Towards Compressive Camera Networks

Kaushik Mitra[1], Ashok Veeraraghavan[1], Aswin C. Sankaranarayanan[2], and Richard G. Baraniuk[1]
[1] ECE Dept., Rice University, Houston, TX
[2] ECE Dept., Carnegie Mellon University, Pittsburgh, PA

Abstract

The scale and scope of multi-camera networks has advanced significantly; camera networks are now found not only in surveillance and access control applications but also in motion capture systems, light stages for reflectance acquisition, and large scale traffic monitoring. While the specifics of the individual applications differ, broader trends transcend these applications; examples of such trends relate to the scalability of the network in terms of the increasing number of cameras and the increasing capabilities of individual cameras. In this article, we study the role of compressive sensing (CS) in multi-camera networks and its central role in enabling scalability of network. Specifically, we focus on the central question of whether recent advances in CS and sparse representations can allow us to solve the scalability challenges in large-scale multi-camera networks. Some of this discussion is speculative, focused on the potential opportunities afforded by fundamentally re-architecting camera networks by exploiting CS based approaches to tackle the data deluge challenge inherent in multi-camera networks.

1. Introduction

Applications that benefit from multi-camera systems cover a wide spectrum --- ranging from the mundane (surveillance) to the exotic (markerless motion capture, crowd dynamics). In most of these applications, scalability of the system to encompass a very large number of cameras faces tremendous challenges in terms of data acquisition, transmission and processing costs. This is further acerbated when sensors are costly such as in high-speed photography and night-vision imaging. As an example, the number of cameras in a network can vary from a few cameras for video conferencing, to tens for monitoring buildings, to hundreds for traffic monitoring, and tens of thousands over a city such as London or New York. Thus, for large camera networks, storing, transmitting, and processing this huge amount of data is extremely challenging. The traditional paradigm of acquiring a high-resolution video feed and then compressing the data for transmission and storage places unreasonable resource requirements on the network. Clearly, there is a need for more efficient techniques for sensing and processing.

Traditional sampling is done on the basis of Shannon-Nyquist sampling theorem, which states that, to avoid loss of information, a signal should be sampled at a rate that is at least twice its signal bandwidth. For applications that involve a network of cameras, this leads to an inordinately high data-rate. This places a significant burden on the associated processing infrastructure. However, much of this is alleviated by the recently proposed theory of compressive sensing (CS) which suggests that we can exploit geometric structures inherent to most real-world signals to sense them at a rate that is often far smaller than the Nyquist rate [1, 2]. For example, natural images are *sparse* in wavelet domain --- in that, most of the energy of the signal is captured a small fraction of its wavelet coefficients. CS exploits this sparsity; it states that we can sample the signal at a rate proportion to the sparsity of the signal and recover the signal.

In this paper, we evaluate the potential for compressive cameras to alleviate sensing and processing requirements in a multi-camera network. We discuss the key concepts in video compressive sensing, which can be enhanced using multi-view constraints to enable sensing at smaller measurement rates. Given that CS provides significant improvements when sensing is costly, we envision that it could play a significant role in enabling multi-modal multi-camera networks of the future. Such networks would be

extremely powerful and capable of imaging scenes and objects in various dimensions of light including spectrum, time, polarization, and angle.

2. Compressive Cameras

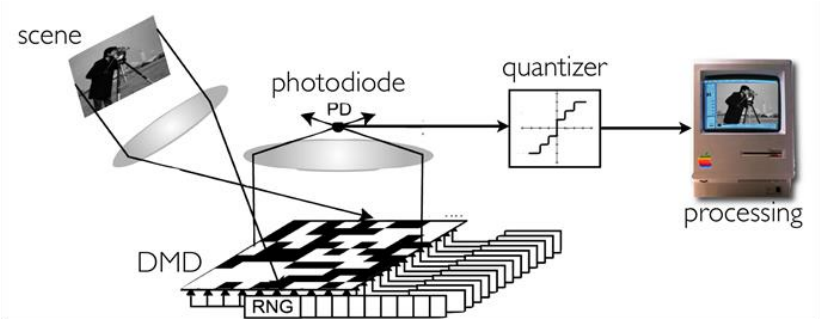
Cameras based on the CS theory hold significant promise to alleviate the data deluge problem. By adopting such cameras in place of traditional Nyquist samplers, we can often obtain significant reductions in processing while simultaneously enabling many desirable features. Let us begin by considering the impact of compressive sensing on a single camera before we address the impact on a network of cameras.

