Rate-adaptive compressive video acquisition with sliding-window total-variation-minimization reconstruction

Ying Liu and Dimitris A. Pados

Department of Electrical Engineering, State University of New York at Buffalo, Buffalo, NY 14260

ABSTRACT

We consider a compressive video acquisition system where frame blocks are sensed independently. Varying block sparsity is exploited in the form of individual per-block open-loop sampling rate allocation with minimal system overhead. At the decoder, video frames are reconstructed via sliding-window inter-frame total variation minimization. Experimental results demonstrate that such rate-adaptive compressive video acquisition improves noticeably the rate-distortion performance of the video stream over fixed-rate acquisition approaches.

Keywords: Compressed sensing, compressive sampling, dimensionality reduction, total variation minimization, temporal DCT transform, video codecs, video streaming

1. INTRODUCTION

By the Nyquist/Shannon sampling theory, to reconstruct a signal without error the sampling rate must be at least twice as much as the highest frequency of the signal. Compressive sampling (CS), also known as compressed sensing, is an emerging line of work that suggests sub-Nyquist sampling of *sparse signals* of interest [1]-[3]. Rather than collecting an entire Nyquist ensemble of signal samples, CS reconstructs sparse signals from a small number of (random [3] or deterministic [4]) linear measurements via convex optimization [5], linear regression [6],[7], or greedy recovery algorithms [8].

An example of a CS application that has attracted interest is the "single-pixel camera" architecture [9] where a still image can be produced from significantly fewer captured measurements than the number of desired/reconstructed image pixels. An important next-step development is compressive video streaming. In this present work, we consider a video transmission system where the transmitter/encoder performs pure direct compressed sensing acquisition without the benefits of the familiar sophisticated forms of video encoding. This set-up is of interest, for example, in problems that involve large wireless multimedia networks of primitive low-complexity, power-limited video sensors. CS is potentially an enabling technology in this context [10], as video acquisition would require minimal or no computational power at all, yet transmission bandwidth would still be greatly reduced. In such a case, the burden of quality video reconstruction will fall solely on the receiver/decoder side. In comparison, conventional predictive encoding schemes (H.264 [11] or HEVC [12]) are known to offer great transmission bandwidth savings for targeted video quality levels, but place strong computational complexity and power consumption demands on the encoder side.

In compressive video streaming literature, frame partitioning into blocks and block-level encoding has been a common approach [10], [13]-[21]. Decoders rely on the utilization of orthonormal bases on which the video frame blocks can be sparsely represented. To reconstruct the video sequence, the decoder minimizes the ℓ_1-norm of the transform domain coefficients. The number of CS measurements required for quality reconstruction is proportional to the signal sparsity captured by the utilized basis.

Since different areas in a video sequence or video frame may have different sparsity, it is reasonable to consider adaptively allocating the available bandwidth (number of CS measurements for each frame block) based on sparsity level. The key issues are to find an effective metric to quantify the sparsity of frame blocks that

Further author information: (Send correspondence to D.A.P)
Y.L.: E-mail: yl72@buffalo.edu, Telephone: 1 716 645 1207
D.A.P: E-mail: pados@buffalo.edu, Telephone: 1 716 645 1150

Compressive Sensing II, edited by Fauzia Ahmad, Proc. of SPIE Vol. 8717, 871703
 · doi: 10.1117/12.2017751

is easily hardware implementable and to address the problem of how encoder and decoder coordinate to use a *scalable* sensing matrix. In existing measurement allocation literature [16], the encoder iteratively requests groups of measurements through a feedback channel and performs ℓ_1-minimization until the estimated decoded quality is high enough. The use of a feedback channel, however, highly increases the complexity of the video streaming system. In [17], the decoder estimates the sparsity of each block by calculating the variance of the reconstructed coefficients. The encoder, then, allocates the measurement rate based on the variance feedback from the decoder. In addition to the feedback channel requirement again, the variance of the reconstructed coefficients cannot in general accurately represent sparsity, especially when the reconstruction quality of the reference frame is not high.

Moving away from simple frame-block encoding and basis-based decoding, grouped multi-frame encoding and total-variation based [22],[23] recovery that preserves intra-frame sharpness/edges and inter-frame small differences has been demonstrated to have excellent reconstruction performance [24],[25]. Although promising, such a system requires complex and expensive spatial-temporal light modulators that make the technique difficult to be implemented in practice. To tackle the implementation problem, a framewise encoder was proposed recently where each frame is encoded independently using compressive sampling followed by a form of inter-frame TV reconstruction [26]. Such framewise video encoder, however, cannot adopt rate adaptive acquisition for efficient bandwidth utilization.

In the work that we describe in the present paper, we pursue rate adaptive CS acquisition for improved bandwidth efficiency. Yet, for ease in implementation we develop an open-loop (no feedback) system consisting of a simple block-level CS encoder and a block-level sliding-window inter-frame TV-minimization decoder [26]. Experimental studies demonstrate the effectiveness of our proposed system.

The rest of the paper is organized as follows. In Section 2, we review briefly TV-based CS signal recovery principles. Section 3 presents the block-level CS video encoder with fixed and adaptive -most importantly- rate acquisition. In Section 4, the inter-frame block-level sliding-window TV minimizing decoder is described in detail. Some experimental results are presented and analyzed in Section 5 and, finally, a few conclusions are drawn in Section 6.

