Motion-adaptive compressive coded apertures

Zachary T. Harmany[a], Albert Oh[a], Roummel Marcia[b], and Rebecca Willett[a]

[a]Department of Electrical and Computer Engineering, Duke University, Durham NC, 27708 USA
[b]School of Natural Science, University of California-Merced, Merced, CA 95343 USA

ABSTRACT

This paper describes an adaptive compressive coded aperture imaging system for video based on motion-compensated video sparsity models. In particular, motion models based on optical flow and sparse deviations from optical flow (i.e. "salient" motion) can be used to (a) predict future video frames from previous compressive measurements, (b) perform reconstruction using efficient online convex programming techniques, and (c) adapt the coded aperture to yield higher reconstruction fidelity in the vicinity of this salient motion.

Keywords: Adaptive coded apertures, MURA, compressed sensing, optical flow, saliency, motion models

1. INTRODUCTION

In this paper we explore compressive sensing (CS) in a video context using compressive coded apertures. As described in Section 2, compressive coded apertures (CCAs) allow compressive measurement in relatively simple optical systems with theoretical performance guarantees. If implemented using spatial light modulators (SLMs), it is possible to change the aperture code over time, potentially adapting to the scene being sensed.

We describe one possible avenue for constructing an adaptive CCA system like this based on video motion models. In particular, we show that motion models can be used to predict future video frames based on past CCA measurements. By tracking where these motion models are least accurate, we can identify regions with time-varying occlusions or otherwise unpredictable or "salient" motion. Once we've identified these regions, we can adapt the SLMs in our architecture to increase our reconstruction accuracy in regions of salient motion, resulting in a compressive, automatic and adaptive foveated video system.

2. COMPRESSIVE CODED APERTURES

The proposed imaging architecture builds upon previous work on compressive coded aperture imaging.[1–3] In this snapshot architecture, as illustrated in Figure 1, the observations we acquire are modeled as

$$y_t = D\left(S_t \odot [M_t * (R_t \odot f_t)]\right) + n_t,$$

where $\odot$ denotes Hadamard (componentwise) multiplication and

- f_t is the scene at time t,
- y_t is the observation vector,
- n_t is noise,
- D corresponds to downsampling induced by FPA and determines how compressive the system is,
- S_t is the subsampler,[3]
- M_t is the CCA mask, and
- R_t is used to set the region of interest (ROI) at time t. (This is the primary adaptive component.)

Further author information, send correspondence to Z. T. Harmany, E-mail: zth@duke.edu Telephone: 1-919-619-9123. This work was supported by DARPA contract no. N66001-11-C-4001.

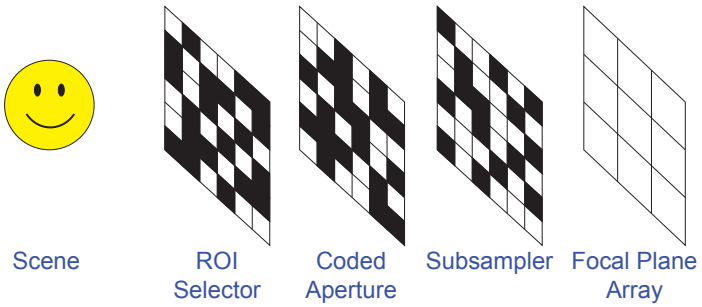


Figure 1: Schematic showing spatial light modulating components of the hypothetical system under study.

These are all linear operations; their concatenation is denoted A_t. The imaging model can be more compactly written (with slight overloading of notation) as

$$y_t = A_t f_t + n_t. \tag{1}$$

The structure of A_t is crucial in two respects. First, a key theorem characterizing the performance of compressive coded apertures was developed independently by Marcia and Willett[1] and Romberg.[3] A pseudo-random mask pattern M_t yields a sensing matrix A_t which satisfies (a weakened version of) the Restricted Isometry Property (RIP)[4] with high probability. Secondly, this structure allows for computationally efficient reconstruction algorithms since fast multiplications with A_t (and its adjoint) can be aided by the two-dimensional FFT.

3. SALIENT MOTION DETECTION AND MOTION ESTIMATION MODEL

We are given indirect, noisy observations $\{y_t\}$ of a dynamic scene $\{f_t\}$, that is

$$y_t = A_t f_t + n_t$$

where A_t is our sensing matrix at time t, and $n_t \sim N(0, \sigma^2 I)$ is sensor noise.

