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

P2C2: Programmable Pixel Compressive Camera for High Speed Imaging

Reddy, D.; Veeraraghavan, A.; Chellappa, R.

TR2011-039 June 2011

Abstract

We describe an imaging architecture for compressive video sensing termed programmable pixel compressive camera (P2C2). P2C2 allows us to capture fast phenomena at frame rates higher than the camera sensor. In P2C2, each pixel has an independent shutter that is modulated at a rate higher than the camera frame-rate. The observed intensity at a pixel is an integration of the incoming light modulated by its specific shutter. We propose a reconstruction algorithm that uses the data from P2C2 along with additional priors about videos to perform temporal super-resolution. We model the spatial redundancy of videos using sparse representations and the temporal redundancy using brightness constancy constraints inferred via optical flow. We show that by modeling such spatio-temporal redundancies in a video volume, one can faithfully recover the underlying high-speed video frames from the observed low speed coded video. The imaging architecture and the reconstruction algorithm allows us to achieve temporal super-resolution without loss in spatial resolution. We implement a prototype of P2C2 using an LCOS modulator and recover several videos at 200 fps using a 25 fps camera.

IEEE Computer Vision and Pattern Recognition (CVPR)

P2C2: Programmable Pixel Compressive Camera for High Speed Imaging

Dikpal Reddy
University of Maryland
College Park, MD, 20742
dikpal@umiacs.umd.edu

Ashok Veeraraghavan
Mitsubishi Electric Research Labs
Cambridge, MA, 02139
veerarag@merl.com

Rama Chellappa
University of Maryland
College Park, MD, 20742
rama@umiacs.umd.edu

Abstract

We describe an imaging architecture for compressive video sensing termed programmable pixel compressive camera (P2C2). P2C2 allows us to capture fast phenomena at frame rates higher than the camera sensor. In P2C2, each pixel has an independent shutter that is modulated at a rate higher than the camera frame-rate. The observed intensity at a pixel is an integration of the incoming light modulated by its specific shutter. We propose a reconstruction algorithm that uses the data from P2C2 along with additional priors about videos to perform temporal super-resolution. We model the spatial redundancy of videos using sparse representations and the temporal redundancy using brightness constancy constraints inferred via optical flow. We show that by modeling such spatio-temporal redundancies in a video volume, one can faithfully recover the underlying high-speed video frames from the observed low speed coded video. The imaging architecture and the reconstruction algorithm allows us to achieve temporal super-resolution without loss in spatial resolution. We implement a prototype of P2C2 using an LCOS modulator and recover several videos at 200 fps using a 25 fps camera.

1. Introduction

Spatial resolution of imaging devices is steadily increasing; mobile phone cameras have $5-10$ megapixels while point-and-shoot cameras have $12-18$ megapixels. But the temporal resolution of video cameras has increased slowly; today's video cameras are mostly $30-60$ fps. High-speed video cameras are technically challenging due to high bandwidth and high light efficiency requirements. In this paper, we present an alternative architecture for acquiring high-speed videos that overcomes both these limitations.

The imaging architecture (Figure 1), is termed Programmable Pixel Compressive Camera (P2C2). P2C2 consists of a normal 25 fps, low resolution video camera, with a high resolution, high frame-rate modulating device such as a Liquid Crystal on Silicon (LCOS) or a Digital Micromirror Device (DMD) array. The modulating device modulates each pixel independently in a pre-determined random fashion at a rate higher than the acquisition frame rate of the camera. Thus, each observed frame at the camera is a coded linear combination of the voxels of the underlying high-speed video frames. Both low frame-rate video cameras and high frame-rate amplitude modulators (DMD/LCOS) are inexpensive and this results in significant cost reduction. Further, the capture bandwidth is significantly reduced due to P2C2's compressive imaging architecture. The underlying high resolution, high-speed frames are recovered from the captured low resolution frames by exploiting temporal redundancy in the form of brightness constancy and spatial redundancy through transform domain sparsity in a convex optimization framework.

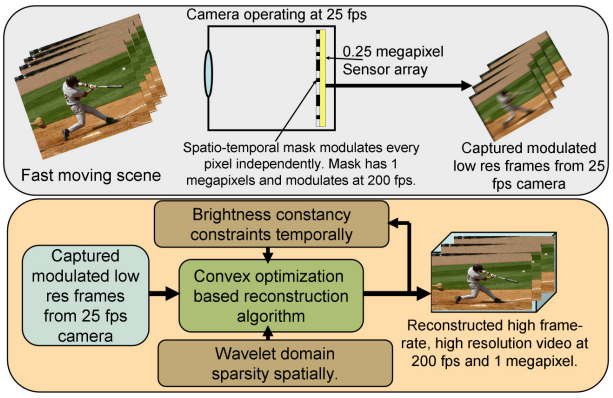

Figure 1. Programmable Pixel Compressive Camera (P2C2): Each pixel of a low frame rate, low resolution camera is modulated independently with a fast, high resolution modulator (LCOS or DMD). The captured modulated low resolution frames are used with accompanying brightness constancy constraints and a wavelet domain sparsity model in a convex optimization framework to recover high resolution, high-speed video.

1.1. Contributions:

- We propose a new imaging architecture 'P2C2' for compressive acquisition of high-speed videos. P2C2 allows temporal super-resolution of videos with no loss in spatial resolution.
- We show that brightness constancy constraints significantly improve video reconstruction. Our algorithm reconstructs high-speed videos from low frame rate

observations over a broad range of scene motions.

