Compressive Acquisition of Dynamic Scenes[*]

Aswin C. Sankaranarayanan[1], Pavan K. Turaga[2], Richard G. Baraniuk[1], and
Rama Chellappa[2]

[1] Rice University, Houston, TX 77005, USA
[2] University of Maryland, College Park, MD 20740, USA

Abstract. Compressive sensing (CS) is a new approach for the acquisition and recovery of sparse signals and images that enables sampling rates significantly below the classical Nyquist rate. Despite significant progress in the theory and methods of CS, little headway has been made in compressive video acquisition and recovery. Video CS is complicated by the ephemeral nature of dynamic events, which makes direct extensions of standard CS imaging architectures and signal models infeasible. In this paper, we develop a new framework for video CS for dynamic textured scenes that models the evolution of the scene as a linear dynamical system (LDS). This reduces the video recovery problem to first estimating the model parameters of the LDS from compressive measurements, from which the image frames are then reconstructed. We exploit the low-dimensional dynamic parameters (the state sequence) and high-dimensional static parameters (the observation matrix) of the LDS to devise a novel compressive measurement strategy that measures only the dynamic part of the scene at each instant and accumulates measurements over time to estimate the static parameters. This enables us to considerably lower the compressive measurement rate considerably. We validate our approach with a range of experiments including classification experiments that highlight the effectiveness of the proposed approach.

1   Introduction

Recent advances in the field of compressive sensing (CS) [4] have led to the development of imaging devices that sense at measurement rates below than the Nyquist rate. Compressive sensing exploits the property that the sensed signal is often sparse in some transform basis in order to recover it from a small number of linear, random, multiplexed measurements. Robust signal recovery is possible from a number of measurements that is proportional to the sparsity level of the signal, as opposed to its ambient dimensionality. While there has

been remarkable progress in CS for static signals such as images, its application to sensing temporal sequences or videos has been rather limited. Yet, video CS makes a compelling application towards dramatically reducing sensing costs. This manifests itself in many ways including alleviating the data deluge problems faced in the processing and storage of videos.

Existing methods for video CS work under the assumption of the availability of multiple measurements at each time instant. To date, such measurements have been obtained using a snapshot imager [20] or by stacking consecutive measurements from a single pixel camera (SPC) [8]. Given such a sequence of compressive measurements, reconstruction of the video has been approached in multiple directions. Wakin et al. [21] use 3D space-time wavelets as the sparsifying basis for recovering videos from snapshots of compressive measurements. Park and Wakin [12] use a coarse-to-fine estimation framework wherein the video, reconstructed at a coarse level, is used to estimate motion vectors that are subsequently used to design dictionaries for reconstruction at a finer level. Vaswani [16] and Vaswani and Lu [17] propose a sequential framework that exploits the similarity of support and the value the signal takes in this support between adjacent frames of a video. All of these algorithms require a large number of measurements at each time instant and, in most cases, the number of measurements is proportional to the sparsity of an individual frame. This is unsatisfactory as at this compression ratio it is possible to stably reconstruct the individual frames by themselves.

Video CS stands to benefit immensely with the use of strong models characterizing the signals. Park and Wakin [12] use MPEG-like block-matching to improve sparsity of the signal by tuning a wavelet. Veeraraghavan et al. [18] propose a compressive sensing framework of periodic scenes using coded strobing techniques. In this paper, we explore the use of predictive/generative signal models for video CS that are characterized by static parameters. Predictive modeling provides a prior for the evolution of the video in both forward and reverse time. By relating video frames over small durations, predictive modeling helps to reduce the number of measurements required at a given time instant. Models that are largely characterized by static parameters help in eliminating problems arising from the ephemeral nature of dynamic events. Under such a model, measurements taken at *all* time instants contribute towards estimation of the static parameters. At each time instant, it is only required to sense at the rate sufficient to acquire the dynamic component of the scene, which could be significantly lower than the sparsity of an individual frame of the video. One dynamic scene model that exhibits predictive modeling as well as high-dimensional static parameters is the linear dynamical system (LDS). In this paper, we develop methods for the CS of dynamic scenes modeled as LDS motivated, in part, by the extensive use of such models in characterizing dynamic textures [5, 7, 14], matching shape sequences [19], and activity modeling and video clustering [15].

In particular, the paper makes the following contributions. We propose a framework called *CS-LDS* for video acquisition using a LDS model coupled with sparse priors for the parameters of the LDS model. The core of the proposed framework is a two-step measurement strategy that enables the recovery of LDS

parameters directly from compressive measurements. We solve for the parameters of the LDS using an efficient recovery algorithm that exploits structured sparsity patterns in the observation matrix. Finally, we demonstrate stable recovery of dynamic textures at very low measurement rates.

2 Background and prior work

Compressive sensing: Consider a signal $\mathbf{y} \in \mathbb{R}^N$, which is *K-sparse* in an orthonormal basis Ψ; that is, $\mathbf{s} \in \mathbb{R}^N$, defined as $\mathbf{s} = \Psi^T \mathbf{y}$, has at most K non-zero components. Compressive sensing [4, 6] deals with the recovery of $\mathbf{y}$ from undersampled linear measurements of the form $\mathbf{z} = \Phi \mathbf{y} = \Phi \Psi \mathbf{s}$, where $\Phi \in \mathbb{R}^{M \times N}$ is the measurement matrix. For $M < N$, estimating $\mathbf{y}$ from the measurements $\mathbf{z}$ is an ill-conditioned problem. Exploiting the sparsity of $\mathbf{s}$, CS states that the signal $\mathbf{y}$ can be recovered exactly from $M = O(K \log(N/K))$ measurements provided the matrix $\Phi \Psi$ satisfies the so-called *restricted isometry property* (RIP) [1].