> *We wish to recover (a) the scene with high (task specific) fidelity, (b) the interframe motion and (c) salient (e.g. unpredictable) motion, and use these to subsequently used to adapt future measurements (A_t).*

In the current formulation, we pose the estimation of salient motion in a predictive context. Building upon existing literature,[5] we make the observation that one could predict the next frame f_{t+1} from the current frame f_t if one had knowledge of the optical flow $v_t = [v_{1,t}; v_{2,t}]$, and any regions of the image that were previously occluded. This prediction would be accurate up to any additional factors that are not explicitly modeled (e.g., changes in illumination, non-Lambertian reflection, deviations from a linear motion model). Mathematically, this model is given by

$$f_{t+1} = f_t - v_{1,t} \odot \Delta_1 f_t - v_{2,t} \odot \Delta_2 f_t + e_{1,t} + e_{2,t} \tag{2}$$

where Δ_1 and Δ_2 are the horizontal and vertical image gradient operations, $e_{1,t}$ are the salient motion regions, and $e_{2,t}$ contains deviations from this model (e.g. higher-order motion terms). Under this model, we make the assumption that the occluded regions have limited spatial extent, although it may have large magnitude. As such, we assume that is the vector $e_{1,t}$ will be sparse in that its ℓ_1 norm, i.e., $\|e_{1,t}\|_1$ will be small. Similarly, the error in the modeling assumptions could have a large spatial extent, but we expect that the size of each component of the error may be small, as such we assume that it will have small ℓ_2 norm, $\|e_{2,t}\|_2$. Finally, we assume that the image f_{t+1} and the optical flow components $v_{1,t}$ and $v_{2,t}$ are piecewise smooth with a small total variation.

We pose our estimation strategy in an online framework. When we collect new measurements at time $t+1$, we wish to produce an estimate of the current scene $\widehat{f}_{t+1}$ along with the optical flow and salient motion ($\widehat{v}_{1,t}$, $\widehat{v}_{2,t}$, and $\widehat{\epsilon}_{1,t}$) relating the previous frame f_t and the current frame f_{t+1}. For the previous frame, we very naturally use the estimate $\widehat{f}_t$ produced at the previous time step.

The estimation problem can be written as a convex program:

$$\underset{f_{t+1},v_t,e_{1,t}}{\text{minimize}}\ \frac{1}{2}\underbrace{\|A_{t+1}f_{t+1}-y_{t+1}\|_2^2}_{\text{data fit}}+\underbrace{\tau\|f_{t+1}\|_{\text{TV}}}_{\text{sparse gradient}}+\underbrace{\lambda\|e_{1,t}\|_1}_{\text{sparse salient motion}}$$
$$+\underbrace{\gamma\|f_{t+1}-\widehat{f}_t+v_{1,t}\odot\nabla_1\widehat{f}_t+v_{2,t}\odot\nabla_2\widehat{f}_t-e_{1,t}\|_2^2}_{\text{motion model fit}}+\underbrace{\mu(\|v_{1,t}\|_{\text{TV}}+\|v_{2,t}\|_{\text{TV}})}_{\text{smooth optical flow}}.$$

The regularization parameters $(\tau, \lambda, \gamma, \mu)$ weight the relative importance placed on the different properties in our model.

Because we have this formulation, we can use convex programming techniques to estimate the various components quickly and accurately. The current algorithm for solving the minimization simply performs a componentwise minimization, fixing all but one quantity while minimizing over the remaining. To facilitate rapid convergence, we initialize the minimization at the current time instance with the solutions from the previous time instance. This 'warm-starting' allows considerable computational savings as only few iterations per time instant are required.

The key ideas are that

- we can accurately predict most pixels in a future frame given an accurate motion model,
- we can estimate an accurate motion model directly from compressive measurements, and
- given this prediction, we can focus our sensing resources on where we anticipate the least accurate predictions.

This last is described in more detail in the following section.

4. ADAPTIVITY TO SALIENT MOTION

The salient motion estimate $e_{1,t}$ tells us where motion is least predictable and can be used to guide adaptive measurements. In particular, we define a region of interest (ROI) as a small neighborhood around large values of $e_{1,t}$ and collect compressive measurements of only this ROI. (This could be accomplished, for example, by using a spatial light modulator in the image plane of system, as sketched in Figure 1.)

This small set of observations can then be used in the reconstruction routine. Because the area of the ROI is generally smaller than the area of the scene, this strategy can be used to adaptively increase resolution in the vicinity of salient motion. Note that the accuracy of this approach hinges on *predictable* motion in the scene and salient motion locations; unexpected salient motion may be missed, making it necessary to alternate between large and small ROIs over time.