- We characterize the benefits and limitations of P2C2 through experiments on high-speed videos. We implement a prototype P2C2 and acquire 200 fps videos of challenging scenes using a 25 fps video camera.

2. Related Work

High speed sensors: Traditional high-speed cameras are expensive due to requirement of high light sensitivity and large bandwidth. Usually these cameras [1] have limited on-board memory with a dedicated bus connecting the sensor. The acquisition time is limited by the on-board memory. For example, FastCam SA5 (a $\$300K$ high-speed camera) can capture atmost 3 seconds of video at 7500 fps and 1 megapixel. Though most videos have significant spatio-temporal redundancy, current high-speed cameras do not exploit them. P2C2 allows us to exploit this, thereby reducing the capture bandwidth significantly. Further, existing cameras use specialized sensors with high light sensitivity and image intensifiers to ensure each frame is above the noise bed. In contrast, P2C2 captures a linear combination of video voxels, thereby naturally increasing the acquisition signal-to-noise ratio and partially mitigating the need for image intensifiers.

Temporal super-resolution: Shechtman et al. [23] perform spatio-temporal super-resolution by using multiple cameras with staggered exposures. Similarly, Wilburn et al. [26] use a dense 30 fps camera array to generate a 1000 fps video. Recently Agrawal et al. [2] showed that combining this idea with per camera flutter shutter (FS) [17] significantly improves the performance of such staggered multi-camera high-speed acquisition systems. While these systems acquire high-speed videos, they require multiple cameras with accurate synchronization and their frame-rate scales only linearly with number of cameras. In contrast, we increase temporal resolution without the need for multiple cameras and also our camera is not restricted to planar scene motion. Ben-Ezra [3] built a hybrid camera where motion is measured using an additional higher frame rate sensor and then used to estimate the point spread function for deblurring. We estimate both motion and appearance from the same sensor measurements.

Video interpolation: Several techniques exist for frame-rate conversion [22]. Recently, [12] showed that explicit modeling of occlusions and optical flow in the interpolation process allows us to extract 'plausible' interpretations of intermediate frames.

Motion deblurring: When a fast phenomenon is acquired via a low frame-rate camera one can either obtain noisy and aliased sharp images using short exposure, or blurred images using long exposures. Motion deblurring has made great progress by incorporating spatial regularization terms within the deconvolution framework [21][6]. Novel hardware architectures [17][10] have also been designed to improve deconvolution. These techniques require the knowledge of motion magnitude/direction and cannot handle general scenes exhibiting complex motion. In contrast, P2C2 can handle complex motion without the need for any prior knowledge.

Compressive sensing (CS) of videos: Existing methods for video CS assume multiple random linear measurements are available at each time instant either using a coded aperture [14] or a single pixel camera (SPC) [5]. [24] shows that videos with slowly changing dynamics need far fewer measurements for subsequent frames once the first frame is recovered using standard number of measurements. [16] presents an algorithm for compressive video reconstruction by using a motion compensated wavelet basis to sparsely represent the spatio-temporal volume. Such methods have achieved only moderate success since (a) the temporal redundancy of videos is not explicitly modeled and (b) the hardware architectures need to be highly engineered and/or are expensive.

In [25], the authors extend FS to videos and build a high-speed camera for periodic scenes. For the class of video that can be adequately modeled as linear dynamical system [20] provides a method for compressively acquiring videos using the SPC architecture. Both approaches can handle only periodic/dynamic texture scenes while P2C2 can capture arbitrary videos.

Spatio-temporal trade-off: Gupta et al. [8] show how per-pixel temporal modulation allows flexible post-capture spatio-temporal resolution trade-off. The method loses spatial resolution for moving elements of the scene, whereas our method preserves spatial resolution while achieving higher temporal resolution. Similarly, Bub et al. [4] propose spatio-temporal trade-off of captured videos but has limited light throughput and unlike [8] lacks flexible resolution trade-off. Gu et al. [7] proposed a coded rolling shutter architecture for spatio-temporal trade-off.

Per-pixel control: Nayar et al. [15] propose a DMD array based programmable imager for HDR imaging, feature detection and object recognition. [8, 4] use DMD array based per-pixel control for spatio-temporal resolution trade-off. Similarly, DMD arrays were used in [19, 18] for phase analysis and shape measurement. While the idea of per-pixel modulation is not new, we propose a sophisticated spatio-temporal modulation using P2C2 for high-speed imaging. Such modulation allows us to achieve higher temporal resolution without loss in spatial resolution.

3. Imaging Architecture

Let the intensity of desired high frame rate video be $x(s,t)$ where $s = (r,c) \in [1\ N] \times [1\ N]$ are the row and column coordinates respectively and $t \in [1\ T]$ the temporal coordinates. We term the higher rate frames x_t as 'sub-frames' since the acquired frames are formed by in-

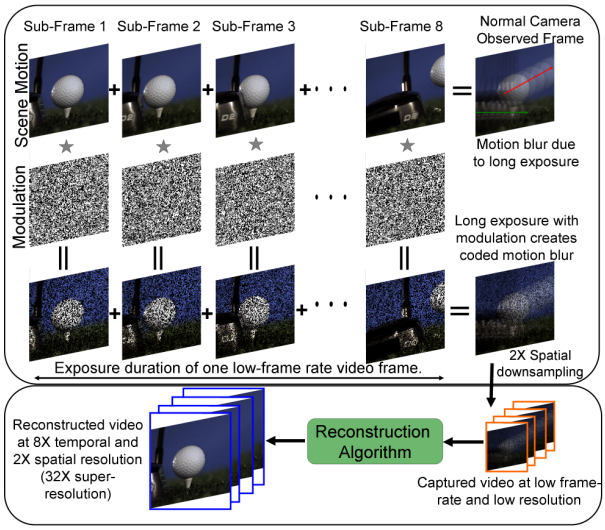

Figure 2. Camera architecture: At every pixel the camera independently modulates the incoming light at sub-frame durations and then integrates it. For example, the 3-D spatio temporal volume of a golf ball is modulated with a random mask at sub-frame durations and then integrated into a frame. A frame captured by our camera has the code embedded in the blur.