Single pixel camera (SPC). Rather than measuring intensity values on a 2D sampling of the scene, CS requires measurement of inner products between the scene and a set of test functions that satisfy the restricted isometry property. When the scene is compressible by an algorithm like JPEG or JPEG2000, CS enables stable reconstruction of an image of the scene from far-fewer measurements than the number of reconstructed pixels (see Figure 1(b)), thus resulting in sub-Nyquist image acquisition. The "single-pixel" CS camera architecture [3] is basically an optical computer (comprising a DMD, two lenses, a single photon detector, and an analog-to-digital (A/D) converter) that optically computes random linear measurements of the scene (see Figure 1(a)). The random CS measurements enable a tradeoff between space and time during image acquisition. Since the camera compresses during the acquisition, it has the capability to efficiently handle high-dimensional data from applications like video and hyper-spectral imaging. The camera design reduces the required size, complexity, and cost of the photon detector array down to a single unit, which enables the use of exotic detectors for sensing in non-visible wavebands that would be impossible in a conventional digital camera.

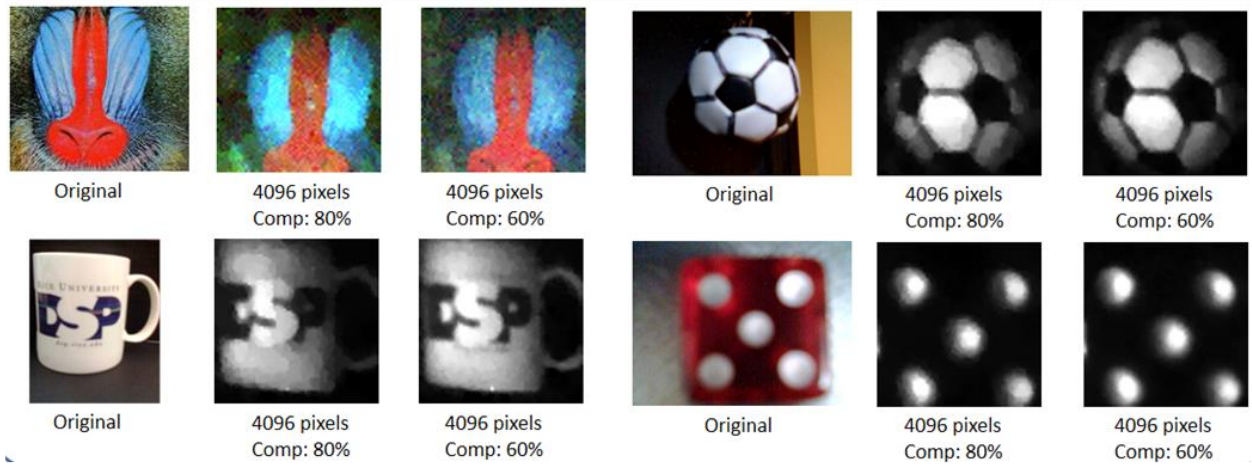
A key advantage of the SPC is in capturing signals beyond the visible spectra. Consumer digital cameras in the mega-pixel range are now ubiquitous thanks to the happy coincidence that the semiconductor material of choice for large-scale electronics integration (silicon) also happens to readily convert photons at visual wavelengths into electrons. In contrast, imaging at wavelengths where silicon is blind is often considerably more expensive. Thus, for comparable resolution, a $50 digital camera for the visible becomes a $50,000 camera for the infrared. The SPC design reduces the required size, complexity, and cost of the photon detector array down to a single unit, which enables the use of exotic detectors that would be impossible in a conventional digital camera that requires an array of sensors.

Video compressive sensing. The SPC can also be used for capturing videos. However, video CS is complicated by the ephemeral nature of dynamic events, which makes direct extensions of standard CS imaging architectures and signal models difficult. This requires the use of more sophisticated signal model and measurement strategies that are specific to the video sensing problem.

Parametric motion model (CS-LDS). One way to address this challenge is to narrow our scope to certain parametric models that are suitable for a broad class of videos; this morphs the video recovery problem to one of parameter estimation and provides a scaffold to address the challenges listed above. We have developed a CS framework for videos [4] modeled as linear dynamical systems (LDSs). Parametric models, like LDSs, offer lower dimensional representations for otherwise high-dimensional videos. This significantly reduces the number of free parameters that need to be estimated and, as a consequence, reduces the amount of data that needs to be sensed. In the context of video sensing, LDSs offer interesting tradeoffs by characterizing the video signal using a mix of dynamic/time-varying parameters and static/time-invariant parameters (see Figure 1(c)). Further, the generative nature of LDSs provides a prior for the evolution of the video in both forward and reverse time. To a large extent, this property helps us circumvent the challenges presented by the ephemeral nature of videos. We use sparse priors for the parameters of the LDS model. The core of the framework is a two-step measurement strategy that enables the recovery of the LDS parameters from compressive measurements by solving a sequence of linear and convex problems. For more details, see [4].



