A DISTRIBUTED COMPRESSIVE SAMPLING APPROACH FOR SCENE CAPTURE USING AN ARRAY OF SINGLE PIXEL CAMERAS

Vikas Ramachandra and Truong Nguyen

University of California, San Diego, La Jolla, CA 92093.

ABSTRACT

This paper presents a method of capturing 3D scene information using an array of single pixel cameras. Based on the recent results for distributed compressive sampling, it is shown here that there could be considerable savings in the measurements required to construct the whole scene, when the correlations between the images captured by the individual cameras in the array is exploited. A technique for doing so for an array of cameras separated by translations along one axis only is illustrated.

Index Terms— Cameras, stereo vision, signal sampling, compressive sensing.

1. INTRODUCTION

Conventional wisdom in image acquisition and sampling dictated that the number of samples in the frequency domain match the resolution of the image, i.e. total number of pixels. This hinged on Nyquist sampling theory, which stated that the number of samples required to reconstruct a signal accurately depends on its bandwidth. Recently, a technique called compressive sampling has shown that signals can be reconstructed fairly accurately from a set of observations far less than the resolution desired. Typical signals and images have some inherent structure, which is exploited by compressive sampling techniques. The same structure is also taken advantage of by image compression methods like transform coding and quantization JPEG. These compression methods use the fact that signals have sparse representation in some chosen basis, and one needs to store only such adaptively chosen transform coefficients instead of all the signal samples.

This usually works as follows: the complete signal is acquired, then transform coded, followed by encoding of the largest transform coefficients; rest of the coefficients are discarded. It is wasteful to retain all signal samples only to later reduce it to a compact representation. Instead, compressive sampling acquires the signal directly in its compact representation [1]. Recently, an actual single pixel camera architecture was proposed in [2] which is based on compressive sampling.

In the case of multiple sensors (Eg: Camera arrays) a distributed compressive sampling (DCS) theory has been developed [3]. In a typical DCS scenario, a number of sensors measure signals that are each individually sparse in some basis and also correlated from sensor to sensor. Each sensor independently encodes its signal by projecting it onto another, incoherent basis (such as a random one) and then transmits just a few of the resulting coefficients to a single collection point. Under the right conditions, a decoder at the collection point can reconstruct each of the signals precisely. The DCS theory tries to exploit the joint sparsity of a signal ensemble.

In this paper, we propose a method for 3D scene capture based on the DCS framework. We propose the use of multiple single pixel cameras to form a camera array. We outline the procedure for the acquisition of the common sparse component as well as the innovative part of the signal at each camera. This paper is organized as follows: Section 2 briefly outlines compressive sampling and section 3 describes the single pixel camera. This is followed by section 4 which describes the DCS model we use. Section 5 explains our proposed method. Results are presented in section 6 and section 7 concludes the paper.

2. COMPRESSIVE SAMPLING (CS)

This section provides a brief summary of the compressive sampling theory [1], [5]. Consider a length N real valued signal (a vectorized signal for images) x of any dimension indexed as $x(n), n \in 1, 2, \cdots, N$. Let the basis $\Psi = [\psi_1, psi_2, \cdots, \psi_N]$ provide a K sparse representation of x, in which case we have:

$$x = \Psi\theta = \sum_{n=1}^{N} \theta(n)\psi_n = \sum_{l=1}^{K} \theta(n_l)\psi_{n_l} \qquad (1)$$

Here, x is written as a linear combination of k vectors chosen from Ψ, n_l being the indices of these vectors, and $\theta(n)$ are the coefficients. In matrix notation, x is a $N \times 1$ column vector, the sparse basis matrix Ψ is $N \times N$ with the basis vectors ψ_n as columns, and θ is a $N \times 1$ columns vector with K nonzero elements, i.e. $\| \theta\|_0 = K$ where $\| \cdot \|_p$ is the l_p norm.

We measure and encode $M < N$ projections of the signal onto a second set of basis functions. In matrix notation, we

Email: {vikas,nguyent}@ucsd.edu, Website: videoprocessing.ucsd.edu

measure

$$y = \phi x \tag{2}$$

where y is a $M \times 1$ column vector and the measurement basis matrix ϕ is $M \times N$ with each row a basis vector ϕ_m. Since $M < N$, recovery of the signal x from the measurements y is ill-posed in general; however the additional assumption of signal sparsity makes recovery possible and practical. The CS theory tells us that when certain conditions hold, namely that the basis ϕ_m cannot sparsely represent the elements of the basis ψ_m (a condition known as incoherence of the two bases), and M being large enough, one can recover the signal.

The recovery of the sparse set of significant coefficients $\theta(n)$ can be achieved using optimization by searching for the signal with l_0-sparsest coefficients $\theta(n)$ that agrees with the M observed measurements in y, i.e. one needs to solve for:

$$\hat{\theta} = \arg\min \|\theta\|_0 \text{ s.t. } y = \phi\psi\theta \tag{3}$$

For a K sparse signal, $K+1$ random measurements need to be made. Since solving Equation (3) is NP complete, the l_0 optimization is replaced with an l_1 norm based optimization, the price paid being that the number of measurements now need to be $M = cK$ where $c > 1$ is an oversampling factor. This is also called basis pursuit.