In practical scenarios with noise, the signal $\mathbf{s}$ (or equivalently, $\mathbf{y}$) can be recovered from $\mathbf{z}$ by solving a convex problem of the form

$$\min \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{z} - \Phi \Psi \mathbf{s}\| \leqslant \epsilon \tag{1}$$

with ϵ a bound on the measurement noise. It can be shown that the solution to (1) is with high probability the K-sparse solution that we seek. The theoretical guarantees of CS have been extended to *compressible* signals [10]. In a compressible signal, the sorted coefficients of $\mathbf{s}$ decay rapidly according to a power-law.

There exist a wide range of algorithms that solve (1) under various approximations or reformulations [4, 3]. Greedy techniques such as CoSAMP [11] solve (1) efficiently with strong convergence properties and low computational complexity. It is also easy to impose structural constraints such as block sparsity into CoSAMP giving variants such as model-based CoSAMP [2].

Dynamic textures and linear dynamical systems: Linear dynamical systems represent a class of parametric models for time-series data. including dynamic textures [7], traffic scenes [5], and human activities [19, 15]. Let $\{\mathbf{y}_t, t = 0, \dots, T\}$ be a sequence of frames indexed by time t. The LDS model parameterizes the evolution of $\mathbf{y}_t$ as follows:

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{w}_t \quad \mathbf{w}_t \sim N(\mathbf{0}, R), R \in \mathbb{R}^{N \times N} \tag{2}$$

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{v}_t \quad \mathbf{v}_t \sim N(\mathbf{0}, Q), Q \in \mathbb{R}^{d \times d} \tag{3}$$

where $\mathbf{x}_t \in \mathbb{R}^d$ is the hidden state vector, $A \in \mathbb{R}^{d \times d}$ the transition matrix, and $C \in \mathbb{R}^{N \times d}$ is the observation matrix.

Given the observations $\{\mathbf{y}_t\}$, the truncated SVD of the matrix $[\mathbf{y}]_{1:T} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$ can be used to estimate both C and A. In particular, an estimate

of the observation matrix C is obtained using the truncated SVD of $[\mathbf{y}]_{1:T}$. Note that the choice of C is unique only up to a $d \times d$ linear transformation. That is, given $[\mathbf{y}]_{1:T}$, we can define $\hat{C} = UL$, where L is an invertible $d \times d$ matrix. This represents our choice of coordinates in the subspace defined by the columns of C. This lack of uniqueness leads to structured sparsity patterns which can be exploited in the inference algorithms.

3 Compressive acquisition of linear dynamical systems

For the rest of the paper, we use the following notation. At time t, the image observation (the t-th frame of the video) is $\mathbf{y}_t \in \mathbb{R}^N$ and the hidden state is $\mathbf{x}_t \in \mathbb{R}^d$ such that $\mathbf{y}_t = C\mathbf{x}_t$, where $C \in \mathbb{R}^{N \times d}$ is the observation matrix. We use $\mathbf{z}$ to denote compressive measurements and Φ and Ψ to denote the measurement and sparsifying matrices respectively. We use ":" subscripts to denote sequences, such as $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ and $[\cdot]_{1:T}$ to denote matrices, such as $[\mathbf{y}]_{1:T}$ is the $N \times T$ matrix formed by $\mathbf{y}_{1:T}$ such that the k-th column is $\mathbf{y}_k$.

One of the key features of an LDS is that the observations $\mathbf{y}_t$ lie in the subspace spanned by the columns of the matrix C. The subspace spanned by C forms a static parameter of the system. Estimating C and the dynamics encoded in the state sequence $\mathbf{x}_{1:T}$ is sufficient for reconstructing the video. For most LDSs, $N \gg d$, thereby making C much higher dimensional than the state sequence $\{\mathbf{x}_t\}$. In this sense, the LDS models the video using high information rate static parameters (such as C) and low information rate dynamic components (such as $\mathbf{x}_t$). This relates to our initial motivation for identifying signal models with parameters that are largely static. The subspace spanned by C is static, and hence, we can "pool" measurements over time to recover C.

Further, given that the observations $\mathbf{y}_t$ are compressible in a wavelet/Fourier basis, we can argue that the columns of C need to be compressive as well, either in a similar wavelet basis. This is also motivated by the fact that columns of C encodes the dominant motion in the scene, and for a large set of videos, this is smooth and has sparse representation in a wavelet/DCT basis or in a dictionary learnt from training data. We can exploit this along the lines of the theory of CS. However, note that $\mathbf{y}_t = C\mathbf{x}_t$ is a bilinear relationship in C and $\mathbf{x}_t$ which complicates direct inference of the unknowns. Towards alleviating this non-linearity, we propose a two-step measurement process that allows to estimate the state $\mathbf{x}_t$ first and subsequently solve for a sparse approximation of C. We refer to this as the *CS-LDS* framework.

