; (c) PSNR for each of the 36 reconstructed video frames of the Monster scene using our proposed Shifted-Normalized mask. Reconstruction quality varies due to varying motion between subsequent frames in the video sequence. OMP and GAP perform best, with CLS performing slightly worse. Note that the selection of a $2 \times 2 \times 36$ patch size for the GAP algorithm is because the code necessitated that the time dimension of the patch would be a multiple of the spatial patch size. The code was further tested with a $7 \times 7 \times 35$ patch and resulted in performance comparable to the one of the CLS/HP algorithm (not reported here for consistency in the presented number of total reconstructed frames). FISTA and L1_LS lead to the worst reconstruction quality; (d) PSNR vs Runtime comparison using 4 algorithms for the reconstruction of the Monster scene. CLS provides the best balance between reconstruction quality and speed.

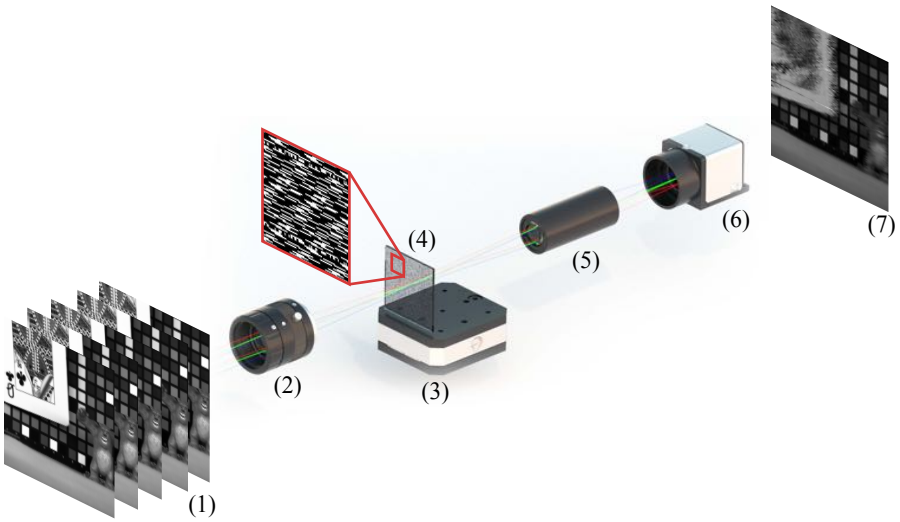


Fig. 4: **System design**: (1) Moving scene; (2) Objective Lens; (3) Piezoelectric stage; (4) Coded optical mask; (5) Relay Lens; (6) Sensor; (7) Coded image. See section 5.1 for details.

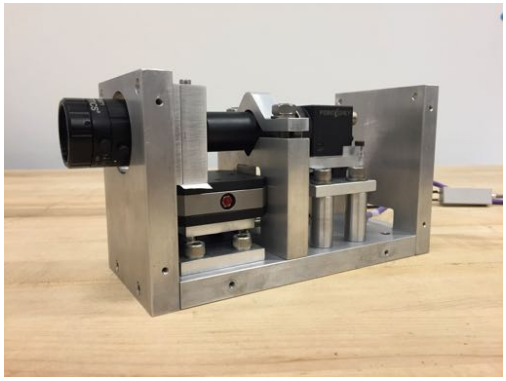


Fig. 5: **Hardware prototype**: Our completed camera system with its dust cover removed.

We further test the performance of our hybrid Normalized-Shifted sampling pattern under different algorithms. We compare the obtained results in terms of PSNR and Runtime in Figs 3(c) and 3(d). Figure 3(c) shows reconstruction quality for each one of the 36 reconstructed video frames of the Monster scene (see Fig. 3 caption for details). Our results show that, while GAP and OMP provide the best reconstruction quality, CLS offers the best tradeoff between reconstruction quality and processing time (see Fig. 3(d)). The reconstruction times for GAP and OMP are $5\times$ and $25\times$ slower, respectively, compared to CLS for the 256×256 pixel Monster scene. Furthermore, reconstruction times scale fast with image size. Currently, the OMP implementation using the $10,000$ atom dictionary of [31] takes approximately two weeks on one CPU for a 2 Mpixel image. On the contrary, the CLS algorithm provides outstanding speed benefits with little sacrifice in reconstruction quality. We believe that the reconstruction speed benefit when using the CLS approach likely outweighs the small decrease in reconstruction quality for many potential applications. Furthermore, our analysis informs decisions on quality versus reconstruction speed tradeoffs that may be made at runtime to best suit a specific application's needs.