tegrating them. P2C2 captures the modulated intensities $y(s_l, t_l)$ where $s_l = (r_l, c_l) \in [1\ N/L_s] \times [1\ N/L_s]$ and $t_l \in [1\ T/L_t]$ are its spatial and temporal coordinates. L_s and L_t are the spatial and temporal sub-sampling factors respectively. The captured frame y_{t_l} is related to sub-frames x_t as

$$y_{t_l} = D\left(\sum_{t=(t_l-1)L_t+1}^{t_l L_t} x_t \phi_t\right) \tag{1}$$

where ϕ is the spatio-temporal modulation function (achieved by LCOS as shown in Figure 3) and $x_t\phi_t$ is modulation of sub-frame at t with mask at t. $D(\cdot)$ denotes a spatial subsampling operation to account for the possibility that camera could also be spatially modulated at sub-pixel resolution. Notice that L_t sub-frames are modulated with L_t independent high resolution random masks and then integrated to produce one spatio-temporally subsampled frame of captured video (as shown in Figure 2). We limit our discussion mostly to temporal downsampling. Nevertheless, the architecture and recovery algorithm presented later easily extend to spatial subsampling as well and we illustrate it through results in experimental section.

Since the observed pixel intensities y are linear combinations of the desired voxels x, with the weights given by modulation function ϕ, the equation (1) can be written in matrix-vector form as,

$$\mathbf{y} = \mathbf{\Phi x} \tag{2}$$

where $\mathbf{\Phi}$ is the matrix representing per pixel modulation followed by integration in time and spatial sub-sampling. $\mathbf{x}$ and $\mathbf{y}$ are the vectorized form of desired high-speed voxels x (eg., $256 \times 256 \times 32$ voxels) and the captured video y ($128 \times 128 \times 4$ video) respectively. The optimization term enforcing fidelity of the recovered sub-frames to the captured frames is given by $E_{data} = \|\mathbf{y} - \mathbf{\Phi x}\|_2^2$.

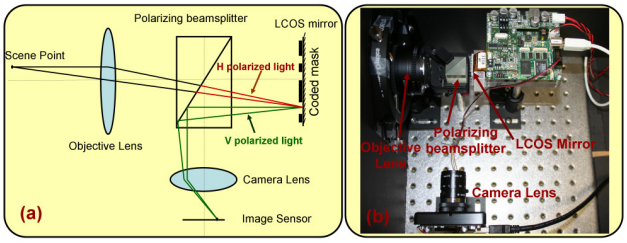

Figure 3. Prototype: Illustration of the optical setup.

3.1. Prototype P2C2

We realize P2C2 with an LCOS mirror SXGA-3DM from Forth Dimension Displays as a spatio-temporal modulator. The mirror has 1280×1024 pixels and each pixel can be fluttered (binary) independently at maximum rate of 3.2 kHz. This imposes an upper limit of 3200 fps on framerate of the recovered video. LCOS works by flipping the polarization state of incoming light and therefore needs to be used with a polarizing beam-splitter and necessary relay optics as shown in Figure 3. The scene is focused on LCOS device which modulates this incoming light. The Pointgrey Dragonfly2 sensor (1024×768 pixels at 25 fps) is in turn focused on LCOS mirror. An LCOS modulator offers a significantly higher contrast ratio (> 100) compared to off-the-shelf graphic LCD attenuators. Further, the primary advantage of LCOS mirror over LCD arrays is the higher fill factor of pixels. LCOS based per-pixel control was used by Mannami et al. [13] for recovering high dynamic range images.

Related architectures: P2C2 architecture is a generalization of previous imaging architectures proposed for high-speed imaging and motion deblurring. Flutter shutter (FS) camera [17] is a special case where all the pixels have same shutter. P2C2 adopts a random spatio-temporal modulation and is a generalized version of architectures for spatio-temporal resolution trade-off [8, 4, 7].

P2C2 is a compressive imaging system and is related to SPC [5]. In P2C2 the mixing of voxel intensities is localized in space and time as opposed to SPC which aims for global mixing of underlying voxel intensities. Our architecture also exploits the cost benefit of current sensors (especially in visible wavelength) by using a pixel array in place of single pixel detector.

4. High speed video recovery

Since the number of unknown pixel intensities is much larger than available equations, (2) is severely underdetermined. To solve for sub-frames $\mathbf{x}$, a prior on spatio-temporal video volume should be incorporated.

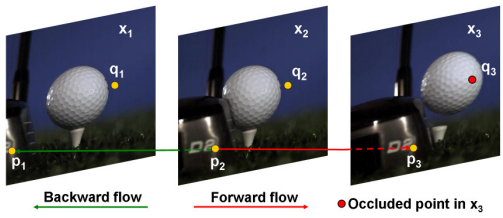

Figure 4. Brightness constancy constraints at p_1, p_2 and p_3 and OF consistency check at q_2 and q_3.