3.1 Outline of the CS-LDS framework

At each time instant t, we take two sets of measurements:

$$\mathbf{z}_t = \begin{pmatrix} \check{\mathbf{z}}_t \\ \tilde{\mathbf{z}}_t \end{pmatrix} = \begin{bmatrix} \check{\Phi} \\ \tilde{\Phi}_t \end{bmatrix} \mathbf{y}_t = \Phi_t \mathbf{y}_t, \tag{4}$$
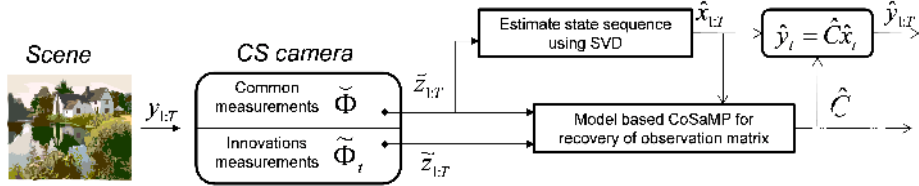
Fig. 1. Block diagram of the CS-LDS framework.

where $\check{\mathbf{z}}_t \in \mathbb{R}^{\widecheck{M}}$ and $\widetilde{\mathbf{z}}_t \in \mathbb{R}^{\widetilde{M}}$, such that the total number of measurements at each frame is $M = \widecheck{M} + \widetilde{M}$. Consecutive measurements from an SPC [8] can be aggregated to provide multiple measurements at each t under the assumption of a quasi-stationary scene. We denote $\check{\mathbf{z}}_t$ as *common* measurements since the corresponding measurement matrix $\check{\Phi}$ is the same at each time instant. We denote $\widetilde{\mathbf{z}}$ as the *innovations* measurements.

The CS-LDS, first, solves for the state sequence $[\mathbf{x}]_{1:T}$ and subsequently, estimates the observation matrix C. The common measurements $[\check{\mathbf{z}}]_{1:T}$ are related to the state sequence $[\mathbf{x}]_{1:T}$ as follows:

$$[\check{\mathbf{z}}]_{1:T} = \begin{bmatrix} \check{\mathbf{z}}_1 \ \check{\mathbf{z}}_2 \ \cdots \ \check{\mathbf{z}}_T \end{bmatrix} = \check{\Phi} C \begin{bmatrix} \mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_T \end{bmatrix} = \check{\Phi} C [\mathbf{x}]_{1:T}. \tag{5}$$

The SVD of $[\check{\mathbf{z}}]_{1:T} = USV^T$ allows us to identify $[\mathbf{x}]_{1:T}$ up to a linear transformation. In particular, the columns of V corresponding to the top d singular values form an estimate of $[\mathbf{x}]_{1:T}$ up to a $d \times d$ linear transformation (the ambiguity being the choice of coordinate). When the video sequence is exactly an LDS of d dimensions, this estimate is exact provided $\widecheck{M} > d$. The estimate can be very accurate, when the video sequence is approximated by a d-dimensional subspace as discussed later in this section.

Once we have an estimate of the state sequence, say $[\hat{\mathbf{x}}]_{1:T}$, we can obtain C by solving the following convex problem:

$$(P1) \ \min \sum_{k=1}^{d} \|\Psi^T \mathbf{c}_k\|_1, \text{ subject to } \|\mathbf{z}_t - \Phi_t C \hat{\mathbf{x}}_t\|_2 \leqslant \epsilon, \forall t \tag{6}$$

where $\mathbf{c}_k$ is the k-th column of the matrix C, and Ψ is a sparsifying basis for the columns of C. In Section 3.3, we show that the specifics of our measurements induce a structured sparsity in the columns of C, and this naturally leads to an efficient greedy solution.

To summarize (see Figure 1), the design of the measurement matrix as in (4) enables the estimation of the state sequence using just the common measurements, and subsequently solving for C using the diversity present in the innovations measurements $[\widetilde{\mathbf{z}}]_t$.

3.2 Random projections of LDS data

As mentioned earlier, when $[\mathbf{y}]_{1:T}$ lies exactly in the (column) span of thematrix C, then $[\check{\mathbf{z}}]_{1:T}$ lies in the span of $\check{\Phi} C$. Hence, the SVD of $[\check{\mathbf{z}}]_{1:T}$ can be used to

recover the state sequence up to a linear transformation, provided $\widetilde{M} \geqslant d$

$$[\check{\mathbf{z}}]_{1:T} = USV^T, \quad [\hat{\mathbf{x}}]_{1:T} = S_d V_d^T \tag{7}$$

where S_d is the $d \times d$ principal sub-matrix of S and V_d is the $T \times d$ matrix formed by columns of V corresponding to the largest d singular values. In practice, the observations $\mathbf{y}_t$ lie close to the subspace spanned by C such that projection of onto C makes for a highly accurate approximation of $\mathbf{y}_t$. In such a case, the estimate of the state sequence from the SVD of $[\check{\mathbf{z}}]_{1:T}$ is accurate only when the observations $\mathbf{y}_t$ are compressible [9]. In our case, this is equivalent to imposing a power-law decay on the singular values. Figure 2 shows the accuracy of the approximation of the estimated state sequence for various values of $\widetilde{M}$. This suggests that, in practice, $\mathbf{x}_t$ can be reliably estimated with $\widetilde{M} \propto d$.