5. Hardware prototype

5.1. System design

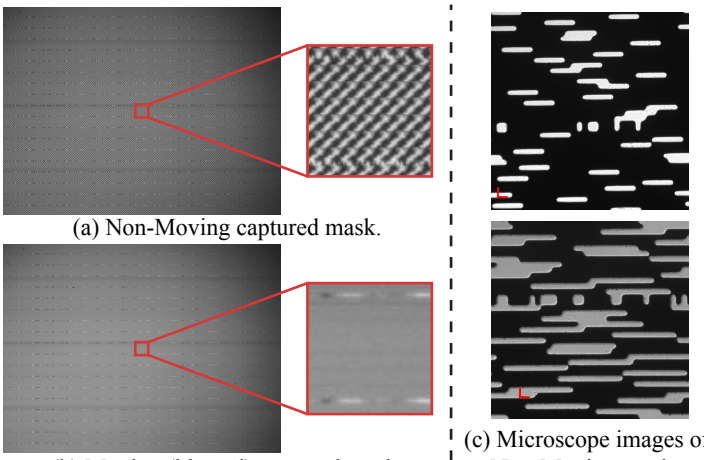
Our hardware prototype consists of five functional components ((2)-(6)) as shown in Fig. 4. A moving scene, represented as a 3D space-time volume, is imaged by an Edmunds Optics 25 mm C-mount objective lens (EO-59-871). We place an optical mask at the back focal plane of the objective lens. The silicon-dioxide optical mask is laser etched (JD Photo-Tools, UK) with the coded patterns described in section 4. The mask is translated laterally across the image plane by a PX 400 stage from Piezosystemjena. An Edmunds Optics 30 mm $f/4$ relay lens (EO-45-759) images the mask and first conjugate plane onto the sensor, a Point Grey Blackfly 20E4M-CS. These components are all mounted in a custom-machined aluminum chassis. Manual micrometer stages allow for fine focal adjustments of the main mask and sensor relative to the two lenses. The complete system is shown in Fig. 5 with its dust cover removed.

5.2. System calibration

One of the disadvantages of using a translating mask system over an SLM-based system is that a mechanism for mask motion must be introduced, and the motion must be precisely controlled to avoid systematic errors that negatively impact image quality. Our system features much higher resolution than other implementations (2 Mpixels instead of .26 Mpixels [1] and .08 Mpixels [31]). However, the small pixel size of our sensor ($4.5 \times 4.5 \mu m^2$) requires much higher mechanical and optical precision, complicating system calibration. A key feature of our system that enables high spatial resolutions is the incorporation of inexpensive CMOS sensors (as opposed to CCD, used in [1]). We initially used a rolling shutter based CMOS sensor but found this system difficult to calibrate due to problems predicting rolling-shutter readout timing. For this reason, we chose to use the Point Grey Blackfly 20E4M-CS CMOS, which is one of the highest resolution available CMOS cameras providing a global shutter exposure mode.

Stage movement is another factor leading to system performance deterioration. We operate the piezoelectric stage at a relatively low frequency of 5 Hz to ensure constant velocity over the duration of exposure. In the future we are planning on using faster translation speeds, which would allow for continuous mode operation, enabling capture of longer video sequences. Our custom built stage ensures precise alignment of sensor, mask, and relay optics. Nonetheless, small mask rotations and non-unity magnification make it difficult to predict the exact mapping from mask to sensor during camera operation. Furthermore, due to the very small pixel size, the mask fabrication process can lead to non-sharp edges, as depicted in the microscope captured images of two different coded masks in Fig. 6(c). In an attempt to calibrate for mask rotation, we placed a set of fiducial patterns every 100 rows of the image, which can be seen in the static mask capture of Fig. 6(a) as well as in the real captured images discussed in section 6.