2. COMPRESSIVE SAMPLING WITH TV MINIMIZATION RECONSTRUCTION

In this section, we briefly review 2D and 3D signal acquisition by compressive sampling and recovery using sparse gradient constraints (TV minimization). If the signal of interest is a 2D image block $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} = vec(\mathbf{X}) \in \mathbb{R}^N$, $N = n^2$, represents vectorization of $\mathbf{X}$ via column concatenation, then CS measurements of $\mathbf{X}$ are collected in the form of

$$\mathbf{y} = \mathbf{\Phi} vec(\mathbf{X}) \tag{1}$$

with a linear measurement matrix $\mathbf{\Phi}_{P\times N}, P \ll N$. Under the assumption that images are mostly pixel-wise smooth in the horizontal and vertical pixel directions, it is natural to consider utilizing the sparsity of the spatial gradient of $\mathbf{X}$ for CS image reconstruction [5],[27]-[31]. If $x_{i,j}$ denotes the pixel in the ith row and jth column of $\mathbf{X}$, the horizontal and vertical gradients at $x_{i,j}$ are defined, respectively, as

$$D_{h;ij}[\mathbf{X}] = \begin{cases} x_{i,j+1} - x_{i,j}, & j < n, \\ 0, & j = n, \end{cases}$$

and

$$D_{v;ij}[\mathbf{X}] = \begin{cases} x_{i+1,j} - x_{i,j}, & i < n, \\ 0, & i = n. \end{cases}$$

The discrete spatial gradient of $\mathbf{X}$ at pixel $x_{i,j}$ can be interpreted as the 2D vector

$$D_{ij}[\mathbf{X}] = \begin{pmatrix} D_{h;ij}[\mathbf{X}] \\ D_{v;ij}[\mathbf{X}] \end{pmatrix} \tag{2}$$

and the anisotropic 2D-TV of $\mathbf{X}$ is simply the sum of the magnitudes of this discrete gradient at every pixel,

$$\mathrm{TV}_{2\mathrm{D}}(\mathbf{X}) \triangleq \sum_{ij} \Big(|D_{h;ij}[\mathbf{X}]| + |D_{v;ij}[\mathbf{X}]| \Big) = \sum_{ij} ||D_{ij}[\mathbf{X}]||_{\ell_1}. \tag{3}$$

To reconstruct $\mathbf{X}$, we can solve the convex program

$$\widehat{\mathbf{X}} = \arg\min_{\widetilde{\mathbf{X}}} \mathrm{TV}_{2\mathrm{D}}(\widetilde{\mathbf{X}}) \qquad \text{subject to} \quad \mathbf{y} = \mathbf{\Phi} vec(\widetilde{\mathbf{X}}). \tag{4}$$

However, in practical situations the measurement vector $\mathbf{y}$ may be corrupted by noise. Then, CS acquisition of $\mathbf{X}$ can be formulated as

$$\mathbf{y} = \mathbf{\Phi} vec(\mathbf{X}) + \mathbf{e} \tag{5}$$

where $\mathbf{e}$ is the unknown noise vector bounded by a presumably known power amount $\|\mathbf{e}\|_{\ell_2} \leq \epsilon$, $\epsilon > 0$. To recover $\mathbf{X}$, we can use 2D-TV minimization as in (4) with a relaxed constraint in the form of

$$\widehat{\mathbf{X}} = \arg\min_{\widetilde{\mathbf{X}}} \mathrm{TV}_{2\mathrm{D}}(\widetilde{\mathbf{X}}) \qquad \text{subject to} \quad \|\mathbf{y} - \mathbf{\Phi} vec(\widetilde{\mathbf{X}})\|_{\ell_2} \leq \epsilon. \tag{6}$$

Moving on now to the needs of the specific CS video work presented in this paper, if the underlying signal is a video signal $\mathbf{X} \in \mathbb{R}^{n \times n \times q}$ representing a stack of q co-located blocks $\mathbf{X}_t \in \mathbb{R}^{n \times n}$, $t = 1, ..., q$, across q successive frames, then concatenating the columns of all $\mathbf{X}_1$, ..., $\mathbf{X}_q$ results to a length $n^2 q$ vector $\mathbf{x} = vec(\mathbf{X})$. If $x_{i,j,t}$ denotes the pixel at the ith row and jth column of frame block $\mathbf{X}_t$, then the horizontal, vertical, and temporal gradient at $x_{i,j,t}$ can be defined, respectively, as

$$D_{h;ij}[\mathbf{X}_t] = \begin{cases} x_{i,j+1,t} - x_{i,j,t}, & j < n, \\ 0, & j = n, \end{cases}$$

$$D_{v;ij}[\mathbf{X}_t] = \begin{cases} x_{i+1,j,t} - x_{i,j,t}, & i < n, \\ 0, & i = n, \end{cases}$$

and

$$D_{t;ij}[\mathbf{X}_t] = \begin{cases} x_{i,j,t+1} - x_{i,j,t}, & t < q, \\ x_{i,j,1} - x_{i,j,t}, & t = q. \end{cases}$$

Correspondingly, the spatial-temporal gradient of $\mathbf{X}$ at $x_{i,j,t}$ can be interpreted as the 3D vector

$$D_{ij}[\mathbf{X}_t] = \begin{pmatrix} D_{h;ij}[\mathbf{X}_t] \\ D_{v;ij}[\mathbf{X}_t] \\ D_{t;ij}[\mathbf{X}_t] \end{pmatrix} \tag{7}$$

and the anisotropic 3D-TV of $\mathbf{X}$ is simply the sum of the magnitudes of this discrete gradient at every pixel

$$\mathrm{TV}_{3\mathrm{D}}(\mathbf{X}) \triangleq \sum_{i,j,t} \Big(|D_{h;ij}[\mathbf{X}_t]| + |D_{v;ij}[\mathbf{X}_t]| + |D_{t;ij}[\mathbf{X}_t]| \Big) = \sum_{i,j,t} ||D_{ij}[\mathbf{X}_t]||_{\ell_1}. \tag{8}$$

To reconstruct the block sequence $\mathbf{X} \in \mathbb{R}^{n \times n \times q}$ from noiseless measurements, we can solve the convex program

$$\widehat{\mathbf{X}} = \arg \min_{\widetilde{\mathbf{X}} \in \mathbb{R}^{n \times n \times q}} \mathrm{TV}_{3\mathrm{D}}(\widetilde{\mathbf{X}}) \qquad \text{subject to} \quad \mathbf{y} = \mathbf{\Phi} vec(\widetilde{\mathbf{X}}). \tag{9}$$