(a) Architecture of the single-pixel camera



(b) Image reconstructions from the single-pixel camera

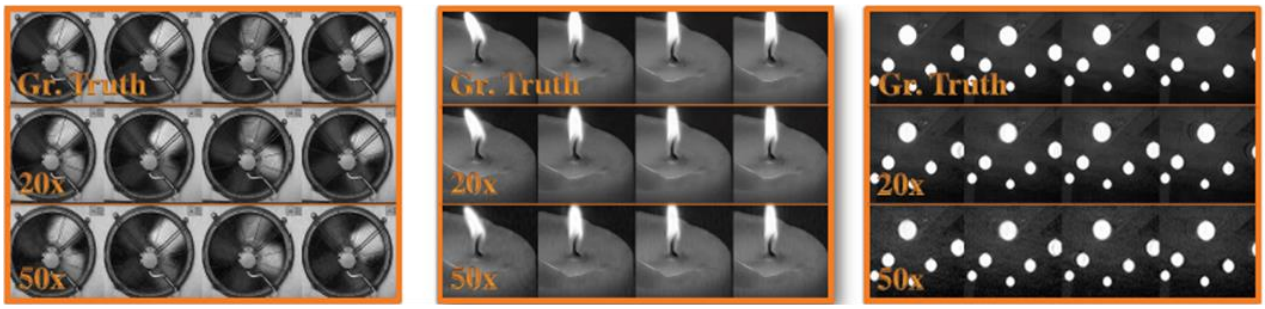


Figure 1. Compressive imaging: (a) Single-pixel camera block diagram [3]. Incident lightfield (corresponding to the desired image) is reflected off a digital micromirror device (DMD) array whose mirror orientations are modulated in the pseudorandom supplied by the random number generators (RNG). Each different mirror pattern produces a voltage at the single photodiode that corresponds to one measurement. (b) Examples of image reconstructions using the SPC. (c) A gallery of reconstruction results using the CS-LDS framework [4]. Each subfigure shows reconstruction results for a different video. The three rows of each subfigure correspond to, from top-bottom, the ground truth video and CS-LDS reconstructions at compression ratios of 20 and 50. Each column is a frame of the video and its reconstruction. Figure courtesy [3. 4].

General motion (CS MUVI): As discussed earlier, the key challenge with sensing videos with SPC is that the scene changes with every compressive measurement obtained. Though LDS can account for simple motion, it cannot handle complex motions. We circumvent this problem by designing special measurement matrices that enable a two-step recovery process; the first step is to estimate motion in the

scene and the second step is to recover the scene in full spatial and temporal resolution [5]. We design measurement matrices that simultaneously satisfy two properties: (i) contains high-spatial frequencies so as to recover videos at full spatial-resolution; and (ii) has close-to-optimal l2-recovery properties when downsampled. Given that Hadamard matrices are optimal, among the space of +/- 1 matrices, for l_2-recovery we design our dual-scale sensing (DSS) matrices by upsampling low-resolution Hadamard matrices and adding random sign-flips to this. This ensures that the DSS matrices satisfy both properties we require. Least-squares recovery using our DSS matrices works extremely well. The key here is that our recovered result is at a lower spatial resolution which has two main advantages: (i) lesser resolution implies a smaller dimensional signal to estimate and hence, lesser measurements; and (ii) in essence, we provide a tradeoff between spatial blur (downsampling) and motion blur. This, coupled with least square recovery (no use of sparse approximation) and Hadamard matrices (that provide optimal linear recovery guarantees) gives us these high-quality initial estimates. We refer to these initial estimates as the preview. These are extremely fast to compute; all it requires is a matrix multiplication with the added advantage of the matrix enjoying a fast transform. In particular, the preview provides insight into the scene and its temporal evolution. Given the preview of a video, we use optical flow to estimate motion field between the preview images (upsampled to full resolution). Optical flow estimates can be written as a linear relationship between images (using bilinear interpolation model). We can now solve an l_1-recovery problem with compressive measurement constraints as well as optical flow constraints between frames. The recovered video has both high-spatial and temporal resolutions. Recently, an alternative sampling matrix and video reconstruction algorithm have been proposed which provides faster preview and removes the need for computing optical flow [6].