Most natural video sequences are spatio-temporally redundant. Spatially, images are compressible in transform basis such as wavelets and this fact is used in image compression techniques such as JPEG2000. Temporally, object and/or camera motion preserves the appearance of objects in consecutive frames and this fact is used in video compression schemes such as MPEG. We exploit both forms of redundancy to solve the system of under-determined equations (2) and recover the high-speed sub-frames.

4.1. Transform domain sparsity

Each sub-frame is sparse in appropriate transform domain and we enforce this property in our recovery through ℓ_1 regularization of its transform coefficients. The regularization term enforcing spatial-sparsity of sub-frames is $E_{spatial} = \sum_{t=1}^{T} \beta \|\mathbf{\Psi}^{-1}\mathbf{x}_t\|_1$, where $\mathbf{x}_t$ is the vectorized sub-frame x_t and $\mathbf{\Psi}$ the transform basis.

4.2. Brightness constancy as temporal redundancy

Unlike spatial redundancy, temporal redundancy in videos is not easily amenable to sparse representation in a transform basis. Hence, regularization of 3-D transform basis coefficients to solve the under-determined system in (2) results in poor reconstruction quality. To overcome this challenge, we propose to keep temporal regularization term distinct from spatial regularization. We exploit brightness constancy(BC) constraint in temporal direction. This constraint is distinct from and in addition to the spatial transform domain sparsity regularization.

Consider three consecutive frames of a club hitting the ball in Figure 4. The points p_1, p_2 and p_3 correspond to the same point on the golf club in frames x_1, x_2 and x_3 respectively. If relative displacement of these points is estimated, then their pixel intensities in (2) can be constrained to be equal i.e. brightness at these pixels is constant $x(p_2, 2) - x(p_1, 1) = 0$ (backward flow) and $x(p_2, 2) - x(p_3, 3) = 0$ (forward flow). This effectively decreases the number of unknowns by 2. The system becomes significantly less under-determined if BC constraints at other points are known as well. The sub-frame BC constraints over entire video volume are then given by

$$\mathbf{\Omega x} = 0 \tag{3}$$

where every row of matrix $\mathbf{\Omega}$ is the relevant BC equation of a spatio-temporal point (s, t). We incorporate these constraints in the optimization by adding a BC regularization term $E_{BC} = \lambda \|\mathbf{\Omega x}\|_2^2$.

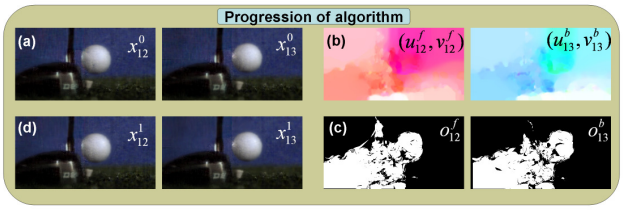

Figure 5. In clockwise direction (a) two sub-frames from the initialization (b) forward and backward OF at respective sub-frames (c) corresponding forward and backward consistency map (d) sub-frames from next iteration incorporate BC constraints only at white pixels from (c).

To create BC constraint at any spatio-temporal point (s, t), we first estimate the optical flow (OF) at sub-frame x_t in forward direction (u_t^f, v_t^f). Then we perform consistency check by estimating the backward flow (u_{t+1}^b, v_{t+1}^b) at sub-frame x_{t+1}. Such a consistency check not only detects points of x_t occluded in sub-frame x_{t+1}, but also prunes the untrustworthy flow in (u_t^f, v_t^f). For example, consider points q_1 and q_2 on blue background in Figure 4. Both points have same spatial coordinates and have same intensity $x(q_2, 2) - x(q_1, 1) = 0$. The fact that both q_1 and q_2 are same points in the scene (here background) is established solely from OF by performing following consistency check: q_1 goes to q_2 according to forward OF and q_2 comes back to q_1 in the backward OF. On the other hand $x(q_2, 2) \neq x(q_3, 3)$ even though $q_2 = q_3$. This is because forward OF suggests q_2 is same as q_3 since q_2 belongs to background and has 0 flow. But the backward OF at q_3 is non-zero and hence $q_3 + (u^b(q_3, 3), v^b(q_3, 3)) \neq q_2$. This implies point q_2 is occluded and/or has unreliable forward OF $(u^f(q_2, 2), v^f(q_2, 2))$. This means the consistency doesn't check at q_2 i.e. $o^f(q_2, 2) = 0$ whereas it checks at q_1 i.e. $o^f(q_1, 1) = 1$. BC constraint is enforced only when consistency checks. We perform consistency check in backward direction as well by checking consistency between (u_t^b, v_t^b) and (u_{t-1}^f, v_{t-1}^f). The process of pruning OF is illustrated in Figure 5. The sub-frames estimated in first iteration of our algorithm (Figure 5a) are used to determine E_{BC} for the next iteration. The results of next iteration are shown in Figure 5d.

The importance of brightness constancy in video recovery is shown in Figure 6. Third column shows reconstruction fidelity obtained by assuming only spatial sparsity. The fourth column shows our reconstruction which incorporates explicit brightness constancy (BC) constraints. This significantly improves reconstruction since the algorithm adapts to the complexity of motion in a particular video.