Our calibration procedure consists of capturing a sequence of images, each one of which corresponds to the sensor integrating the incoming light through the mask while the mask is translating between two consecutive stage positions. Following such procedure, we effectively measure a set of mask patterns $\mathbb{M}_i$, which are then used to populate the sensing matrix Φ using Eq. (2). This method is performed in hope of eliminating the need for exact alignment between mask and sensor pixels. However, the resulting Φ matrix is no longer identical to the designed mask pattern, introducing small differences in performance between simulations and experiments. Specifically, Fig. 6(a) shows a picture of an example real non-moving mask together with its moving captured counterpart in Fig. 6(b). As expected, the moving mask exhibits non-binary values due to integration during mask translation. Such blurred measured masks were utilized for filling in the diagonals of matrix Φ, making therefore the degradation and recovery processes consistent.



(a) Non-Moving captured mask.

(b) Moving (blurred) captured mask.

(c) Microscope images of Non-Moving masks.

Fig. 6: **Coded mask detail**: (a) Real capture of a static mask (with $P = 4$ and $T = 10$) where fiducial calibration lines are visible, as described in section 5.2; (b) Real capture while the mask moves by 10 pixels or $45\mu m$ horizontally (full duration of acquisition); (c) Microscope images of two different coded masks (upper part for $P = 6$ and $T = 36$; lower part for $P = 12$ and $T = 36$) showing imperfections of the fabrication process.

6. Results

In order to determine the best mask for use in our hardware prototype we perform an experiment comparing reconstruction performance for different values of P and T. For each selected number of frames to be reconstructed (T) we select the optimal number of samples per pixel (P) based on reconstruction performance. Some indicative results are presented in Table 1. The best results were obtained using a Shifted-Normalized mask for 10 frame recovery sampling each pixel 4 times while being translated at 10 discrete locations during acquisition.

Table 1: Reconstruction performance for the Monster scene using ℓ_1 minimization (FISTA/DCT) and the proposed mask (see Fig. 2(d)) for different values of P and T.

T	4	10	16	26	36	50	66
P	1	4	5	9	12	17	22
Mean PSNR (dB)	14.43	**22.23**	21.75	20.63	19.59	19.06	18.51

We present high resolution reconstructions that recover 10 frames of video from a single captured, coded image employing the Shifted-Normalized mask described above and presented in Fig. 6(a). The exposure time was set to be 13.456 ms, producing a frame-rate of $\frac{10}{13.456ms} = 743$ fps. Our first example is the Metronome scene, shown in Fig. 7. This scene consists of a single "Jack" playing card placed on a metronome beating at 88 beats per minute. The metronome speed is carefully chosen leading to small amounts of linear translating motion. The top portion of Fig. 7 shows the captured coded image. The middle portion presents 3 of the 10 reconstructed frames using the CLS algorithm with insets below showing closeups of the motion of the "Jack" card. High-frequency details in the "Jack" are visible in the reconstructions, and motion is clearly visible between frames. Note, in particular, the location of the letter "J" and the heart relative to the grid lines. The total motion of the card is approximately 15 pixels over the entire 13.456 ms exposure. Please see the complete video in Media1.

Figure 8 shows another scene consisting of several balls falling to the ground. This example depicts high-speed linear translating motion. Again, the top subfigure shows the captured coded image, while the bottom subfigures show 3 of the 10 reconstructed frames with closeups, using the CLS algorithm. Motion of the soccer ball in the closeups is clearly recovered, but the motion is significantly greater than one pixel per frame on the sensor. Specifically, the total motion of the soccer ball is approximately 60 pixels over the entire 13.456 ms exposure, resulting in a motion blur of several pixels for each of the reconstructed frames. This motion blur could theoretically be eliminated using a system with higher compression ratio (*e.g.*, 100× instead of 10×); however reconstruction quality would invariably suffer as a result. The reconstructed video also exhibits some ghosting artifacts, which we believe could be eliminated in future algorithmic implementations. Please see the complete video in Media2. For comparison purposes, we also provide the reconstruction using the OMP/Dictionary approach in Media3.