3.3 Structured sparsity and recovery with modified CoSAMP

The SVD of the common compressive measurements $\check{\mathbf{z}}_t$ introduces an ambiguity in the estimates of the state sequence in the form of $[\hat{\mathbf{x}}]_{1:T} \approx L^{-1}[\mathbf{x}]_{1:T}$, where L is an invertible $d \times d$ matrix. Solving (P1) using this estimate will, at best, lead to an estimate $\hat{C} = CL$ satisfying $\mathbf{z}_t = \Phi_t \hat{C} \hat{\mathbf{x}}_t$. This ambiguity introduces additional concerns in the estimation of C. Suppose the columns of C are K-sparse (equivalently, compressible for a certain value of K) each in Ψ with support $\mathcal{S}_k$ for the k-th column. Then, the columns of CL are potentially dK-sparse with identical supports $\mathcal{S} = \bigcup_k \mathcal{S}_k$. The support is exactly dK-sparse when the $\mathcal{S}_k$ are disjoint and L is dense. At first glance, this seems to be a significant drawback, since the overall sparsity of $\hat{C}$ has increased to $d^2 K$. However, this apparent increase in sparsity is alleviated by the columns having identical supports. The property of identical supports on the columns of CL can be exploited to solve (P1) very efficiently using greedy methods.

Given the state sequence, we use a modified CoSAMP algorithm to estimate C. The modification exploits the structured sparsity induced by the columns of C having identical support. In this regard, the resulting algorithm is a particular instance of the model-based CoSAMP algorithm [2]. One of the key properties of model-based CoSAMP is that stable signal recovery requires only a number of measurements that is proportional to the model-sparsity of the signal, which in our case is equal to dK. Hence, we can recover the observation matrix from $O(dK \log(Nd))$ measurements [2]. Figure 3 summarizes the model-based CoSAMP algorithm used for recovering the observation matrix C.

3.4 Performance and measurement rate

For a stable recovery of the observation matrix C, we need in total $O(dK \log(Nd))$ measurements. In addition to this, for recovering the state sequence, we need a number of common measurements proportional to the dimensionality of the state vectors

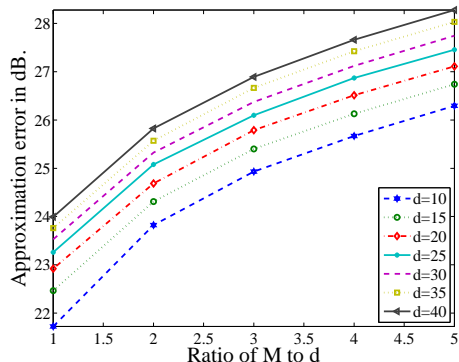$$MT \quad \propto \quad dK \log(Nd), \qquad \widetilde{M} \quad \propto \quad d. \tag{8}$$

Fig. 2. Average error in estimating the state sequence from common measurements for various values of state dimension d and the ratio $\widetilde{M}/d$. Statistics were computed using 114 videos of 250 frames taken from the DynTex database [13].

Compared to Nyquist sampling, we obtain a measurement rate (M/N) given by

$$\frac{M}{N} \quad \propto \quad \frac{dK \log(Nd)}{NT}. \tag{9}$$

This indicates extremely favorable operating scenarios for the CS-LDS framework, especially when T is large (as in a high frame rate capture). Consider a segment of a video of *fixed* duration observed at various sampling rates. The effective number of frames, T, changes with the sampling rate, f_s (in frames per second), as $T \propto f_s$. However, the complexity of the video measured using the state space dimension d does not change. Hence, as the sampling rate f_s increases, $\widecheck{M}$ can be decreased while keeping Mf_s constant. This will ensure that (8) is satisfied, enabling a stable recovery of C. This suggests that as the sampling rate f_s increases our measurement rate decreases, a very desirable property for high-speed imaging.

3.5   Extensions

Mean + LDS: In many instances, a dynamical scene is modeled better as an LDS over a static background, that is, $\mathbf{y}_t = C\mathbf{x}_t + \mu$. This can be handled with two minimal modifications to the algorithm described above. First, the state sequence $[\hat{\mathbf{x}}]_{1:T}$ is obtained by performing SVD on the matrix $[\widecheck{z}]_{1:T}$ modified such that the each row sums to zero. This works under the assumption that the sample mean of $\widecheck{z}_{1:T}$ is equal to $\widecheck{\Phi}\mu$, the compressive measurement of μ. Second, we use model-based CoSAMP to estimate both C and μ simultaneously. However, only the columns of C enjoy the structured sparsity model. The support of μ is not constrained to be similar to that of C.

$\widehat{C} = \text{CoSaMP\_Model\_Sparsity}(\Psi, K, \mathbf{z}_t, \widehat{\mathbf{x}}_t, \Phi_t, t = 1, \dots, T)$

Notation:
$\text{supp}(vec; K)$ returns the support of K largest elements of vec
$A_{|\Omega,\cdot}$ represents the submatrix of A with rows indexed by Ω and all columns.
$A_{|\cdot,\Omega}$ represents the submatrix of A with columns indexed by Ω and all rows.

$\forall t, \Theta_t \leftarrow \Phi_t \Psi$
$\forall t, \mathbf{v}_t \leftarrow \mathbf{0} \in \mathbb{R}^M$
$\Omega_{\text{old}} \leftarrow \phi$
While (stopping conditions are not met)

$\qquad R = \sum_t \Theta_t^T \mathbf{v}_t \widehat{\mathbf{x}}_t^T \qquad\qquad\qquad \left(R \in \mathbb{R}^{N \times d}\right)$
$\qquad k \in [1, \dots, N], \mathbf{r}(k) = \sum_{i=1}^d R^2(k, i) \qquad \left(\mathbf{r} \in \mathbb{R}^N\right)$

$\qquad \Omega \leftarrow \Omega_{\text{old}} \bigcup \text{supp}(\mathbf{r}; 2K)$

$\qquad \text{Find } A \in \mathbb{R}^{|\Omega| \times d} \text{ that minimizes } \sum_t \|\mathbf{z}_t - (\Theta_t)_{|\cdot,\Omega} A \widehat{\mathbf{x}}_t\|_2$