3. CS BASED SINGLE PIXEL CAMERA

In the system in [2], a Texas Instruments (TI) digital micro mirror device (DMD) is used. The DMD consists of an array of electrostatically actuated micro-mirrors where each mirror the array is suspended above an individual SRAM cell. The DMD micro-mirrors form a pixel array of size 1024×768. Each mirror rotates about a hinge and can be positioned in one of two states (+12 degrees and -12 degrees from horizontal); thus light falling on the DMD may be reflected in two directions depending on the orientation of the mirrors. The light from a given configuration of the DMD mirrors is summed at the photodiode to yield an absolute voltage that yields a coefficient $y(m)$ for that configuration. The output is amplified through an op-amp circuit and then digitized by a 12-bit analog-to-digital converter. M such measurements are taken with M different (random) mirror configurations to get y. The image is recovered from these M measurements using l_1 norm based optimization.

4. DISTRIBUTED COMPRESSIVE SAMPLING (DCS)

In the model proposed in [3], all signals share a common sparse component while each individual signal contains a sparse innovation component; that is,

$$x_j = z_C + z_j, j \in 1, 2, , \cdots, J \tag{4}$$

with $z_C = \Psi\theta_C, \|\theta_C\|_0 = K$ and $z_j = \Psi\theta_j, \|\theta_j\|_0 = K$.

Thus, the signal z_C is common to all of the x_j and has sparsity K in basis. The signals z_j are the unique portions of the x_j and have sparsity K_j in the same basis.

It was shown in [3] that for 2 signals, given the $(K+K_1)c$ measurements for x_1 as side information, and assuming that the partitioning of x_1 into z_C and z_1 is known, cK_2 measurements that describe z_2 should allow reconstruction of x_2. Also, $(K + K_1 + K_2)c$ coefficients should suffice to reconstruct both x_1 and x_2, since we have $K + K_1 + K_2$ nonzero elements in x_1 and x_2. This can be extended to arbitrary number of signals.

5. THE PROPOSED DCS SINGLE PIXEL CAMERA ARRAY FRAMEWORK

The DCS model explained above fits our approach, where projections of the scene are captured as images on different cameras, which are all $K_1, K_2, \cdots$ sparse. Consider the simple case where the camera are all aligned along a single baseline (i.e. their relative positions vary by translations along one coordinate only). Although this is a fairly restrictive model, it is a very common setup for camera arrays. In this case, the projected images are related to each other through disparity maps (ignoring occlusions). The disparity maps are again K sparse in the same basis set as the individual projected images. (An example of such a basis is wavelets).

Our approach for the case of 2 single pixel cameras (which can be trivially extended to the N camera case) is explained below:

- Using camera 1, we take $K_1^{'}$ measurements of the image of the 3D scene as seen by that camera. Note that the relationship between K_1 and $K_1^{'}$ is that, $K_1^{'} = K_1 + K$. Here, it was assumed that the disparity map is K sparse in the basis. In other words, $K_1^{'}$ measurements enable us to reconstruct projected image of camera one accurately, without the need of extra measurements from other cameras.

- Using camera 2, we take K measurements of the scene as projected onto that camera. It was assumed that the disparity map is K sparse in the basis. Note: Here, it is implicitly assumed that the disparity map is far less complex than each of the projected camera images themselves (i.e. $K < K_1, K < K_2$), which is almost always the case. The disparity map almost always has a smaller dynamic range as well as smoother regions than the actual images.

- Based on the $K_1^{'}$ measurements, we reconstruct image for camera 1 accurately (Call it Image 1). Based on K measurements,we can reconstruct only a blurry version of the projected image onto camera 2 (Call it image 2'). This is because an accurate image would require $K_2^{'} = K_2 + K$ measurements.

- In order to get the full scene, our goal is to now reconstruct an accurate version of the projected image onto camera 2 (Call it image 2). Now we take advantage of the correlations between the 2 camera images as follows. We re-encode the Image 1 (with K nonzero coefficients) to obtain a blurred version called Image 1'. Using image 1' and Image 2', we calculate the disparity map D [4]. Using D and accurate Image 1, we reconstruct (accurate) image 2 by simply compensating Image 1 with the motion(s) indicated by the map D in different regions. Now, we can reconstruct the scene using accurate images Image 1 and Image 2.

Here, we outline how our method leads to a saving in the number of measurements jointly taken by taking advantage of the correlations between the images, as opposed to applying the compressed sampling framework for individual images captured by each of the single cameras in the camera array. In the case of N camera images, we would need $K_1^{'} + K_2^{'} + \cdots + K_N^{'}$ measurements if each image was compressive sampled independent of the other. In the case of the proposed framework, we would need only $K_1^{'} + (N-1)K$ measurements. For an illustration, in the case that each image is roughly of the same complexity, i.e. all the $K_i^{'}$'s are equal to K', then our framework results in a saving of $(N-1)K^{'} - K$ measurements.