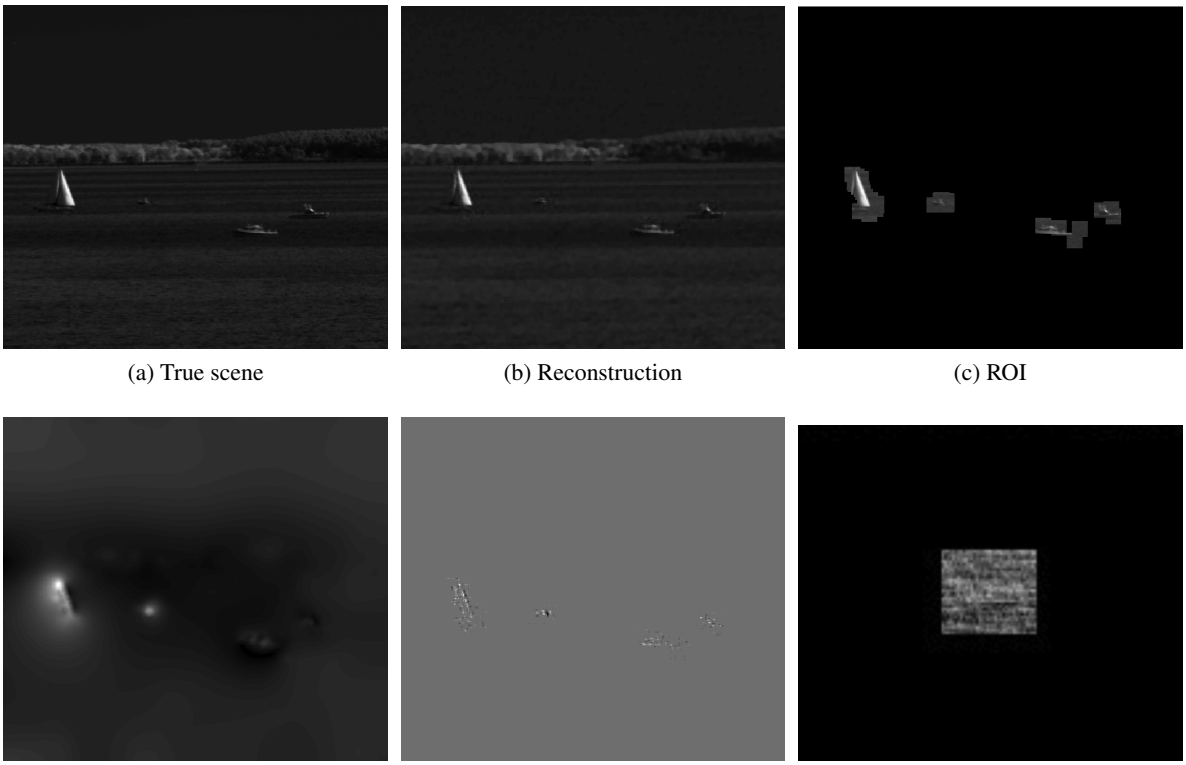
5. EXPERIMENTAL RESULTS

These methods were explored via simulations on a short-wave infrared (SWIR) video courtesy of Jon Nichols, NRL. Figure 2 displays the various components of (3) at frame 23. We can see that the reconstruction is accurate despite the limited amount of data collected. The ratio of reconstructed pixels to FPA elements is $16:1$. Furthermore, the estimated ROI is concentrated around the moving objects in the scene, and the velocity magnitude is largest in the vicinity of those moving objects.

Quantitative performance is plotted in Figure 3. We also zoom in on two moving objects for comparison purposes in Figure 4. This figure demonstrates that (a) the CCA approach yields significant resolution improvements over conventional downsampling sensors in low-noise environments, and (b) adapting the ROI increases the resolution and accuracy around moving objects in the scene.

6. CONCLUSIONS

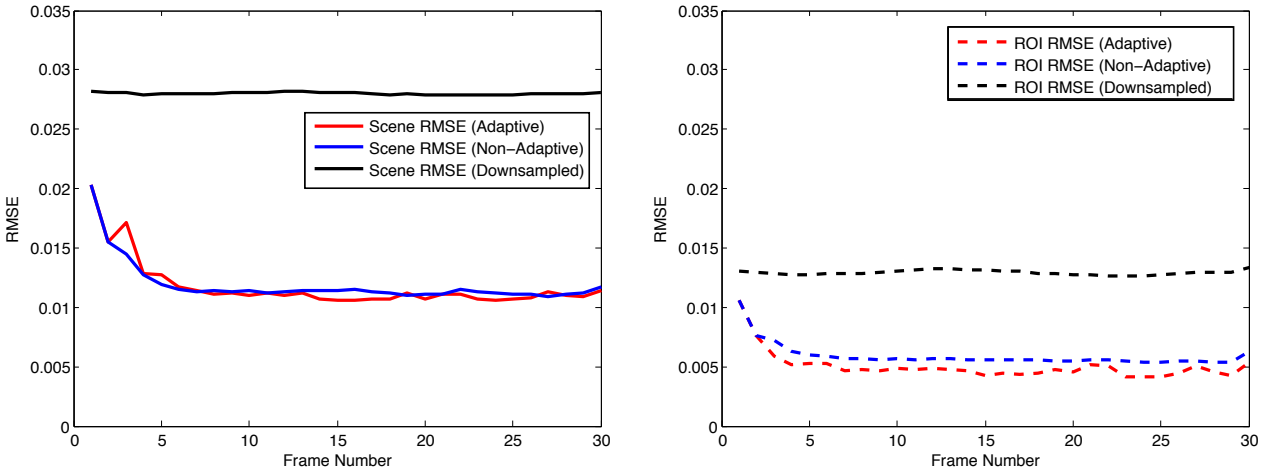
Motion models based on optical flow can be extracted from CCA data using online convex programming. These models allow us to distinguish salient (unpredicted) motion and adapt our aperture masks to the scene. The result is essentially motion-adaptive foveated imaging, with increase accuracy and resolution in the vicinity of salient motion. Finally, while ideas are explored in the context of compressive sampling, algorithms extend trivially to non-compressive settings.