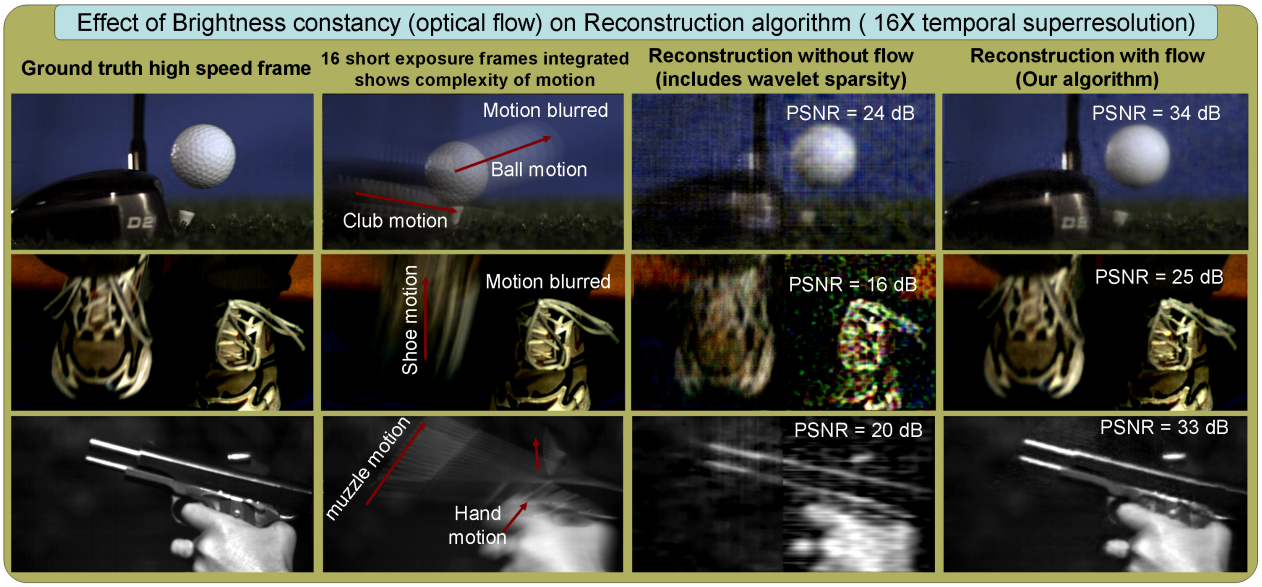

Figure 6. Importance of Brightness Constancy: All results are 16X temporal super-resolution. Shown are the original high-speed frames, motion blurred frames and reconstructions with and without BC. Notice the huge improvement in reconstruction SNR due to BC. The results in column 4 and its necessary OF were computed in an alternating fashion using an iterative procedure on the observations. OF was not assumed to be available. 'Golf' and 'Recoil' high-speed video credit TECH IMAGING.

4.2.1 Recovery Algorithm

Initialization: Given optical flow, BC constraints are incorporated through E_{BC}. But OF can be determined only when the sub-frames are available. Hence, we iteratively determine the sub-frames and the optical flow in an alternating fashion. We begin by estimating the sub-frames without any BC constrains. In the first iteration we trade-off spatial resolution to recover frames at desired temporal resolution. We assume that each sub-frame x_t is an upsampled version of a lower spatial resolution frame: $\mathbf{x}_t = U(\mathbf{z}_t)$ where $\mathbf{z}_t$ is a vectorized $[\frac{N}{L_s\sqrt{L_t}} \times \frac{N}{L_s\sqrt{L_t}}]$ image and $U(\cdot)$ is a linear upsampling operation such as bilinear interpolation. The initial estimate is given by solving

$$\mathbf{z}^0 = \arg\min \sum_{t=1}^{T} \beta \|\mathbf{\Psi}^{-1} U(\mathbf{z}_t)\|_1 + \|\mathbf{y} - \mathbf{\Phi} U(\mathbf{z})\|_2^2. \quad (4)$$

The estimate $\mathbf{x}^0 = U(\mathbf{z}^0)$ doesn't capture all the spatial detail and is noisy but it preserves the motion information accurately as shown in Figure 5a. We estimate OF [11] on initial estimate (Figure 5b) and perform consistency check to prune the flow (Figure 5c) as described in section 4.2. Only the consistent flow is used to build BC constraint matrix $\mathbf{\Omega}^0$ for next iteration.

Optimization: We minimize the total energy function which also includes the term E_{BC} built using matrix $\mathbf{\Omega}^{k-1}$ from previous iteration.

$$\mathbf{x}^k = \arg\min \sum_{t=1}^{T} \beta \|\mathbf{\Psi}^{-1}\mathbf{x}_t\|_1 + \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \|\mathbf{\Omega}^{k-1}\mathbf{x}\|_2^2 \quad (5)$$

The above problem is convex but the problem size is significantly large. Even for a moderate sized video of $256X256$ pixels and 32 frames, we need to solve for 2 million variables. We use a fast algorithm designed for large systems, based on fixed point continuation [9], to solve the optimization problem. In all our experiments we fix the parameters at $\beta = 10^{-5}$ and $\lambda = 10^{-1}$. In practice, our algorithm converges in 5 iterations.

5. Experimental Results

We rigorously evaluate the performance and reconstruction fidelity on several challenging datasets. First, we simulate P2C2 in software by capturing fast events with a standard high-speed camera.

5.1. Simulation on high speed videos

Figure 6 shows example reconstructions of high-speed sub-frames at 16X temporal super-resolution. Notice that while normal camera frames are highly blurred, the reconstruction retains sharpness and high frequency texture detail is maintained. Several of our examples contain complex and non-linear motion. Most examples also contain several objects moving independently causing occlusion and disocclusions. To better understand the quality of reconstruction with varying spatio-temporal compression factors, examine Figure 8. This video has highly complex motion, where different dancers are performing different motions. There is significant non-rigidity in motion and challenging occlusion effects. Notice that our reconstruction retains high fidelity even at high compression factors. Even a compression factor of $4 \times 4 \times 4 = 64$ produces acceptable visual quality and $24dB$ PSNR.