$\qquad B_{|\Omega,\cdot} \leftarrow A$
$\qquad B_{|\Omega^c,\cdot} \leftarrow 0$

$\qquad k \in [1, \dots, N], \mathbf{b}(k) = \sum_{i=1}^d B^2(k, i) \qquad \left(\mathbf{b} \in \mathbb{R}^N\right)$
$\qquad \Omega \leftarrow \text{supp}(\mathbf{b}; K)$
$\qquad S_{|\Omega,\cdot} \leftarrow B_{|\Omega,\cdot} \qquad S_{|\Omega^c,\cdot} \leftarrow 0$
$\qquad \Omega_{\text{old}} \leftarrow \Omega$
$\qquad \widehat{C} \leftarrow \Psi B$

$\qquad \forall t, \mathbf{v}_t \leftarrow \mathbf{z}_t - \Theta_t S \widehat{\mathbf{x}}_t$

Fig. 3. Pseudo-code of the model-based CoSAMP algorithm for CS-LDS.

4 Experimental validation

We present a range of experiments validating various aspects of the CS-LDS framework. Our test dataset comprises of videos from DynTex [13] and data we collected using high speed cameras. For most experiments, we chose $\widetilde{M} = 2d$, with d and K chosen appropriately. We used the mean+LDS model for all the experiments with the 2D DCT as the sparsifying basis for the columns of C as well as the mean. Finally, the entries of the measurement matrix were sampled from iid standard Gaussian distribution. We compare against *frame-by-frame* CS where each frame of the video is recovered separately using conventional CS techniques. We use the term *oracle LDS* for parameters and video reconstruction obtained by operating on the original data itself. The oracle LDS estimates the parameters using a rank-d approximation to the ground truth data. The reconstruction SNR of the oracle LDS gives an upper bound on achievable SNR. Finally, the ambiguity in observation matrix (due to non-uniqueness of the SVD
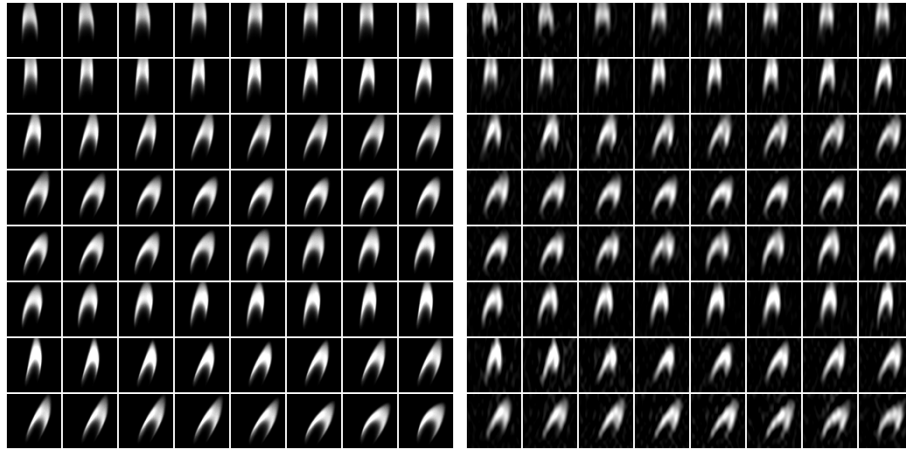
Fig. 4. Reconstruction of $T = 1024$ frames of a scene of resolution $N = 64 \times 64$ pixels shown as a mosaic. The original data was collected using a high speed camera operating at 1000 fps. Compressive measurements were obtained with $\tilde{M} = 30$ and $\check{M} = 20$, thereby giving a measurement rate $M/N = 1.2\%$. Reconstruction was performed using an LDS with $d = 15$ and $K = 150$. Shown above are 64 uniformly sampled frames from the ground truth (left) and the reconstruction (right).

based factorization) as estimated by oracle LDS and CS-LDS is resolved for visual comparison in Figures 5 and 6.

Reconstruction: Figure 4 shows reconstruction results from data collected from a high speed camera of a candle flame. Figure 5 shows the estimated observation matrix as well as the state sequence.

Figure 6 shows video reconstruction of a dynamic texture from the DynTex dataset [13]. Reconstruction results are under a measurement rate $M/N = 1/234$ (about 0.42%), an operating point where a frame-to-frame CS recovery is completely infeasible. However, the dynamic component of the scene is relatively small ($d = 20$) which allows us to recover the video from relatively few measurements. The SNR of the reconstructions shown are as follows: Oracle LDS = 24.97 dB, frame-to-frame CS: 11.75 dB and CS-LDS: 22.08 dB.

Performance with measurement noise: It is worth noting that the video sequences used in the experiments have moderate model fit error at a given value of d. The columns of C with larger singular values are, inherently, better conditioned to deal with this model error. The columns corresponding to the smaller singular values are invariably estimated at higher error. This is reflected in the estimates of the C matrix in Figures 5 and 6.