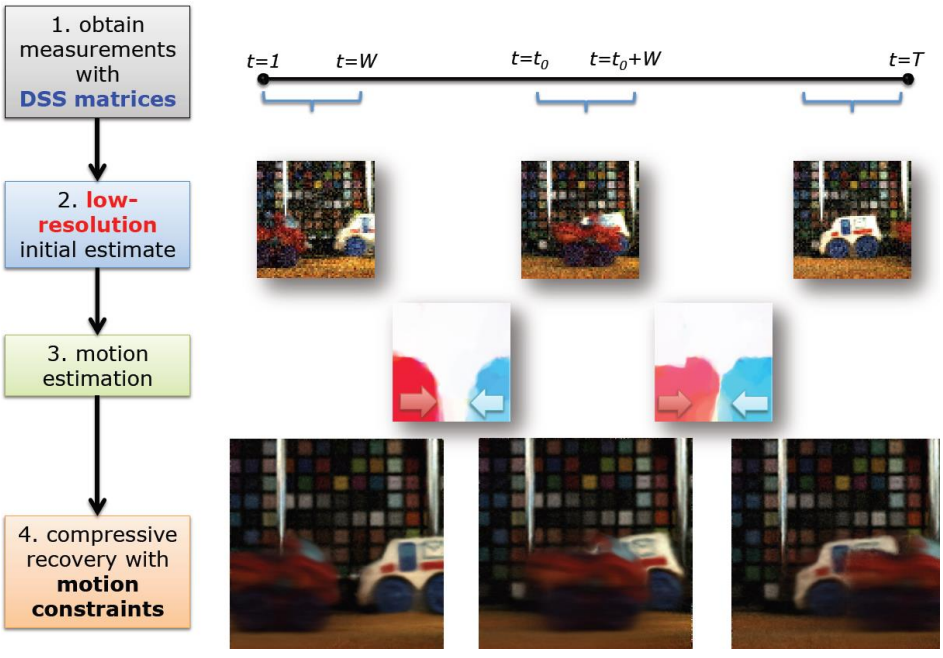


Figure 2. CS-MUVI, a framework for sensing compressed videos: The key challenge with sensing videos with SPC is that the scene changes with every compressive measurement obtained. Traditional l1-recovery methods fail in the presence of fast motion. We circumvent this problem by designing special measurement matrices that enable a two-step recovery process; the first step is to estimate motion in the scene and the second step is to recover the scene in full spatial and temporal resolution. Figure courtesy [5].

3. Compressive Cameras Networks

Given that CS based cameras can reduce the sampling requirements on individual cameras, they provide a potential solution to the data deluge in large scale camera networks. In addition to the direct sampling

advantages offered by compressive sensing to independent cameras, if the different cameras in a network have overlapping fields of view, then further reductions in sampling rate may be obtained by carefully treating the entire network as a single compressive imaging system rather than treating each camera as an independent compressive imager.

When a number of CS cameras are connected into a network, we can further reduce the data bandwidth by exploiting the redundancy across overlapping field of views (FOV) [12]. Consider the scenario shown in Figure 3, where a large area is being imaged using many single pixel cameras. In Figure 3(a) we show FOV of some of the single pixel cameras in the network. During reconstruction, instead of doing independent SPC reconstruction for each camera, we can perform joint reconstruction of the whole scene by first geometrically registering the images. The advantage of joint reconstruction is that pixels that are common to multiple cameras need fewer measurements. Suppose we want to reconstruct N pixels per SPC and there are M cameras with overlapping FOVs. If we perform independent reconstruction for each camera, then the total number of unknown parameters is *MN*. In contrast, if we perform joint reconstruction, with a fraction f of pixels being common to all camera, then the number of unknown is *MN(1-f)+Nf*. For large *N*, the number of unknown in joint reconstruction is a fraction *(1-f)* of the independent reconstruction case. Since the number of measurements required increases with the number of unknowns, we need fewer measurements for joint reconstruction. Thus, we can further reduce the data bandwidth. Registration across views is done using geometric transformations such as homography. To estimate homography, we need to perform independent reconstruction of a single frame at each SPC. After computing homography from this frame, we can use the joint reconstruction scheme. Figure 3(b-c) show the advantage of joint reconstruction (26.2 dB) over independent reconstruction (20.2 dB) for compression ratio of 4 at each SPC camera. Figure 3(d) shows the SNR vs. compression ratio for independent and joint reconstruction, from which it is clear that joint reconstruction scheme can be used for reducing data bandwidth. A similar approach has been considered in [7], where the images are considered as points in a low-dimensional manifold and a manifold lifting algorithm has been proposed for joint reconstruction.