The reconstruction of $\mathbf{X} \in \mathbb{R}^{n \times n \times q}$ from noisy measurements can be formulated as the 3D-TV decoding

$$\widehat{\mathbf{X}} = \arg \min_{\widetilde{\mathbf{X}} \in \mathbb{R}^{n \times n \times q}} \mathrm{TV}_{3\mathrm{D}}(\widetilde{\mathbf{X}}) \qquad \text{subject to} \quad \|\mathbf{y} - \mathbf{\Phi} vec(\widetilde{\mathbf{X}})\|_{\ell_2} \leq \epsilon. \tag{10}$$

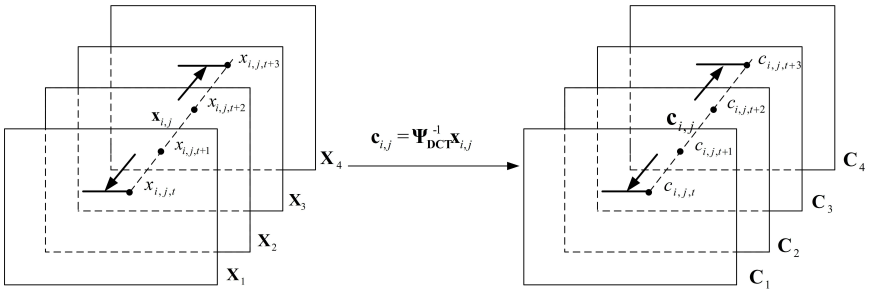

Figure 1. Illustration of pixelwise temporal DCT ($q = 4$).

If the individual blocks $\mathbf{X}_1$, ..., $\mathbf{X}_q$ in $\mathbf{X}$ are highly time-correlated, then a pixelwise temporal DCT generally improves sparsity. As illustrated in Fig. 1, each temporal-length q ($q = 4$ for example) vector $\mathbf{x}_{i,j} = [x_{i,j,1}, ..., x_{i,j,q}]^{\mathrm{T}}$, $i = 1, ..., n$, $j = 1, ..., n$, consisting of the pixels at spatial position (i, j) across q successive co-located blocks, can be represented as

$$\mathbf{x}_{i,j} = \mathbf{\Psi}_{\mathrm{DCT}}\mathbf{c}_{i,j} \tag{11}$$

where $\mathbf{\Psi}_{\mathrm{DCT}}$ is the 1D-DCT basis and $\mathbf{c}_{i,j}$ is the transform-domain coefficient vector. The resulting coefficient matrix $\mathbf{C}_1$ represents the frequency component that remains unchanged over time (dc) and the subsequent coefficient matrices $\mathbf{C}_t$, $t = 2, ..., q$, represent frequency components of increasing time variability. Since each matrix $\mathbf{C}_t$, $t = 1, ..., q$, is expected to have small TV, they can be jointly recovered in the form of

$$\widehat{\mathbf{C}}_1, ..., \widehat{\mathbf{C}}_q = \arg \min_{\widetilde{\mathbf{C}}_1, ..., \widetilde{\mathbf{C}}_q} \sum_{t=1}^{q} \mathrm{TV}_{2\mathrm{D}}(\widetilde{\mathbf{C}}_t) \qquad \text{subject to} \quad \|\mathbf{y} - \mathbf{\Phi} vec(\mathrm{DCT}^{-1}(\widetilde{\mathbf{C}}_1, ..., \widetilde{\mathbf{C}}_q))\|_{\ell_2} \leq \epsilon \tag{12}$$

where $\mathrm{DCT}^{-1}(\widetilde{\mathbf{C}}_1, ..., \widetilde{\mathbf{C}}_q)$ stands for pixelwise inverse 1D-DCT. Subsequently, the complete block sequence $\mathbf{X} \in \mathbb{R}^{n \times n \times q}$ can be reconstructed simply as

$$\widehat{\mathbf{X}} = \mathrm{DCT}^{-1}(\widehat{\mathbf{C}}_1, ..., \widehat{\mathbf{C}}_q). \tag{13}$$

In the sequel, we will refer to this form of inter-frame CS reconstruction as *TV-DCT decoding*.

3. PROPOSED CS VIDEO ENCODER

3.1 Block-level CS Video Encoder with Fixed Rate Acquisition

In the simple compressive video encoding block diagram shown in Fig. 2, each frame $\mathbf{F}_t$, $t = 1, 2, ...$, is virtually partitioned into M non-overlapping blocks of pixels with each block $\mathbf{X}_t^m$ viewed as a vectorized column of length N, $\mathbf{x}_t^m \in \mathbb{R}^N$, $m = 1, ..., M$, $t = 1, 2, ...$ Compressive sampling is performed by projecting $\mathbf{x}_t^m$ onto a $P \times N$ random measurement matrix $\mathbf{\Phi}_t^m$

$$\mathbf{y}_t^m = \mathbf{\Phi}_t^m \mathbf{x}_t^m \tag{14}$$

where $\mathbf{\Phi}_t^m$, $m = 1, ..., M$, $t = 1, 2, ..., T$, is generated by randomly permuting the columns of an order-k, $k \geq N$ and multiple-of-four, Walsh-Hadamard (WH) matrix followed by arbitrary selection of P rows from the k available WH rows (if $k > N$, only N arbitrary columns are utilized). This class of WH measurement matrices has the advantage of easy implementation (antipodal ± 1 entries), fast transformation, and satisfactory reconstruction performance as we will see later on. A richer class of matrices can be found in [32],[33]. To quantize the elements of the resulting measurement vector $\mathbf{y}_t^m \in \mathbb{R}^P$ (block $\mathbf{Q}$ in Fig. 2), in this work we follow a simple adaptive quantization approach of two codeword lengths. A positive threshold $\eta > 0$ is chosen such that 1% of the elements in $\mathbf{y}_1^1$ have magnitude above η. For every measurement vector $\mathbf{y}_t^m$, $m = 1, ..., M$, $t = 1, 2, ...$, 16-bit uniform scalar quantization is used for elements with magnitudes larger than η and 8-bit uniform scalar quantization is used for the remaining elements. The resulting quantized values $\widetilde{\mathbf{y}}_t^m$ are then indexed and transmitted to the decoder.