(a) True scene (b) Reconstruction (c) ROI

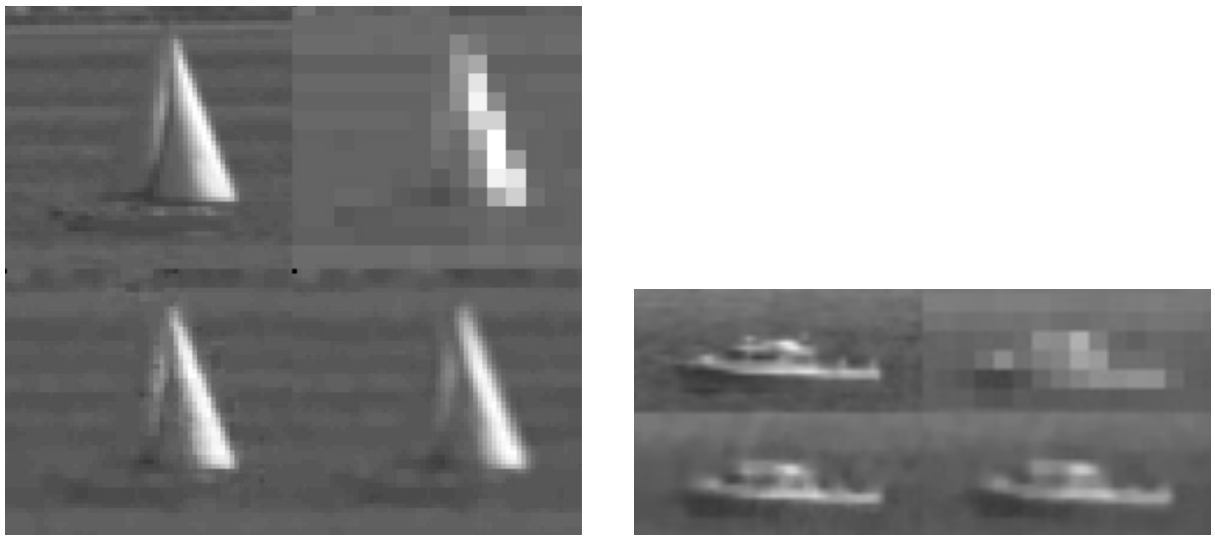
(d) Velocity magnitude (e) Salient motion (f) CCA observations

Figure 2: Snapshot of components of proposed method at 23rd frame.



(a) RMSE over entire frame (b) RMSE in ROI around sailboat

Figure 3: RMSE performance of proposed adaptive method, proposed reconstruction method with fixed, full-frame ROI, and uncoded, conventional downsampling.



(a) Sailboat (b) Speedboat

Figure 4: Comparison of methods displayed around two moving objects. Upper-left: Original. Upper-right: uncoded, convensional downsampling. Lower-left: proposed adaptive ROI method. Lower-right: proposed reconstruction with fixed, full-frame ROI.

REFERENCES

[1] R. Marcia and R. Willett, "Compressive coded aperture superresolution image reconstruction," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2008.

[2] R. Marcia, R. Willett, and Z. Harmany, *Optical and Digital Image Processing Fundamentals and Applications*, ch. Compressive Optical Imaging: Architectures and Algorithms. Wiley-VCH, 2011.

[3] J. Romberg, "Compressive sampling by random convolution," *SIAM Journal on Imaging Sciences* **2**(4), 2009.

[4] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory* **15**, pp. 4203–4215, December 2005.

[5] A. Ayvaci, M. Raptis, and S. Soatto, "Occlusion detection and motion estimation with convex optimization," in *Neural Information Processing Systems*, 2010.