(a) FOV of some SPC camera (b) Independent SPC reconstruction (20.2 dB) (c) Joint SPC reconstruction (26.2 dB) (d) SNR vs. compression ratio

Figure 3. Data bandwidth reduction in network of compressive cameras: (a) The original image is of size 512*512 and is being observed by 9 SPC cameras (Figure courtesy Skybox Imaging, Inc.). The boxes show FOV of 3 such cameras. (b) Independent reconstruction of each SPC produce poor reconstruction (20.2 dB) for compression ratio of 4 at each SPC. (c) Joint reconstruction of all the SPC after geometric registration produces better result (26.2 dB) for the same compression ratio. (d) SNR vs. compression ratio plot clearly show the advantage of joint SPC reconstruction over independent reconstruction. Thus, using joint SPC reconstruction, we can reduce the data bandwidth.

4. Inference in Compressive Camera Networks

The ultimate goal of most camera networks is to perform an inference task such as tracking, object recognition, or activity recognition. The usual way to perform inference in CS network would be to

recover the original video at each SPC using one of recovery algorithms discussed in Section 2. However, recovering the original video is computationally taxing. A better approach would be to perform inference directly on the compressed measurements. We provide example of two inference tasks: background subtraction and object recognition, where we perform inference directly from compressed samples.

Direct background subtraction from compressive measurements. Background subtraction is fundamental in automatically detecting and tracking moving objects with applications in surveillance, teleconferencing and even 3D modeling. Usually, the foreground or *the innovation* of interest occupies a sparse spatial support, as compared to the background and may be caused by the motion and the appearance change of objects within the scene. For CS camera network, it is desirable to directly reconstruct the sparse foreground innovations within a scene without any intermediate image reconstruction. The main idea is that the background-subtracted images can be represented sparsely in the spatial image domain and hence the CS reconstruction theory should be applicable for directly recovering the foreground. We use CS theory to directly recover the sparse innovations (foreground) of a scene [8]. We show that the object silhouettes (binary background subtracted images) can be recovered as a solution of a convex optimization or an orthogonal matching pursuit problem. In our method, the object silhouettes are learned directly using the compressive samples without any auxiliary image reconstruction. We can also simultaneously recover the appearance of objects using the compressive measurements. However, in this case, it may be necessary to reconstruct one auxiliary image. Figure 4 shows background subtraction experimental results using an SPC.

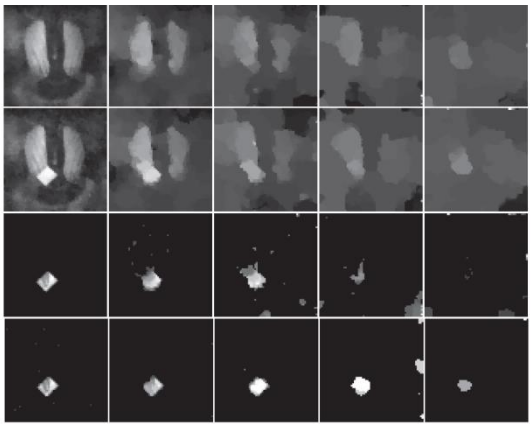
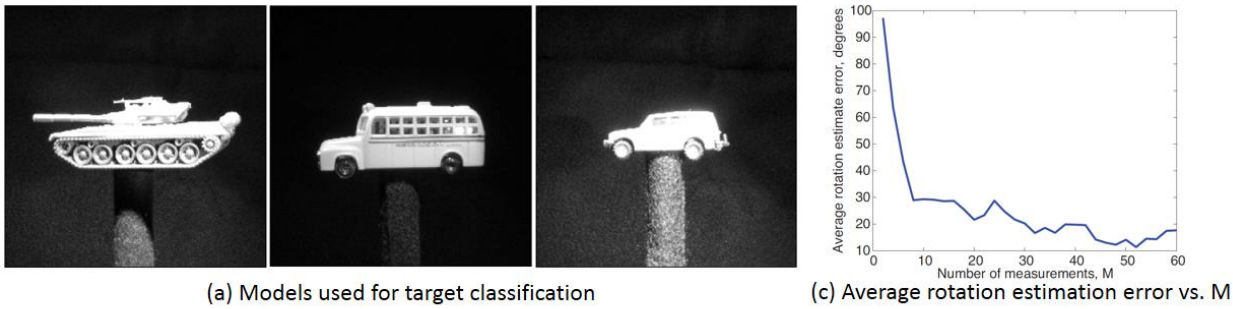


Figure 4. Background subtraction experimental results using an SPC. Reconstruction of background image (top row) and test image (second row) from compressive measurements. Third row: conventional subtraction using the above images. Fourth row: reconstruction of difference image directly from compressive measurements. The columns correspond to measurement rates of 50%, 5%, 2%, 1% and 0.5%, from left to right. Background subtraction from compressive measurements is feasible at lower measurement rates than standard background subtraction. Figure courtesy [8].