As an example, consider N 256x256 images of the same scene captured from different positions. Each projected image would independently require about 30000 measurements for good quality reconstruction, making a total of $30000N$ measurements. In our framework, the first image would require 30000 measurements. It was found that to get a fairly accurate disparity map, about 20000 measurements suffice for the rest of the image, which makes it $30000 + (20000)(N-1)$ measurements, which for N=10, results in savings of $\frac{300000-(30000+20000*9)}{300000} \times 100$ = 30 percent of the measurements. Lesser number of measurements results in lesser time for data acquisition as well as faster communication of the information from the sensors to the main processor, which are important considerations for any sensor network.

6. RESULTS

In this section, we provide results for simulations of the camera array with 2 cameras; position of camera 2 varies from that of camera 1 by a translation along one axis only. Figures 1 and 5 show the original left and right images for the scenes considered ('Coke' and 'Dog'), we cropped regions of size 128x128 from each of the images for faster testing, which are shown in Figures 2 and 6 for either scene considered. 4000 measurements were taken to reconstruct the left image Image L fairly accurately. For the right image, 2000 measurements were taken and the blurry image Image R' was reconstructed. To get the disparity map, we use a blurred version of the left image (Image L') with image R'. The disparity map was constructed using the method outlined in [4]. The disparity map contains values from -8 to $+8$, which is a much smaller dynamic range as compared to the images themselves which have pixels ranging from 0 to 255. Figure 3 shows images R' and the blurred version L' for the 'Coke' scene. Figure 4 shows accurate image L along with the accurately reconstructed right image R using the disparity map's values to motion compensate the pixels in image L (all for the scene 'Coke'). Similarly, Figure 7 shows images R' and the blurred version L' for the 'Dog' scene. Figure 8 shows accurate image L along with the accurately reconstructed right image R using the disparity map's values to motion compensate the pixels in image L (all for the scene 'Dog').

Fig. 1. The actual full sized 'coke' L and R images

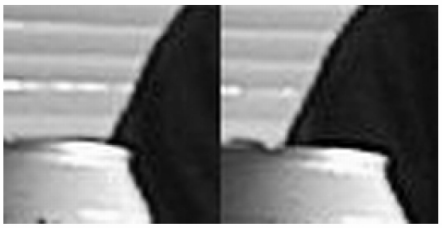

Fig. 2. Original cropped 'coke' L and R images

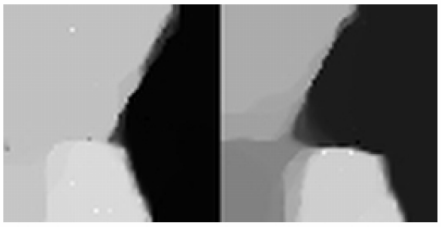

Fig. 3. Blurry reconstructed 'Coke' images: Image L' and Image R', used to build the disparity map.

From the results, it is clear that a fairly accurate reconstruction can be made for both the left and right images by using the correlations between them, as proposed.

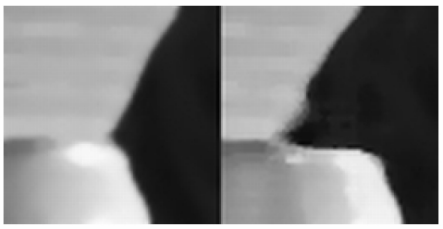

Fig. 4. Accurately reconstructed 'Coke' images using disparity map: Images L and R.

Fig. 5. The actual full sized 'Dog' L and R images

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a single pixel camera array framework which offers considerable savings in the number of measurements (using the theory of distributed compressed sampling) as compared to using the compressive sampling framework at each individual single pixel camera independent of others in the camera array. Although our framework is restricted right now to camera arrays in which the cameras are separated by a translation along one axis only, we believe that it is still useful since this array configuration is by far the most common. Future work could involve extension of the framework to camera arrays with general relationships between the component camera positions. Also, thus far, we have not taken occlusions into account, which are an important practical consideration. This could be addressed in the future.

8. REFERENCES

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

[2] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," *Proc. Computational Imaging IV at SPIE Electronic Imaging,* San Jose, CA, pp. 43-52, January 2006.

[3] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, "An information-theoretic approach to distributed compressed sensing," *Proc. 43rd Allerton Conf. Communication, Control, and Computing,* Monticello, IL, September 2005.

[4] A. S. Ogale and Y. Aloimonos, "Robust contrast invariant stereo correspondence," *Proc. IEEE Conf. on Robotics and Automation (ICRA),* pp. 819-825, April 2005.

[5] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory,* 52(4), pp. 1289 - 1306, April 2006.

Fig. 6. Original cropped 'Dog' L and R images

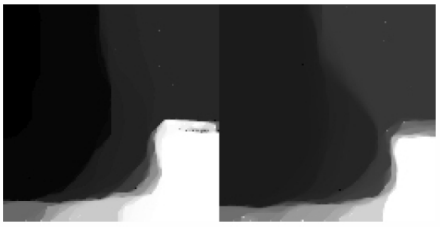

Fig. 7. Blurry reconstructed 'Dog' images: Image L' and Image R', used to build the disparity map.

Fig. 8. Accurately reconstructed 'Dog' images using disparity map: Images L and R.

View publication stats