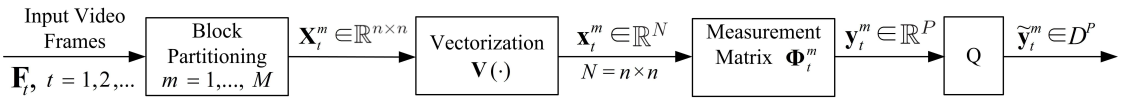

Figure 2. A simple block-level compressed sensing (CS) video encoder system with quantization alphabet $\mathcal{D}$.

3.2 Block-level CS Video Encoder with Adaptive Rate Acquisition

To exploit different sparsity levels at different regions of the video sequence, we now propose adaptive rate acquisition for our block-level CS video encoder. The entire video sequence is viewed as a sequence of groups of q successive frames. Adaptive rate acquisition is carried out within each group of frames. As illustrated in Fig. 3, after block partitioning, the q co-located blocks in the same group $d = 1, q+1, 2q+1, ...$ are concatenated to form a cube $\mathbf{X}_{d:d+q-1}^m$. Then, encoder-only-based (no feedback) measurement rate allocation is performed as shown in Fig. 4. Let the total number of CS measurements be P_{total} per group of frames. Each block in data cube $\mathbf{X}_{d:d+q-1}^m$ receives number of measurements

$$P_t^m = P_{\text{total}} \times \frac{(\text{TV}_{3\text{D}}(\mathbf{X}_{d:d+q-1}^m))^\alpha}{\sum_{m=1}^{M} (\text{TV}_{3\text{D}}(\mathbf{X}_{d:d+q-1}^m))^\alpha} \times \frac{1}{q} \tag{15}$$

where α is a design constant between 0 and 1. For implementation, sensing matrices $\mathbf{\Phi}_t^m \in \mathbb{R}^{P_t^m \times N}$, $N = n^2$, are generated by WH methods as described before and CS encoding of block $\mathbf{x}_t^m$ in $\mathbf{X}_{d:d+q-1}^m$, $t = d, ..., d+q-1$, is carried out by

$$\mathbf{y}_t^m = \mathbf{\Phi}_t^m \mathbf{x}_t^m. \tag{16}$$

Afterwards, each measurement vector $\mathbf{y}_t^m \in \mathbb{R}^{P_t^m}$ is quantized, indexed and transmitted to the decoder as per Fig. 2.

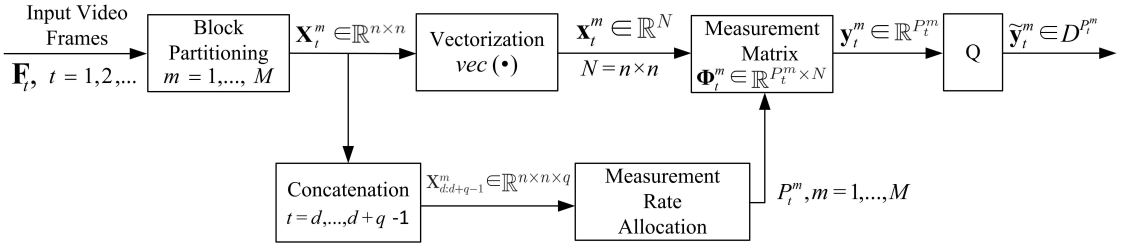

Figure 3. A block-level adaptive-rate compressed sensing (CS) video encoder system with quantization alphabet $\mathcal{D}$.

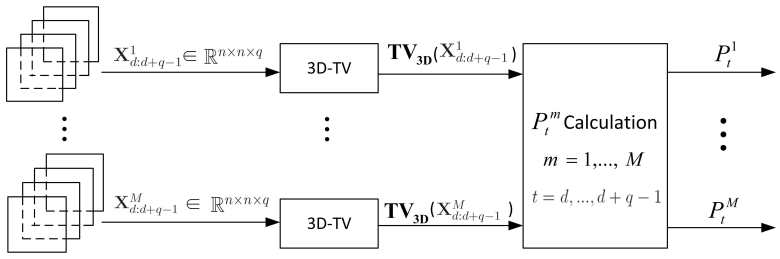

Figure 4. Illustration of measurement rate allocation.

4. PROPOSED CS VIDEO DECODER

To reconstruct the independently encoded CS video frames, a simplistic idea is to recovery each frame block independently via 2D-TV decoding by eq. (4). However, such a decoding scheme does not exploit the inter-frame similarities of a video sequence. We propose, instead, to jointly recover *individually* encoded blocks within the same cube $\mathbf{X}^m_{d:d+q-1}$ via inter-frame block-level TV minimization.

As shown in Fig. 5, to jointly recover the blocks in the m^{th} cube $\mathbf{X}^m_{d:d+q-1}$, the proposed interframe CS video decoder collects and concatenates q dequantized measurement vectors $\widehat{\mathbf{y}}^m_t \in \mathbb{R}^{P^m_t}$, $t = d, ..., d+q-1$, to create vector $\widehat{\mathbf{y}}^m_{d:d+q-1} \in \mathbb{R}^{\sum_{t=d}^{d+q-1} P^m_t}$. Because each dequantized vector is of the form of $\widehat{\mathbf{y}}^m_t = \mathbf{\Phi}^m_t \mathbf{x}^m_t + \mathbf{e}^m_t$ with noise $\mathbf{e}^m_t$, $\widehat{\mathbf{y}}^m_{d:d+q-1}$ can be represented as

$$\widehat{\mathbf{y}}^m_{d:d+q-1} = \widetilde{\mathbf{\Phi}}^m_{d:d+q-1} \mathbf{x}^m_{d:d+q-1} + \mathbf{e}^m_{d:d+q-1} \tag{17}$$

where $\widetilde{\mathbf{\Phi}}^m_{d:d+q-1} \in \mathbb{R}^{\sum_{t=d}^{d+q-1} P^m_t \times (qN)}$ is the block diagonal matrix

$$\widetilde{\mathbf{\Phi}}^m_{d:d+q-1} = \begin{pmatrix} \mathbf{\Phi}^m_d & & & \\ & \mathbf{\Phi}^m_{d+1} & & \\ & & \ddots & \\ & & & \mathbf{\Phi}^m_{d+q-1} \end{pmatrix}, \tag{18}$$