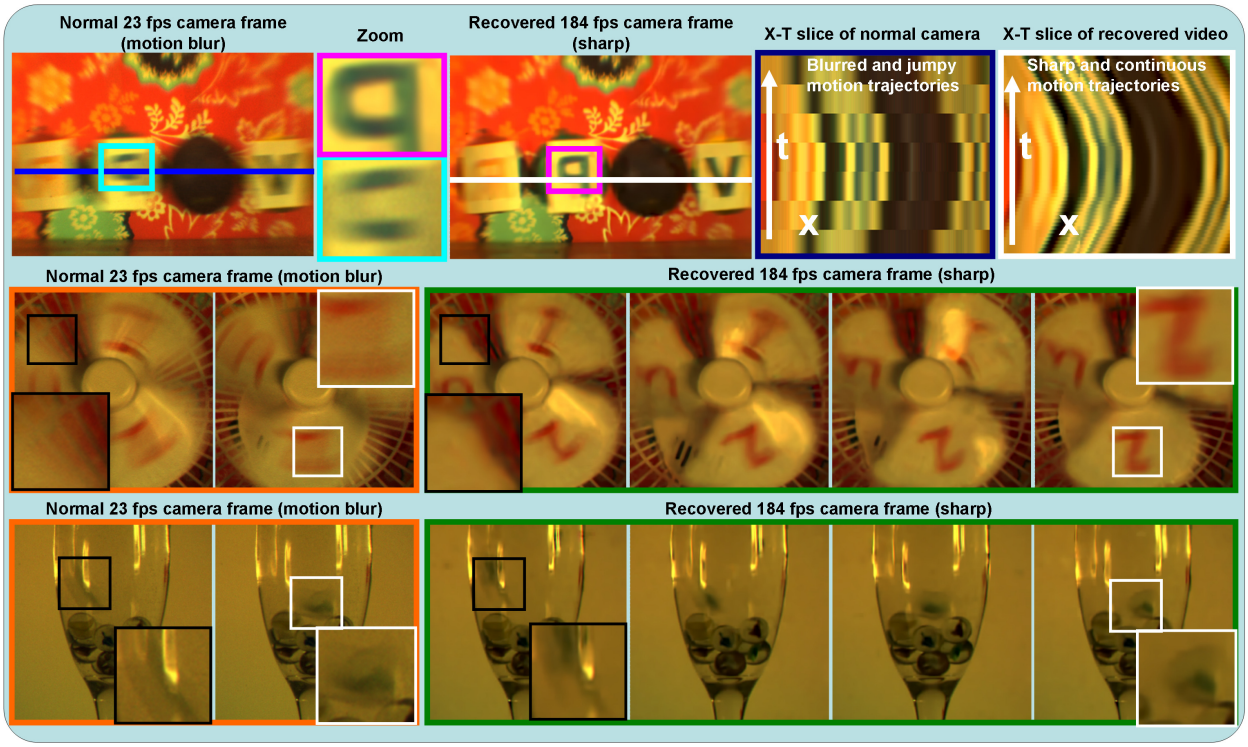

Figure 7. Results on LCOS prototype: For dataset on top, one frame from normal 23 fps camera, our recovered video and zoom-in insets are shown. The fourth and fifth column shows the X-T slices of original and recovered video. For middle and bottom datasets two images from normal 23 fps camera and four recovered images are shown.

Figure 8. Effect of spatio-temporal upsampling factors on Dancers video. Notice that our reconstructions retain visual fidelity even in presence of complex non-rigid multi-body motions and occlusions. High-speed video credit TECH IMAGING.

5.2. Results on P2C2 prototype datasets

We captured several datasets using our prototype device. The camera was operated at 23 fps and 8 different masks were flipped during the integration time of sensor. This allows us to reconstruct the sub-frames at a frame rate of 184 fps (23×8). We note that in our experimental setup we were limited by the field of view since the beamsplitter size forced us to use a lens with a large focal length .

In Figure 7, we show three different datasets. In the pendulum dataset, four letters 'CVPR' were affixed to four out of five balls and the pendulum was swung. As shown in X-T slices the balls and the letters had significant acceleration and also change in direction of motion. The recovered frames are much sharper than the original 23 fps frames as shown in inset. Note that the characters are much clearer in the reconstructed frame despite a 40 pixel blur in each captured frame. Also, the X-T slice clearly shows reconstruction quality. On the other hand, a fine feature such as the thread is blurred out since the flow corresponding to it is hard to recover.

Next, we rotate a fan and capture it at 23 fps and reconstruct sharper and clearer frames at 184 fps as shown in white inset. During recovery, we do not assume that motion is rotational. Note that the normal camera frame has intensities integrated from both fan blade and the background mesh as shown in black inset. We can handle this sub-frame occlusion in our recovery as indicated by the clear background and foreground in the recovered frame.

Finally, we drop a marble in water and capture it at 23 fps and reconstruct at 184 fps. Again, we do not assume any motion path but still recover the curved path of the marble.

Note that, despite specularities in the scene our algorithm is robust.