Compressive classification and target recognition. We propose a framework for compressive classification that operates directly on the compressive measurements without first reconstructing the image [9]. We dub the resulting dimensionally reduced matched filter the smashed filter. We map traditional maximum likelihood hypothesis testing into the compressive domain and find that the number of measurements required for a given classification performance level does not depend on the sparsity or compressibility of the images but only on the noise level. We then apply the generalized maximum likelihood method to deal with unknown transformations such as the translation, scale, or viewing angle of a target object. We exploit the fact the set of transformed images forms a low-dimensional, nonlinear manifold in the high-dimensional image space. We find that the number of measurements required for a given classification performance level grows linearly in the dimensionality of the manifold but only logarithmically in the number of pixels/samples and image classes.

We evaluate the smashed filter in an image target classification setting. We consider three classes, each for a different vehicle model: a tank, a school bus, and a truck, see Figure 5(a). All images are of size 128×128 pixels and all measurement matrices are binary ortho-projectors obtained from a random number generator. We use real data from SPC and perform target recognition with unknown rotations of the three targets in the z-axis in R^3. We assume that we do not know the explicit structure of the three manifolds; hence we use training data to provide an estimate of the manifold structure. We acquired a training set of compressive measurements for each vehicle for rotation angles that are multiples of 10◦ (10◦, 20◦, . . . , 360◦). We first estimated the most likely rotation angle for each class by computing the nearest neighbor



(a) Models used for target classification (c) Average rotation estimation error vs. M

Figure 5. Direct target recognition from compressed measurement. **(a)** M**odels used for target recognition experiment. We performed classification experiments for different numbers of compressive measurements, varying from M = 2 to 60. (b)** C**onfusion matrices for M = 2, 4, and 6. The confusion matrices summarize the distributions for elements belonging to a given class (one per row: tank, school bus, or truck) being assigned a given class label (one per column: tank, school bus, or truck). The diagonal elements show the probabilities of correct classification for each of the classes. The matrices show that performance improves as M increases. Specifically, for M ≥ 6, the classification rate remains at 100%. (c)** A**verage rotation estimate error as a function of the number of measurements. Figure courtesy [9].**

from each class and then performed nearest neighbor classification. We evaluate the performance of the smashed filter classifier using leave-one-out testing. The measurements for each rotation/class combination were classified using a smashed filter trained on all other available data points. We performed classification experiments for different numbers of compressive measurements, varying from M = 2 to 60. Figure 5(b) provides confusion matrices for M = 2, 4, and 6. The diagonal elements show the probabilities of correct classification for each of the classes. The matrices show that performance improves as M increases. Specifically, for M ≥ 6, the classification rate remains at 100%. Figure 5(c) plots the average rotation estimate error as a function of the number of measurements.

5. The Future of Compressive Camera Networks

Recognition tasks such as object or activity recognition are very difficult machine vision problems. Given that ultimately we want to perform such tasks, it is imperative that we not just collect more data but collect data with different modalities such as hyper-spectral data. We can use camera network for capturing different modalities of the visual signal such as multi-spectral data and high dynamic range. Multi-modal data is best captured using a camera array with small spacing between the cameras. We call such an array of camera as Generalized Assorted Cameras (GAC) [10]. We place a mosaic of filters is placed in front of an array of cameras. Since the different cameras in a GAC do not share a single viewpoint, correspondence across cameras needs to be established. Luckily, the recent success in multi-view stereo and structure from motion has shown that sub-pixel dense correspondences can be reliably obtained. We leverage these existing state-of-the-art techniques to establish dense correspondence across a GAC array and warp the data obtained from the different cameras to the viewpoint of a reference camera. The GAC then essentially acts like a GAP camera [11], where diverse filters are placed directly on the photo-detectors. The primary advantage of the GAC architecture is that the external filter mosaic can be customized for particular applications which require spectral selectivity, high dynamic range etc.