The final scene consists of a deck of cards hit by a ball (Fig. 9) such that each card is thrown towards a random direction. The scene depicts large amounts of arbitrary motion. Reconstructing motion in this scene is a particularly challenging problem for our compressive video camera. Closeups in the bottom two rows show rotating motion faithfully recovered, but artifacts are clearly visible. Please see the complete video in Media4.

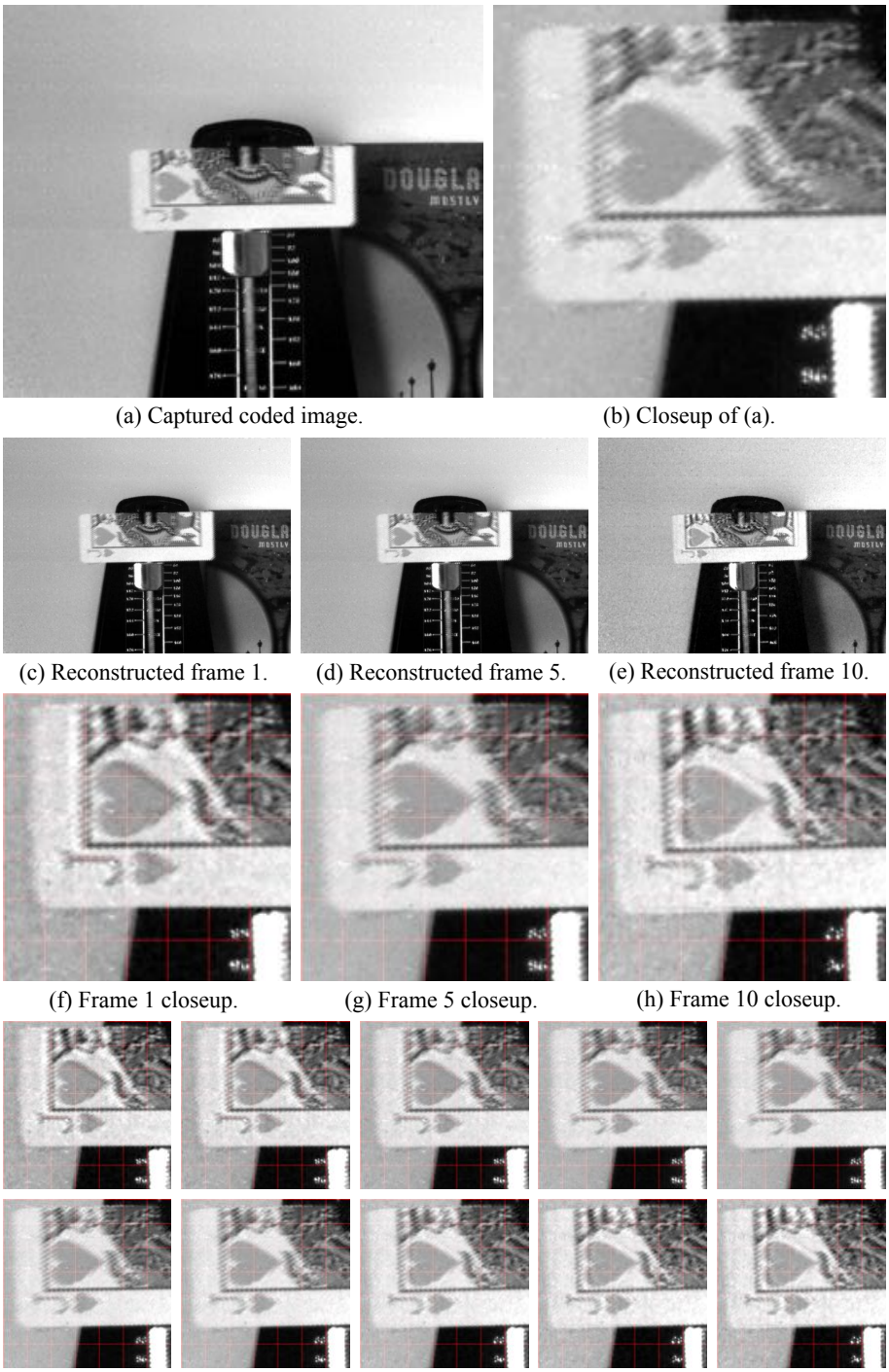
7. Discussion

We have presented a prototype compressive video camera capable of 2 Mpixel, 743 fps reconstructions. To the best of our knowledge, this is the largest frame size to date for any compressive video camera. However, analyzing spatial resolution for compressive video cameras is a very challenging problem since performance depends equally on the mask pattern, reconstruction algorithm, calibration of the experimental setup as well as the amount of motion in the scene. In future work, we plan to perform a thorough analysis to establish the maximum spatial resolution that our camera can achieve. In addition, we introduced a new translating mask pattern that leads to improved performance compared to the translating masks previously used by Llull et al. [1]. We demonstrated several examples of high-resolution videos reconstructed from data captured with our system, testing performance for a variety of motion types.

The reconstruction algorithms evaluated in this paper only consider a single captured coded image. There is potentially performance gain in formulating the reconstruction algorithm to enforce temporal smoothness over multiple captured frames. This is not possible with the current global shutter sensor since the