$\mathbf{x}^m_{d:d+q-1}$ is the concatenation of q vectorized blocks

$$\mathbf{x}^m_{d:d+q-1} = [\mathbf{x}^m_d; \quad \mathbf{x}^m_{d+1}; \quad \dots \quad \mathbf{x}^m_{d+q-1}], \tag{19}$$

and $\mathbf{e}^m_{d:d+q-1}$ is the concatenation of the noise vectors in the form of

$$\mathbf{e}^m_{d:d+q-1} = [\mathbf{e}^m_d; \quad \mathbf{e}^m_{d+1}; \quad \dots \quad \mathbf{e}^m_{d+q-1}]. \tag{20}$$

The decoder then performs 3D-TV decoding on the q blocks (Fig. 5(a)) by

$$\widehat{\mathbf{X}}^m_{d:d+q-1} = \arg\min_{\widetilde{\mathbf{X}}} \mathrm{TV}_{3\mathrm{D}}(\widetilde{\mathbf{X}}) \qquad \text{subject to} \quad \|\widehat{\mathbf{y}}^m_{d:d+q-1} - \widetilde{\mathbf{\Phi}}^m_{d:d+q-1} \mathbf{V}(\widetilde{\mathbf{X}})\|_{\ell_2} \leq \epsilon. \tag{21}$$

Although (21) may be considered a powerful joint 3D-TV recovery procedure for general 2D CS-acquired video, for highly temporally correlated video frames, better reconstruction quality may be achieved via TV-temporal-DCT decoding (Fig. 5(b)) in the form of

$$\widehat{\mathbf{C}}^m_d, ..., \widehat{\mathbf{C}}^m_{d+q-1} = \arg \min_{\widetilde{\mathbf{C}}^m_d, ..., \widetilde{\mathbf{C}}^m_{d+q-1}} \sum_{t=d}^{d+q-1} \mathrm{TV}_{2\mathrm{D}}(\widetilde{\mathbf{C}}^m_t)$$

$$\text{subject to} \quad \|\widehat{\mathbf{y}}^m_{d:d+q-1} - \widetilde{\mathbf{\Phi}}^m_{d:d+q-1} vec(\mathrm{DCT}^{-1}(\widetilde{\mathbf{C}}^m_d, ..., \widetilde{\mathbf{C}}^m_{d+q-1}))\|_{\ell_2} \leq \epsilon. \tag{22}$$

$\mathbf{X}^m_{d:d+q-1}$ can then be reconstructed simply by

$$\widehat{\mathbf{X}}^m_{d:d+q-1} = \mathrm{DCT}^{-1}(\widehat{\mathbf{C}}^m_d, ..., \widehat{\mathbf{C}}^m_{d+q-1}). \tag{23}$$

In (22), (23), we carried out inter-frame block level decoding for each independent group of q blocks $\mathbf{X}^m_{d:d+q-1}$, $d = 1, q+1, 2q+1, 3q+1, ...$ To further exploit inter-frame similarities and capture local motion among adjacent groups of co-located blocks, we now propose a sliding-window TV-DCT decoder. The concept of such a decoder is depicted in Fig. 6. Initially, the decoder performs TV-DCT decoding on the first q ($q = 4$, for example) blocks, $\mathbf{X}^m_1$, ..., $\mathbf{X}^m_q$, specified by a decoding window of length q (Fig. 6(a)) using the block diagonal matrix $\widetilde{\mathbf{\Phi}}^m_{1:q}$ with

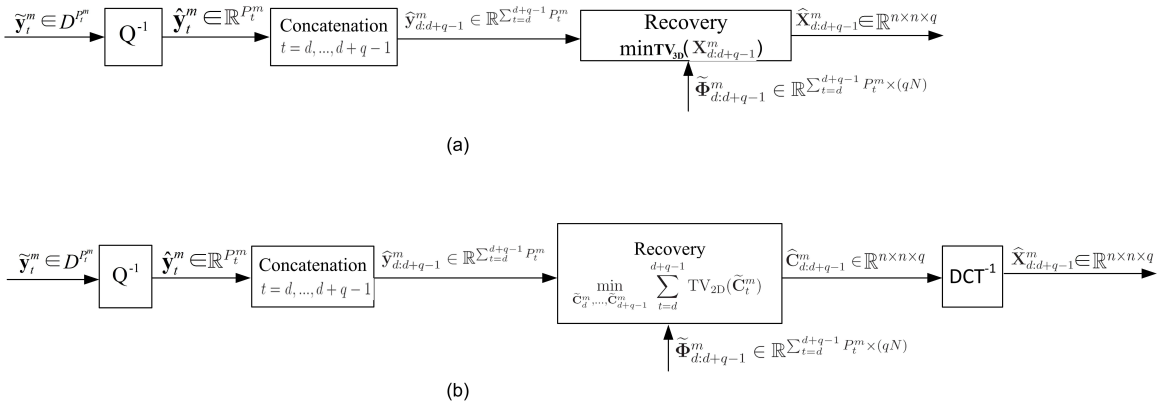

Figure 5. (a) Proposed 3-D total variation (TV). (b) TV-DCT CS decoder on individually encoded video blocks.