GAC assumes that the measurements of a scene point's radiance from cameras with slight different viewpoints are identical. As long as the cameras are close together relative to the depths of the different objects in the scene, this requirement is usually true for a broad range of scene reflectance. If we obtain dense pixel correspondence between all the cameras, then we can warp the views from the different sensors onto a canonical "central" camera, transforming it into a multi-dimensional image sensor. The central camera is usually the sensor whose location is closest to the median camera location. The set of warped images forms a cube of data, much like a hyper-spectral cube. Note that multi-view reconstruction

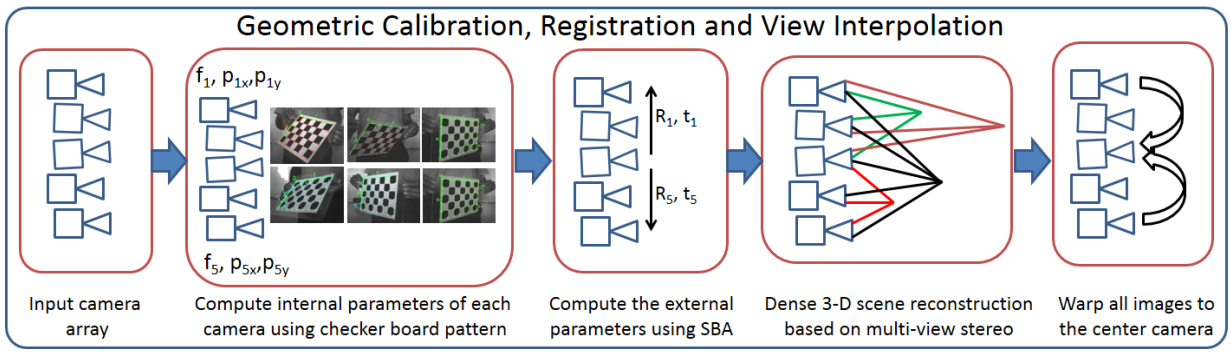


Figure 6. View interpolation in video. Calibration prior to video capture: 1) compute the internal parameters (focal length and principal points) of all the cameras separately, 2) compute the external parameters (rotation and translation w.r.t. the center camera). For each frame: 3) obtain a dense 3-D point cloud using groups of camera, and 4) warp all the images to the center camera based on the 3-D point cloud. Figure courtesy [10].

works best when all the camera characteristics are almost identical. Yet, GAC requires that the characteristics of our sensors are diverse and assorted. Therefore, some near-identical subset of the cameras must be available for scene reconstruction. To warp all images to a central camera with sub-pixel accuracy we perform the procedure outlined in Figure 6, namely: 1) for all cameras calibrate the intrinsic parameters, such as focal length and principle points, 2) compute the extrinsic parameters, rotation and translation, of all the cameras with respect to the central camera, 3) perform dense 3-D reconstruction using groups of cameras, 4) using the 3-D point reconstruction warp all of the images to the viewpoint of the central camera.

Multi-spectral imaging. Traditional cameras are unable to differentiate between metameric scene points, which have the same RGB values but different spectral composition. The GAC framework allows for increased spectral resolution to capture these differences by using narrowband filters in front of some cameras in the array. In general, the specifications of the filters such as bandwidth and the central wavelengths of the filters are application dependent. For example, suppose that we wish to capture multi-spectral information within the range 410nm-710nm. Using a 5x5 camera array, it would be possible to use 16 filters with a bandwidth of 10nm while leaving the remaining cameras vacant to compute point correspondences. The center wavelengths of the filters are chosen to span the range of interest and are shown in Fig. 7(b). We implement our GAC using a ProFUSION 5x5 camera array distributed by PointGrey. Each camera has a resolution of 640x480 pixels, a Bayer color filter array, and a throughput of 15 frames per second. Additionally, image acquisition is synchronized but each camera offers independent control of gain and exposure duration. In our experiments the gain was held constant across all cameras.

An image of the camera array and the filter configuration used to directly sample multiple spectral bands is shown in Figure 7(a) and (b). We first capture multi-spectral images of an assortment of fruit in front of a multi-colored background. Since the ProFUSION sensor contains a Bayer mosaic, the captured images are demosaiced and then converted into a luminance image. The nine spectral bands are shown in Figure 5(e). Notice that the citrus fruits and bananas are dark in the blue and green wavelengths (below 580nm) and bright in the yellow, orange and red wavelengths, while the green apple is brightest at 550nm. After capturing the spectral images, false RGB images can be constructed for ease of display.