6. Analysis

Choice of modulation masks: There are two requirements on modulation masks to obtain high fidelity reconstruction. Firstly, the temporal code at a given pixel should have broadband frequency response [17], such that none of the scene features are blurred irrevocably. Secondly, the temporal code at a local neighborhood of pixels should be different. This along with spatial smoothness assumption provides sufficient conditioning in a neighborhood to recover low resolution sub-frames during the initialization process (4). This initialization is important to extract optical flow estimates which are then propagated forward using the iterative framework. On the other hand, when brightness constancy constraints are available, a modulation mask with well-conditioned matrix is desirable. Given ground-truth BC constraints, we reconstruct sub-frames of Golf dataset at 16X temporal super-resolution under three different masks (Table 1). We see that random mask of P2C2 offers significant advantage over the 'all one' mask and flutter shutter. We believe that proper theoretical analysis will lead to the design of optimal modulation masks and this is an area of future work.

	P2C2	**'All one'**	**FS**
PSNR in dB	26.2	21	16

Table 1. Reconstructing Golf dataset at 16X temporal super-resolution with different masks with ground truth BC constraints.

Comparison with prior art: We compare P2C2 with flexible voxels (FV) [8] on a fast phenomenon shown in Figure 9. Flexible voxels reconstruction suffers from two disadvantages: spatial smoothness is introduced in moving parts of the scene leading to blurred features and since the coding sequence for flexible voxels is mostly zeros, it leads to a highly light-inefficient capture method leading to performance degradation in the presence of noise. In $16\times$ temporal upsampling example shown in Figure 9, the high temporal resolution reconstruction of FV is noisier than our reconstruction.

Effect of spatio-temporal upsampling: To evaluate the impact of varying upsampling factors, we perform statistical evaluation of sub-frame reconstructions using P2C2 on several challenging high-speed videos. These videos have very different spatial and motion characteristics. We carefully selected the dataset to ensure that it spans a large range of videos in terms of spatial texture, light level, motion magnitude, motion complexity, number of independent moving objects, specularities, varying material properties. Shown in Figure 10 is a plot of reconstruction PSNR (in dB) as a function of spatial and temporal upsampling for various datasets. From our visual inspection we note that reconstructions with PSNR of 30dB or greater have sufficient

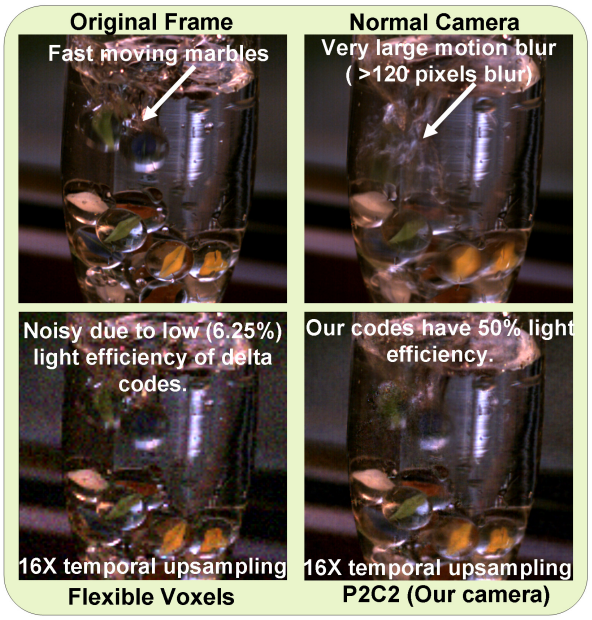

Figure 9. One frame of a video sequence of marble dropped in water. 40dB sensor noise was added. Our reconstruction is less noisy (zoom for better view) than those of flexible voxels due to higher light efficiency of P2C2.

textural sharpness and motion continuity to be called good quality reconstructions. From the figure, it is clear that we can achieve $8-16$X temporal upsampling and retain reconstruction fidelity. Also, when we perform 32X spatio-temporal super-resolution using P2C2 (2X2X8 or 4X4X2) we obtain acceptable reconstruction fidelity. Few frames from the reconstructions and their corresponding PSNR values are also shown in Figure 6 and 8 to relate visual quality to PSNR.

Benefits: Our imaging architecture provides three advantages over conventional imaging architectures. It significantly reduces the bandwidth requirement at the sensor by exploiting the compressive sensing paradigm. It improves light throughput of the system compared to acquiring a short exposure low frame-rate video and allows acquisition at low light levels. These are significant advantages since the prohibitive cost of high-speed imagers, is essentially due to the requirement for high bandwidth and high light sensitivity. Finally, the imaging architecture is flexible allowing incorporation of several other functionalities including high dynamic range (HDR) [13], assorted pixels [27] and flexible voxels [8].

Limitations: P2C2 exploits spatio-temporal redundancy in videos. Scenes such as a bursting balloon cannot be directly handled by the camera. Since the spatio-temporal redundancy exploited by traditional compression algorithms and our imaging architecture are very similar, as a thumb rule one can assume that scenes that are compressed efficiently can be captured well using our method. Our prototype uses a binary per-pixel shutter and this causes a 50%