Figure 7 shows the performance of the recovery algorithm for various levels of measurement noise. The effect of the measurement noise on the reconstructions is perceived only at much lower SNR. This is, in part, due to the model fit error dominating the performance of the algorithm when the measurement noise SNR is very high. As the measurement SNR drops significantly below the model fit
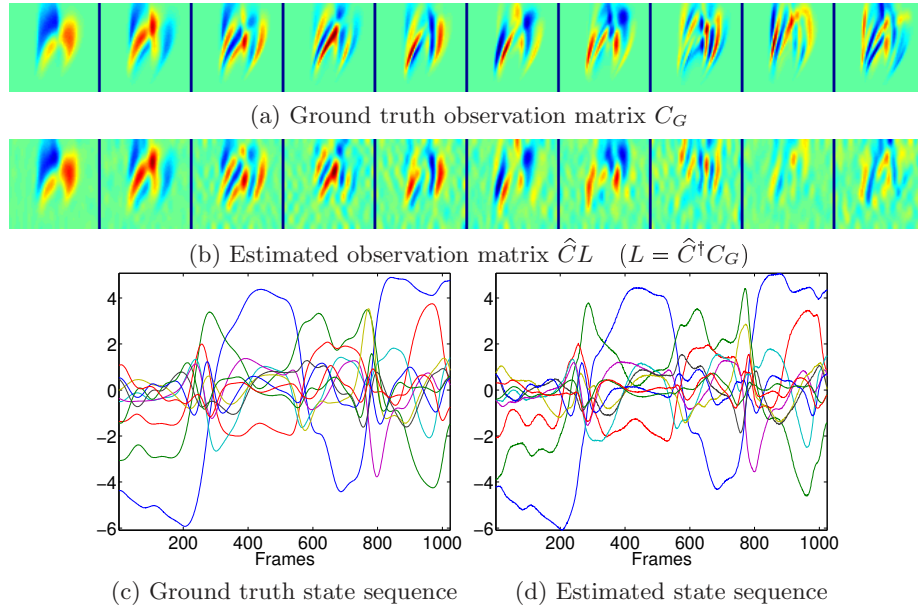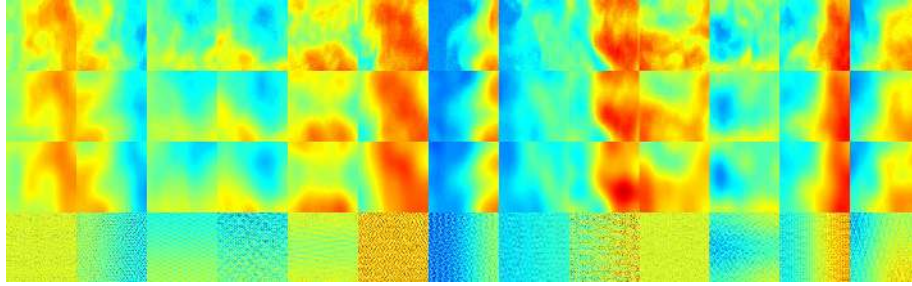
(a) Ground truth observation matrix C_G

(b) Estimated observation matrix $\widehat{C}L$ $(L = \widehat{C}^{\dagger}C_G)$

(c) Ground truth state sequence   (d) Estimated state sequence

Fig. 5. Ground truth and estimated parameters corresponding to Figure 4. Shown are the top 10 columns of the observation matrix and state sequences. Matlab's "jet" colormap (red= +large and blue= −large) is used in (a) and (b).

error, predictably, it starts influencing the reconstructions more. This provides a certain amount of flexibility in the design of potential CS-LDS cameras especially in scenarios where we are not primarily interested in visualization of the sensed video.

Sampling rate: Figure 8 shows reconstruction plots of the candle sequence (of Figure 4) for 1 second of video at various sampling rates. We use (9) to predict the required measurement rates at various sampling rates to maintain a constant reconstruction SNR. As expected, the reconstruction SNR remains the same, while the measurement rate decreases significantly with a linear increase in the sampling rate. This makes the CS-LDS framework extremely promising for high speed capture applications. In contrast, most existing video CS algorithms have measurement rates that, at best, remain constant as the sampling rate increases.

Application to scene classification: In this experiment, we study feasibility of classification problems on the videos sensed and reconstructed under the CS-LDS framework. We consider the UCSD traffic database used in [5]. The dataset consists of 254 videos of length 50 frames capturing traffic of three types: light, moderate, heavy. Figure 9 shows reconstruction results on a traffic sequence from the dataset. We performed a classification experiment of the videos into these three categories. There are 4 different train-test scenarios provided with the dataset. Classification is performed using the subspace-angles based metric

(a) Mosaic of frames of a video, with each column a different time instant, and each row a different algorithm. (top row to bottom) ground truth, oracle LDS, CS-LDS, and frame-by-frame CS.
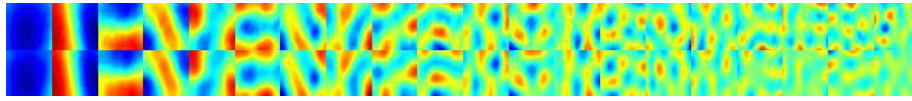
(b) Mosaic of ground truth (top) and estimated (bottom) observation matrix