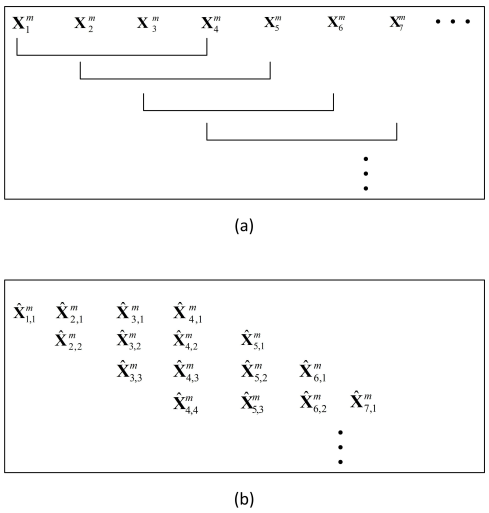

Figure 6. Proposed sliding-window TV-DCT CS decoder system.

diagonal elements $\mathbf{\Phi}_1^m$, ..., $\mathbf{\Phi}_q^m$. The reconstructed blocks are called $\widehat{\mathbf{X}}_{1,1}^m$, $\widehat{\mathbf{X}}_{2,1}^m$, ..., $\widehat{\mathbf{X}}_{q,1}^m$ (Fig. 6(b)) where $\widehat{\mathbf{X}}_{t,l}^m$ represents the l^{th} reconstruction of the m^{th} block in frame t. Then, the decoding window shifts one frame to the right, performs TV-DCT decoding on $\mathbf{X}_2^m$, ..., $\mathbf{X}_{q+1}^m$ using the matrix $\widetilde{\mathbf{\Phi}}_{2:q+1}^m$ with diagonal elements $\mathbf{\Phi}_2^m$,..., $\mathbf{\Phi}_{q+1}^m$, and produces the reconstructed blocks $\widehat{\mathbf{X}}_{2,2}^m$, $\widehat{\mathbf{X}}_{3,2}^m$,..., $\widehat{\mathbf{X}}_{q,2}^m$, $\widehat{\mathbf{X}}_{q+1,1}^m$. The decoder continues on with sliding-window TV-DCT decoding until the last group of blocks $\mathbf{X}_{T-q+1}^m$,..., $\mathbf{X}_T^m$ is recovered. Final reconstruction of each block $\widehat{\mathbf{X}}_t^m$ is executed by taking the average of all different decodings by

$$\widehat{\mathbf{X}}_t^m = \begin{cases} \frac{1}{t}\sum_{l=1}^{t} \widehat{\mathbf{X}}_{t,l}^m, & 1 \le t \le q, \\ \frac{1}{q}\sum_{l=1}^{q} \widehat{\mathbf{X}}_{t,l}^m, & q \le t \le T-q+1, \\ \frac{1}{T-t+1}\sum_{l=1}^{T-t+1} \widehat{\mathbf{X}}_{t,l}^m, & T-q+2 \le t \le T. \end{cases} \tag{24}$$

Compared to the simple (non-sliding-window) TV-DCT decoder of (22), (23), the sliding-window TV-DCT decoder enforces sparsity for *any successive* q co-located blocks in the video sequence. Hence, it protects sharp

Table 1. Empirical q values for Container

average $\frac{P_t^m}{N}$	0.125	0.25	0.375	0.5	0.625
fixed $\mathbf{\Phi}_t^m$	20	20	20	20	20
varying $\mathbf{\Phi}_t^m$	2	4	20	20	20

Table 2. Empirical q values for Highway

average $\frac{P_t^m}{N}$	0.125	0.25	0.375	0.5	0.625
adaptive rate sliding-window TV-DCT	4	4	4	20	20
adaptive rate 3D-TV	20	20	20	20	20
fixed rate sliding-window TV-DCT	4	4	4	20	20
fixed rate 3D-TV	20	20	20	20	20

temporal changes for pixels that have fast motion in any q-frame-sequence and smoothes intensities for static or slow motion pixels in the same decoding window.

5. EXPERIMENTAL RESULTS

In this section, we study experimentally the performance of the developed CS video systems by evaluating the peak-signal-to-noise ratio (PSNR) (as well as the perceptual quality) of reconstructed video sequences. Two test sequences, Container and Highway, with CIF resolution 352×288 pixels and frame rate of 30 frames/sec are used. Processing is carried out only on the luminance component.

At our CS encoder side, each frame is partitioned into non-overlapping blocks of 32×32 pixels. The frame group size q for adaptive rate allocation, in accordance with the decoding window size, is set empirically to the values shown in Table 1 and Table 2. After adaptive rate allocation, each block $\mathbf{x}_t^m$ is handled as a vectorized column of length $N = 1024$ multiplied by a $P_t^m \times N$ randomized partial WH matrix $\mathbf{\Phi}_t^m$. In our experiments, average $P_t^m = 128, 256, 384, 512, 640$ is used to produce the corresponding bit rates of $3071.7, 6143.4, 9215.1, 12287,$ and 15358 kbps*. It has been demonstrated [26] that for slow motion sequences (Container), varying diagonal elements of $\widetilde{\mathbf{\Phi}}_{d:d+q-1}^m$ in (18) enhances the performance of inter-frame TV minimization reconstruction. For fast motion sequences (Highway), fixed diagonal elements of $\widetilde{\mathbf{\Phi}}_{d:d+q-1}^m$ offers better performance. In the block-level CS video decoder presented in this work, each video cube of size $n \times n \times q$ is decoded independently, which implies that for the Container sequence $\mathbf{\Phi}_t^m$ needs to be varying within each cube only. Therefore, we propose to start

*Considering the quantization scheme described in Section III and frame rate 30 fps, the bit rate can be calculated as $(16 \times 0.01P + 8 \times 0.99P)$ bits per block $\times 99$ blocks per frame $\times 30/1000$ kbps.