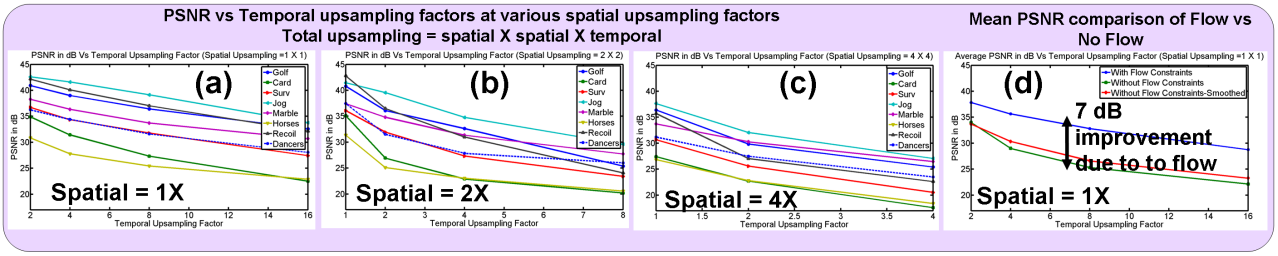

Figure 10. PSNR vs compression factors. (a) Spatial compression is kept at 1 and temporal compression is varied (b) Spatial compression is 2 in both dimensions. (c) Spatial compression is 4 in both dimensions. Notice that the video reconstruction fidelity remains high (PSNR$>$30dB) even at total compression factors of $16-32$. (d) Brightness Constancy constraints significantly improves the reconstruction.

reduction in light throughput. Since most sensors already have the ability to perform 'dual mode' integration (i.e., change the gain of pixels) we imagine the possibility of non-binary modulation in future. The algorithm is not real-time and this precludes the direct-view capability.

Conclusion: We presented Programmable Pixel Compressive Camera (P2C2), a new imaging architecture for high-speed video acquisition, that (a) reduces capture bandwidth and (b) increases light efficiency compared to related works. We also highlighted the importance of explicitly exploiting the brightness constancy constraints.

Acknowledgement

The authors thank John Barnwell for his invaluable help with the hardware. Thanks to Nitesh Shroff, Ming-Yu Liu, Petros Boufounos, Mohit Gupta, Amit Agrawal, Yuichi Taguchi, Jay Thornton, Jaishanker Pillai and Visesh Chari for helpful discussions. This research was conducted at MERL with support from MERL. Rama Chellappa was supported by MURI from the Army Research Office under the Grant W911NF-09-1-0383.

References

[1] www.photron.com.

[2] A. Agrawal, M. Gupta, A. Veeraraghavan, and S. Narasimhan. Optimal coded sampling for temporal super-resolution. In *IEEE Conference on CVPR*, pages 599–606, 2010.

[3] M. Ben-Ezra and S. K. Nayar. Motion-based motion deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:689–698, 2004.

[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(3):209–211, 2010.

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

[6] R. Fergus, B. Singh, A. Hertzmann, S. Roweis, and W. Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006*.

[7] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar. Coded rolling shutter photography: Flexible space-time sampling. In *IEEE International Conference on Computational Photography*, pages 1 –8, 2010.

[8] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. Narasimhan. Flexible Voxels for Motion-Aware Videography. *ECCV*, 2010.

[9] E. Hale, W. Yin, and Y. Zhang. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. *CAAM TR07-07, Rice Univ.*, 2007.

[10] A. Levin, P. Sand, T. Cho, F. Durand, and W. Freeman. Motion-invariant photography. In *ACM SIGGRAPH 2008*.

[11] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Cambridge, MA, USA, 2009.

[12] D. Mahajan, F. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur. Moving gradients: a path-based method for plausible image interpolation. In *ACM SIGGRAPH 2009*.

[13] H. Mannami, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Adaptive dynamic range camera with reflective liquid crystal. *J. Vis. Comun. Image Represent.*, 18:359–365, October 2007.

[14] R. F. Marcia and R. M. Willett. Compressive coded aperture video reconstruction. In *EUSIPCO 2008*, Lausanne, Switzerland, August 2008.

[15] S. K. Nayar, V. Branzoi, and T. E. Boult. Programmable Imaging: Towards a Flexible Camera. *International Journal on Computer Vision*, Oct 2006.

[16] J. Park and M. Wakin. A multiscale framework for compressive sensing of video. In *IEEE Picture Coding Symposium, 2009*, pages 1–4.

[17] R. Raskar, A. Agrawal, and J. Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *ACM SIGGRAPH 2006*.

[18] S. Ri, M. Fujigaki, T. Matui, and Y. Morimoto. Accurate pixel-to-pixel correspondence adjustment in a digital micromirror device camera by using the phase-shifting moiré method. *Applied optics*, 45(27):6940–6946, 2006.

[19] S. Ri, Y. Matsunaga, M. Fujigaki, T. Matui, and Y. Morimoto. Development of DMD reflection-type CCD camera for phase analysis and shape measurement. In *Proceedings of SPIE*, volume 6049, 2005.

[20] A. Sankaranarayanan, P. Turaga, R. Baraniuk, and R. Chellappa. Compressive Acquisition of Dynamic Scenes. *ECCV*, pages 129–142, 2010.

[21] Q. Shan, J. Jia, A. Agarwala, et al. High-quality motion deblurring from a single image. *ACM SIGGRAPH 2008*.

[22] Q. Shan, Z. Li, J. Jia, and C. Tang. Fast image/video upsampling. In *ACM SIGGRAPH Asia 2008*.

[23] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:531–545, 2005.

[24] N. Vaswani. Kalman filtered compressed sensing. In *IEEE International Conference on Image Processing*, pages 893–896, 2008.

[25] A. Veeraraghavan, D. Reddy, and R. Raskar. Coded strobing photography: Compressive sensing of high speed periodic videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):671 –686, 2011.

[26] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM SIGGRAPH 2005*.

[27] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar. Generalized Assorted Pixel Camera: Postcapture Control of Resolution, Dynamic Range, and Spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.