Fig. 6. Reconstruction of a fire texture of length 250 frames and resolution of $N = 128 \times 128$ pixels. Compressive measurements were obtained at $\widetilde{M} = 30$ and $\widetilde{M} = 40$ measurements per frame, there by giving a measurement rate of 0.42% of Nyquist. Reconstruction was performed with $d = 20$ and $K = 30$. Frames of the videos are shown in false-color for better contrast.
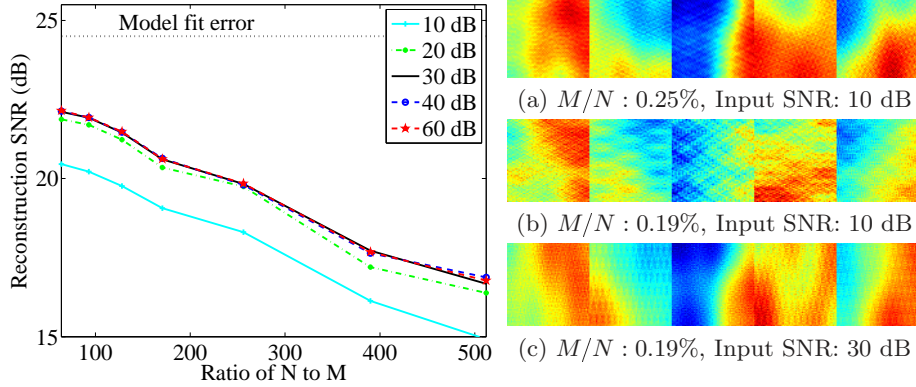

(a) $M/N : 0.25\%$, Input SNR: 10 dB

(b) $M/N : 0.19\%$, Input SNR: 10 dB

(c) $M/N : 0.19\%$, Input SNR: 30 dB

Fig. 7. Resilience of the CS-LDS framework to measurement noise. (Left) Reconstruction SNR as a function of measurement rates and input SNR levels computed using 32 Monte-Carlo simulations. The "black-dotted" line shows the reconstruction SNR for an $d = 20$ oracle LDS. (Right) Snapshots at various operating points. The dynamic texture of Figure 6 was used for this result.

with a nearest-neighbor classifier on the LDS parameters [14]. The experiment was performed using the parameters estimated directly without reconstructing the frames. For comparison, we also perform the same experiments with fitting
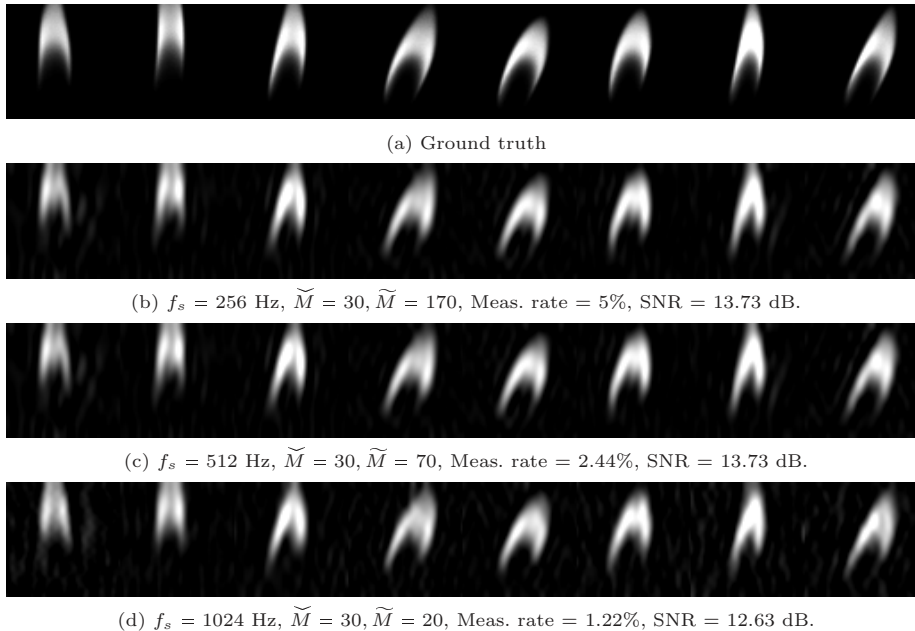
(a) Ground truth

(b) $f_s = 256$ Hz, $\breve{M} = 30$, $\widetilde{M} = 170$, Meas. rate = 5%, SNR = 13.73 dB.

(c) $f_s = 512$ Hz, $\breve{M} = 30$, $\widetilde{M} = 70$, Meas. rate = 2.44%, SNR = 13.73 dB.

(d) $f_s = 1024$ Hz, $\breve{M} = 30$, $\widetilde{M} = 20$, Meas. rate = 1.22%, SNR = 12.63 dB.

Fig. 8. As the sampling frequency f_s increases, we maintain the same reconstruction capabilities for significantly lesser number of measurements. Shown are reconstructions for $N = 64 \times 64$ and various sampling frequencies, achieved measurement rates, and reconstruction SNRs.

Table 1. Classification results (in %) on the traffic databases for two different values of state space dimension d. Results are over a database of 254 videos, each of length 50 frames at a resolution of 64×64 pixels under a measurement rate of 4%.

(a) $d = 10$

	Expt 1	Expt 2	Expt 3	Expt 4
Oracle LDS	85.71	85.93	87.5	92.06
CS-LDS	84.12	87.5	89.06	85.71

(b) $d = 5$

	Expt 1	Expt 2	Expt 3	Expt 4
Oracle LDS	77.77	82.81	92.18	80.95
CS-LDS	85.71	73.43	78.1	76.1

the LDS model on the original frames (oracle LDS). Table 4 shows classification results. We see that we obtain comparable classification performance using the proposed CS-LDS recovery algorithm to the oracle LDS. This suggests that the CS-LDS camera is extremely useful in a wide range of applications not tied to video recovery.