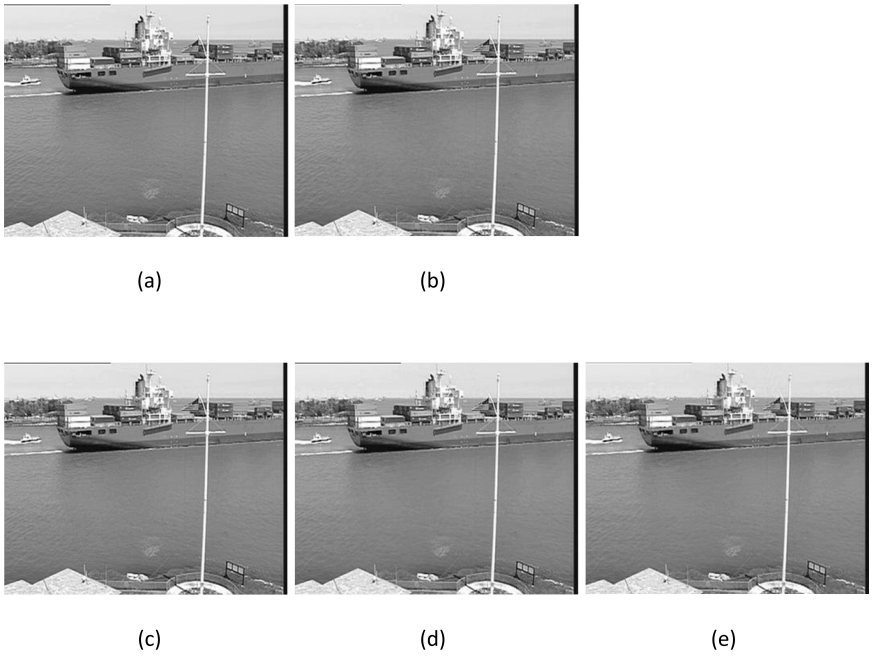

Figure 7. Different sliding-window TV-DCT decodings of the $28th$ frame of Container (average $\frac{P_t^m}{N} = 0.625$): (a) Original frame; (b) adaptive rate acquisition with varying $\mathbf{\Phi}_t^m$ ($q = 20$); (c) fixed rate acquisition with varying $\mathbf{\Phi}_t^m$ ($q = 20$); (d) adaptive rate acquisition with fixed $\mathbf{\Phi}_t^m$ ($q = 20$); and (e) fixed rate acquisition with fixed $\mathbf{\Phi}_t^m$ ($q = 20$).

with an order-N WH matrix, followed by permutating independently its columns and rows q times and storing the permutation indices at both the encoder and decoder. When encoding blocks $\mathbf{x}_t^m$, $t = d, ..., d+q-1$, in the cube, the columns of the original WH matrix are permutated based on the $(t-d+1)^{th}$ column permutation indices, followed by selecting the rows indicated by the first P_t^m indices in the $(t-d+1)^{th}$ row permutation to form a sensing matrix $\mathbf{\Phi}_t^m$. For the Highway sequence, we generate only once the column and row permutations for an order-N WH matrix and store them at both the encoder and decoder. When encoding each block $\mathbf{x}_t^m$, $t = d, ..., d+q-1$, in the cube, the columns of the original WH matrix are permutated as indicated by the column permutation indices. Then, the rows indicated by the first P_t^m indices in row permutation are chosen from the resulting matrix to form a sensing matrix $\mathbf{\Phi}_t^m$. The elements of each captured P_t^m-dimensional measurement vector are quantized and then transmitted to the decoder. In our simulation, the largest cube size is $32 \times 32 \times 20$, which requires $q = 20$ column and row permutations of an order-N ($N = 1024$) WH matrix to generate a varying sensing matrix $\mathbf{\Phi}_t^m$ for the whole Container sequence resulting to 50k bytes[†] storage memory at both encoder and decoder. To encode/decode the Highway sequence with a fixed sensing matrix, the storage requirement is reduced to 2.5k bytes.

At the decoder side, we choose the TVAL3 software [24],[25] for reconstruction motivated by its low-complexity and satisfactory recovery performance characteristics. In our experimental studies for the slow-motion Container sequence, four CS video systems are examined: (i) varying $\mathbf{\Phi}_t^m$ adaptive rate acquisition with sliding-window TV-DCT decoding; (ii) varying $\mathbf{\Phi}_t^m$ fixed rate acquisition with sliding-window TV-DCT decoding; (iii) fixed $\mathbf{\Phi}_t^m$ adaptive rate acquisition with sliding-window TV-DCT decoding; and (iv) fixed $\mathbf{\Phi}_t^m$ fixed rate acquisition with sliding-window TV-DCT decoding. For the fast-motion Highway sequence, we show results with fixed $\mathbf{\Phi}_t^m$ for CS acquisition and (i) adaptive rate acquisition with sliding-window TV-DCT decoding; (ii) adaptive rate acquisition with 3D-TV decoding; (iii) fixed rate acquisition with sliding-window TV-DCT decoding; and (iv) fixed rate acquisition with 3D-TV decoding.

[†]The bits required to store the indices of one row/column permutation for an order-N WH matrix is calculated by $(\log_2 N)$ bits per index $\times$ N indices.

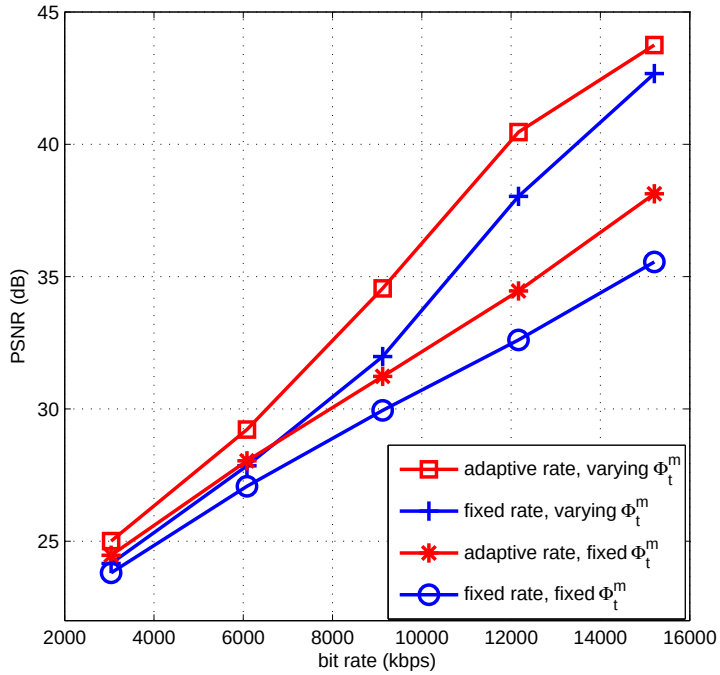

Figure 8. Rate-distortion studies on the Container sequence (sliding-window TV-DCT decoding).