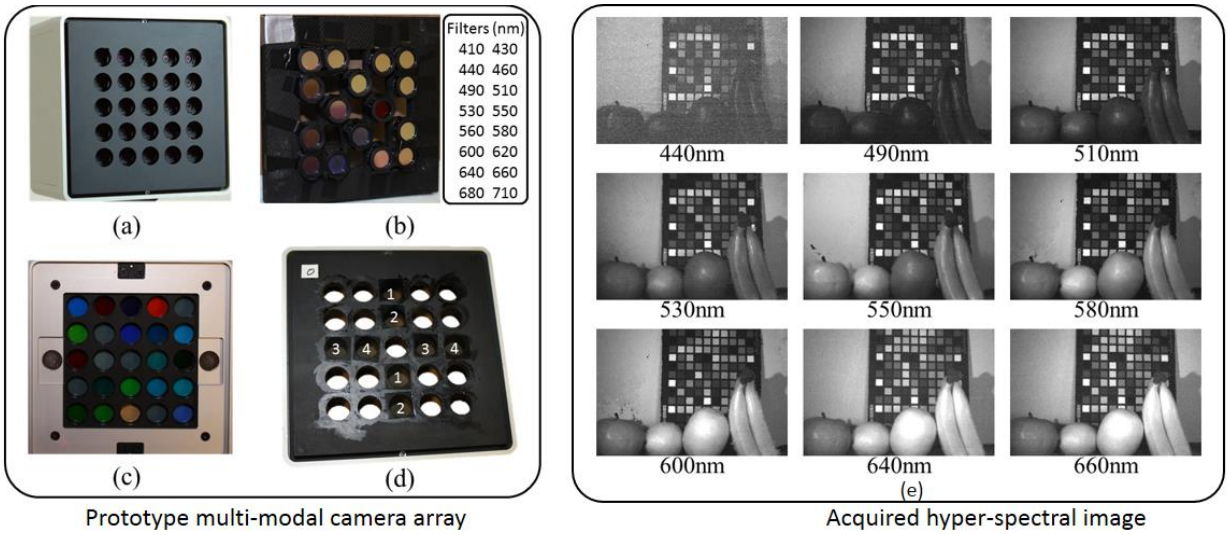


Figure 7. Prototype GAC: (a) The ProFUSION 5x5 color camera array records video at 15 fps. (b) The location of narrowband filters used to directly measure multi-spectral video. (c) Broadband filters used to capture multiplexed spectral images. (d) Location of pairs of linear polarization filters. The right subfigure shows 9 of the 16 spectral bands recorded using narrowband filters in front of the camera array. Images are displayed from the viewpoint of the central camera. From left the fruits are a green apple, a lemon, an orange, and bananas. Figure courtesy [10].

6. Conclusions

Networks of camera are becoming ubiquitous and their proliferation poses a data deluge challenge. In this paper, we hypothesize that compressive sensing can potentially tackle this resulting data deluge. Further, we envision, that this may lead to a future in which camera networks capture higher dimensional visual information such as hyper-spectral and light-field data, rather than just videos.

References

[1] M. B. Wakin and E. J. Candes, *An introduction to compressive sensing*, IEEE Signal Processing Magazine, 2008.

[2] R. G. Baraniuk, *Compressive sensing*, IEEE Signal Processing Magazine, 24(4), pp. 118-121, July 2007.

[3] 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 Processing Magazine, March 2008.

[4] A. C. Sankaranarayanan, P. K. Turaga, R. Chellappa and R. G. Baraniuk, *Compressive Acquisition of Linear Dynamical Systems*, SIAM Journal on Imaging Sciences, 2013.

[5] A. C. Sankaranarayanan, C. Studer, R. G. Baraniuk, *CS-MUVI: Video Compressive Sensing for Spatial-Multiplexing Cameras*, IEEE Intl. Conf. Computational Photography, 2012.

[6] T. Goldstein, L. Xu, K. F. Kelly and R. G. Baraniuk, *The STONE transform: Multi-resolution Image enhancement and real-time compressive video*, arXiv abs/1311.3405, 2013.

[7] J. Y. Park and M. B. Wakin, *A geometric approach to multi-view compressive imaging*, EURASIP J. Adv. Sig. Proc., 2012.

[8] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, R. Chellappa, *Compressive sensing for background subtraction*, ECCV, 2008.

[9] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly and R. G. Baraniuk, *The smashed filter for compressive classification and target recognition*, Proc. IS&T/SPIE Symposium on Electronic Imaging: Computational Imaging, 2007.

[10] J. Halloway, *Increasing temporal, structural, and spectral resolution in images using exemplar based priors*, M.S. thesis, Department of ECE, Rice University, 2013.

[11] F. Yasuma, T. Mitsunaga, D. Iso and S. K. Nayar, *Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum*, IEEE Transactions on Image Processing, 2010.

[12] M. A. Davenport, C. Hegde, M. F. Duarte, and R. G. Baraniuk, *Joint manifolds for data fusion*, IEEE Transactions on Image Processing, vol. 19, no. 10, pp. 2580-2594, Oct., 2010.