5   Discussion

In this paper, we proposed a framework for the compressive acquisition of dynamic scenes modeled as LDSs. We show that the strong scene model for the
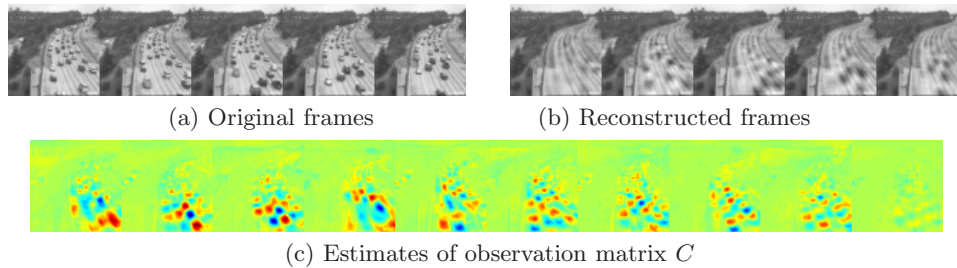
(a) Original frames

(b) Reconstructed frames

(c) Estimates of observation matrix C

Fig. 9. Reconstructions of a traffic scene of $N = 64 \times 64$ pixels at a measurement rate 4%, with $d = 15$ and $K = 40$. The quality of reconstruction and LDS parameters is sufficient for capturing the flow of traffic.

video enables stable reconstructions at very low measurement rates. In particular, this emphasizes the power of video models that are predictive as well as static.

Extensions of the CS-LDS framework: The CS-LDS algorithm proposed in this paper requires, at best, $O(d)$ measurements per time instant. This roughly corresponds to the number of degrees of freedom in the dynamics of the video under a d-dimensional LDS model. However, the state transition model of the LDS further constrains the dynamics by providing a model for the evolution of the signal. Incorporating this might help in reducing the number of measurements required at each time instant. Another direction for future research is in fast recovery algorithms that operate at multiple spatio-temporal scales, exploiting the fact that a global LDS model induces a local LDS model as well. Finally, much of the proposed algorithm relies on sparsity of the observation matrix C. Wavelets and Fourier (DCT) bases do not sparsify videos where the motion is localized in space. This suggests the use of dyadic partition methods such as platelets [22], which have been shown to have success in modeling bounded shapes.

Newer models for video CS: While the CS-LDS framework makes a compelling case study of LDSs for video CS, its applicability to an arbitrary video is limited. The LDS model is well-matched to a large class of dynamic textures such as flames, water, traffic etc. but does not extend to simple non-stationary scenes such as people walking. The importance of video models for CS motivates the search for models that are more general than LDS. In this regard, a promising line of future research is to leverage our new understanding of video models for compression algorithm-based CS recovery.

References

1. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constructive Approximation 28(3), 253–263 (2008)

2. Baraniuk, R., Cevher, V., Duarte, M., Hegde, C.: Model-Based Compressive Sensing. IEEE Transactions on Information Theory 56(4), 1982–2001 (2010)
3. van den Berg, E., Friedlander, M.P.: Probing the pareto frontier for basis pursuit solutions. SIAM Journal on Scientific Computing 31(2), 890–912 (2008)
4. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on information theory 52(2), 489–509 (2006)
5. Chan, A.B., Vasconcelos, N.: Probabilistic kernels for the classification of auto-regressive visual processes. In: IEEE Conf. on Computer Vision and Pattern Recognition. pp. 846–851 (2005)
6. Donoho, D.: Compressed sensing. IEEE Transactions on Information Theory 52(4), 1289–1306 (2006)
7. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. International Journal of Computer Vision 51(2), 91–109 (2003)
8. Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K., Baraniuk, R.: Single-pixel imaging via compressive sampling. IEEE Signal Processing Magazine 25(2), 83–91 (2008)
9. Fowler, J.: Compressive-projection principal component analysis,. IEEE Transactions on Image Processing 18(10) (October 2009)
10. Haupt, J., Nowak, R.: Signal reconstruction from noisy random projections. IEEE Transactions on Information Theory 52(9), 4036–4048 (2006)
11. Needell, D., Tropp, J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis 26(3), 301–321 (2009)
12. Park, J., Wakin, M.: A multiscale framework for compressive sensing of video. In: Picture Coding Symposium. pp. 197–200 (May 2009)
13. Péteri, R., Fazekas, S., Huiskes, M.: DynTex: A Comprehensive Database of Dynamic Textures. (to appear) p. URL: http://projects.cwi.nl/dyntex/ (2010)
14. Saisan, P., Doretto, G., Wu, Y., Soatto, S.: Dynamic texture recognition. In: CVPR. vol. 2, pp. 58–63 (December 2001)
15. Turaga, P., Veeraraghavan, A., Chellappa, R.: Unsupervised view and rate invariant clustering of video sequences. CVIU 113(3), 353–371 (2009)
16. Vaswani, N.: Kalman filtered compressed sensing. In: ICIP (2008)
17. Vaswani, N., Lu, W.: Modified-CS: Modifying compressive sensing for problems with partially known support. In: Intl. Symposium on Information Theory (2009)
18. Veeraraghavan, A., Reddy, D., Raskar, R.: Coded strobing photography: Compressive sensing of high-speed periodic events. TPAMI ((to appear))
19. Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R.: Matching shape sequences in video with applications in human movement analysis. TPAMI 27, 1896–1909 (2005)
20. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. Applied Optics 47(10), 44–51 (2008)
21. Wakin, M., Laska, J., Duarte, M., Baron, D., Sarvotham, S., Takhar, D., Kelly, K., Baraniuk, R.: Compressive imaging for video representation and coding. In: Picture Coding Symposium (April 2006)
22. Willett, R., Nowak, R.: Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging. IEEE Transactions on Medical Imaging 22(3), 332–350 (2003)