readout time between consecutive captured frames spans the duration of multiple subsequent frames in the captured scene that would lead to large gaps in the reconstructed video sequence. This may actually be a reason to return to a rolling shutter readout mode, though it would pose challenges in mask calibration.

Future success of our compressive video sensing architecture relies on elimination of systematic errors that negatively affect reconstructed image quality. The most practical hardware setup is one that would not require precise alignment and synchronization between the mask and sensor. An ideal system would apply spatio-temporal modulation electronically using sensor readout circuitry. Such as sensor would provide the same compressive video capture without the need for any extra optics (*e.g.*, coded mask, stage, or relay lens), enabling compressive video capture to be implemented in an extremely compact package. To the best of our knowledge, no such sensor is currently commercially available, however, it is feasible that one could be built. We hope to develop such a custom sensor in the near future.

Finally, increasing frame-rate necessarily requires fewer photons to be collected by each frame, whether using native high frame-rate cameras or compressive video capture. As future work enables faster reconstruction frame-rates, the limits imposed by noise on reconstruction quality will need to be explored more thoroughly.



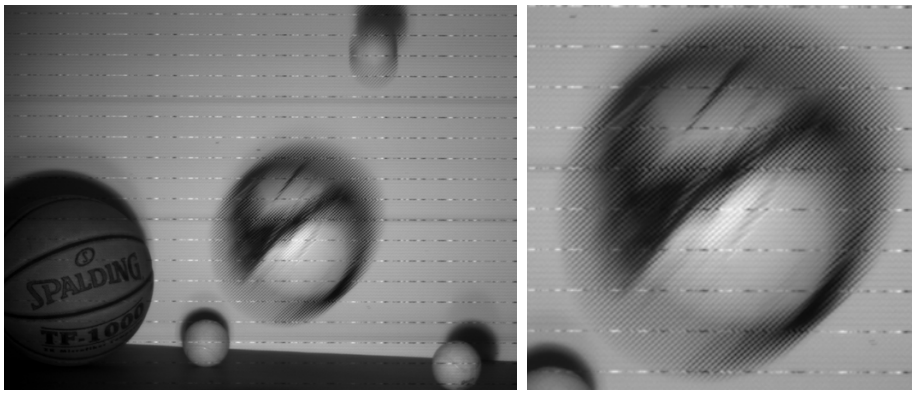
(a) Captured coded image. (b) Closeup of (a).

(c) Reconstructed frame 1. (d) Reconstructed frame 5. (e) Reconstructed frame 10.

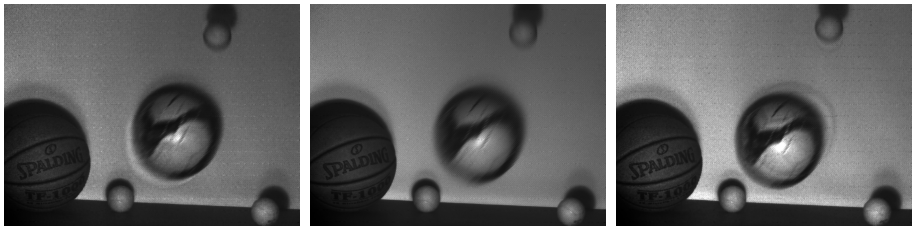
(f) Frame 1 closeup. (g) Frame 5 closeup. (h) Frame 10 closeup.

(i) Closeups of all 10 reconstructed frames.

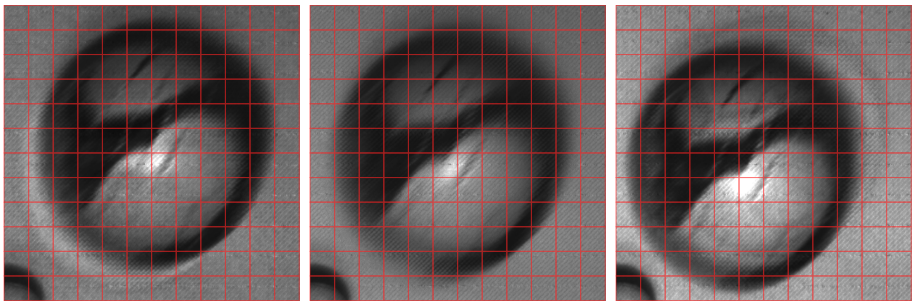
Fig. 7: Reconstruction of the Metronome scene using the Constrained Least-Squares method with a high-pass filter. The scene consists of small amounts of translating motion. Parts (a)-(b) show the captured image; Parts (c)-(h) present 3 of the 10 reconstructed frames; Part (i) depicts closeups on the translating "Jack". Please see the complete video in Media1.



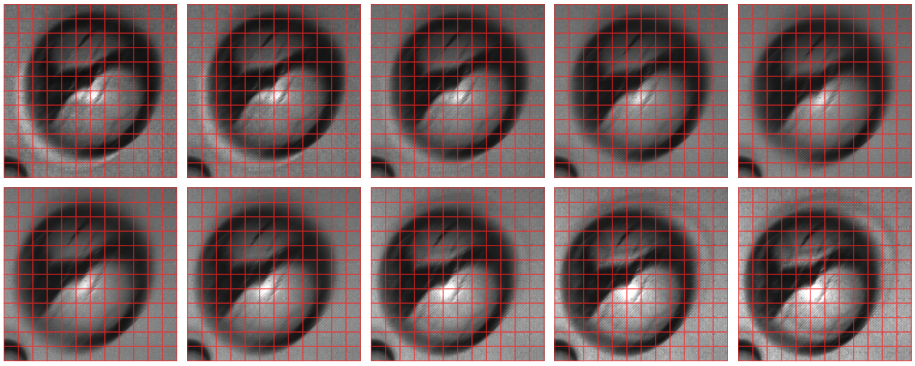
(a) Captured coded image. (b) Closeup of (a).



(c) Reconstructed frame 2. (d) Reconstructed frame 5. (e) Reconstructed frame 9.



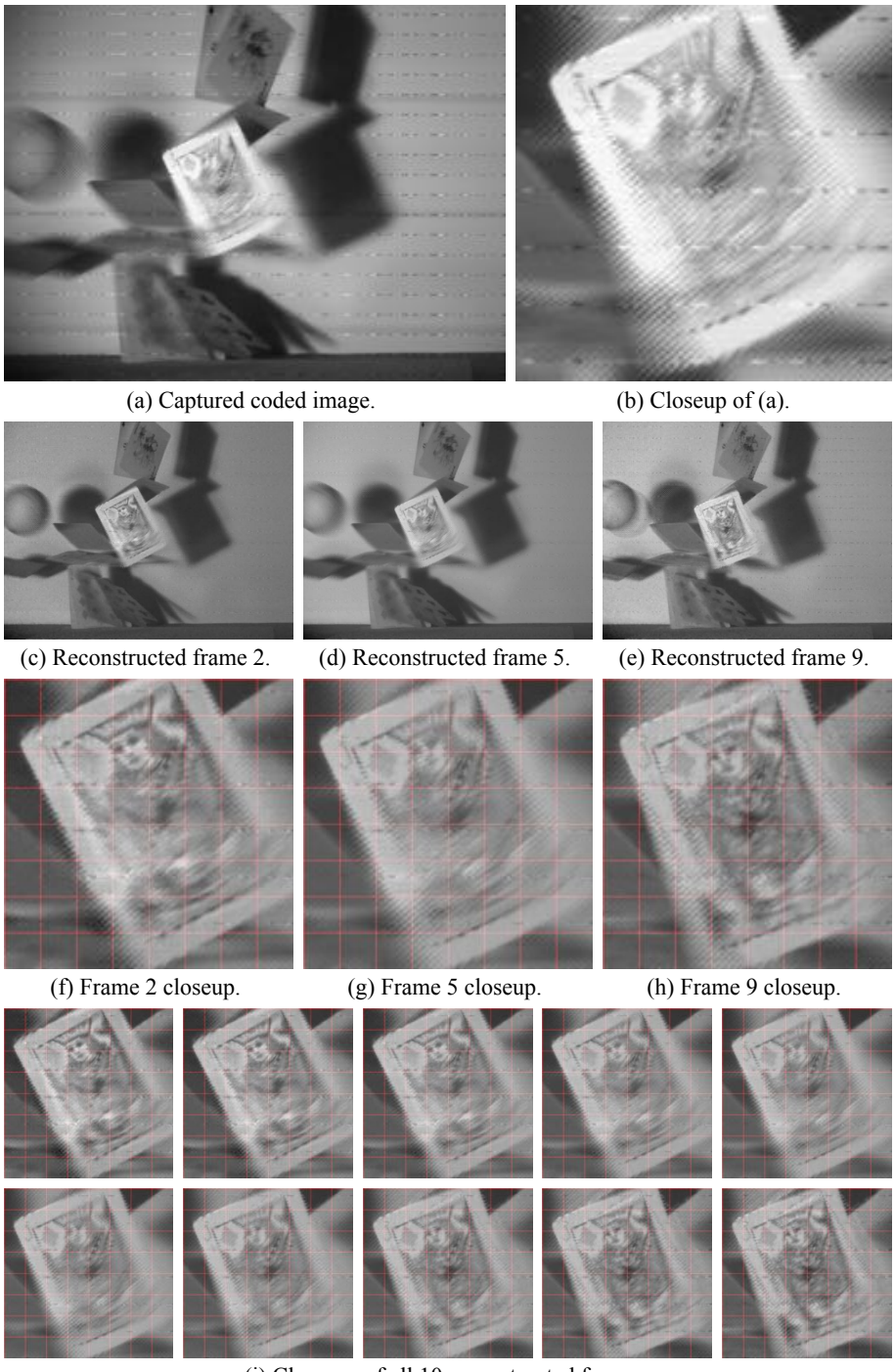
(f) Frame 2 closeup. (g) Frame 5 closeup. (h) Frame 9 closeup.



(i) Closeups of all 10 reconstructed frames.

Fig. 8: Reconstruction of the Ball scene using the Constrained Least-Squares method with a high-pass filter. The scene consists of large amounts of translating motion. Parts (a)-(b) show the captured image; Parts (c)-(h) present 3 of the 10 reconstructed frames; Part (i) depicts close-ups of the falling soccer ball. Please see the complete video in Media2 and the corresponding Orthogonal Matching Pursuit/Dictionary reconstruction in Media3

.



(a) Captured coded image. (b) Closeup of (a).

(c) Reconstructed frame 2. (d) Reconstructed frame 5. (e) Reconstructed frame 9.

(f) Frame 2 closeup. (g) Frame 5 closeup. (h) Frame 9 closeup.

(i) Closeups of all 10 reconstructed frames.

Fig. 9: Reconstruction of the Deck of Cards scene using the Constrained Least-Squares method with a high-pass filter. The scene consists of large amounts of arbitrary motion. Parts (a)-(b) show the captured image; Parts (c)-(h) present 3 of the 10 reconstructed frames; Part (i) depicts closeups of the rotating "Jack". Please see the complete video in Media4.