Fig. 7 shows the sliding-window TV-DCT decodings with window size $q = 20$ of the $28th$ frame of Container produced by adaptive rate acquisition with varying $\mathbf{\Phi}_t^m$ (Fig. 7(b)), fixed rate acquisition with varying $\mathbf{\Phi}_t^m$ (Fig. 7(c)), adaptive rate acquisition with fixed $\mathbf{\Phi}_t^m$ (Fig. 7(d)), and fixed rate acquisition with fixed $\mathbf{\Phi}_t^m$ (Fig. 7(e)). It can be observed, in Fig. 8, that for both varying $\mathbf{\Phi}_t^m$ and fixed $\mathbf{\Phi}_t^m$, adaptive rate acquisition demonstrates considerable reconstruction quality improvement[‡] compared to its fixed rate acquisition counterpart. The demonstrated superiority of varying $\mathbf{\Phi}_t^m$ is consistent with the belief that varying $\mathbf{\Phi}_t^m$, $t = d, ..., d+q-1$, in (16) results in a joint block-diagonal recovery matrix $\widetilde{\mathbf{\Phi}}_{d:d+q-1}^m$ that is more likely to satisfy the restricted-isometry-property (RIP) [3] for a given data sparsity level. Adaptive rate acquisition with varying $\mathbf{\Phi}_t^m$ TV-DCT decoding outperforms fixed rate acquisition with varying $\mathbf{\Phi}_t^m$ TV-DCT decoding for all P values, with gains as much as 2.5dB at the median bit rate range. For fixed $\mathbf{\Phi}_t^m$ TV-DCT decoding, adaptive rate acquisition performs better than fixed rate acquisition as well, with gains as much as 2.5dB at the high bit rate range. All PSNR values are averages over the 100 frames of the video sequence.

For the Highway sequence with always fixed $\mathbf{\Phi}_t^m$ and window size $q = 20$, Fig. 9 shows the decodings of the $54th$ frame produced by adaptive rate acquisition and sliding-window TV-DCT decoding (Fig. 9(b)), adaptive rate acquisition and 3D-TV decoding (Fig. 9(c)), fixed-rate acquisition and sliding-window TV-DCT decoding (Fig. 9(d)), and fixed-rate acquisition and 3D-TV decoding (Fig. 9(e)). By Fig. 10, the proposed adaptive-rate acquisition and sliding-window TV-DCT decoding system outperforms adaptive-rate acquisition with 3D-TV decoding and the two fixed-rate acquisition CS video systems.

[‡]As usual, pdf formatting of the present article tends to dampen perceptual quality differences between Figs. 7(a)-(e) that are quite pronounced in video playback. Fig. 8 is the usual attempt to capture average differences quantitatively.

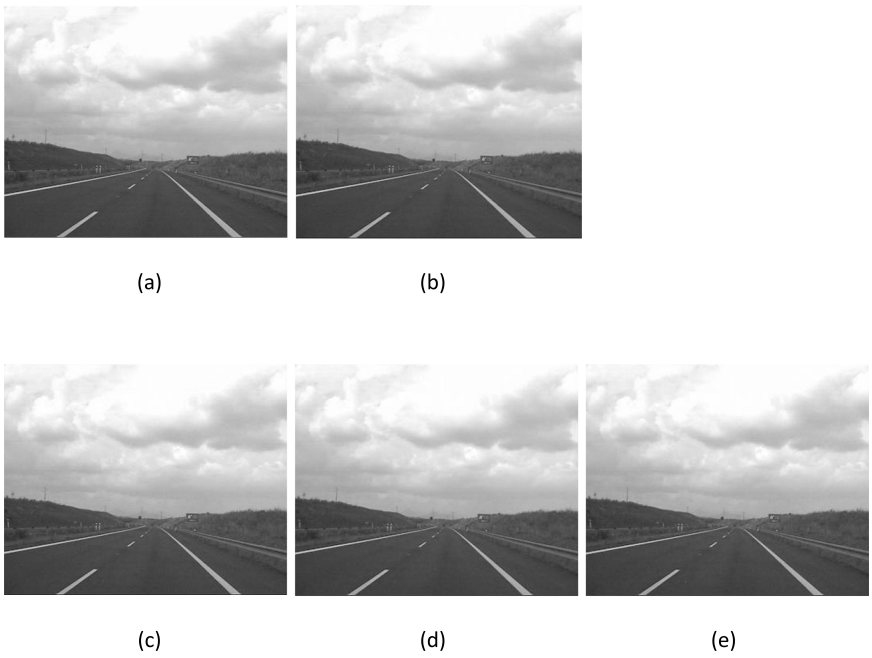

Figure 9. Different fixed Φ_t^m decodings of the $54th$ frame of Highway (average $\frac{P_t^m}{N} = 0.625$): (a) Original frame; (b) adaptive rate acquisition with sliding-window TV-DCT decoder ($q = 20$); (c) adaptive rate acquisition with 3D-TV decoder ($q = 20$); (d) fixed rate acquisition with sliding-window TV-DCT decoder ($q = 20$); and (e) fixed rate acquisition with sliding-window TV-DCT decoder ($q = 20$).

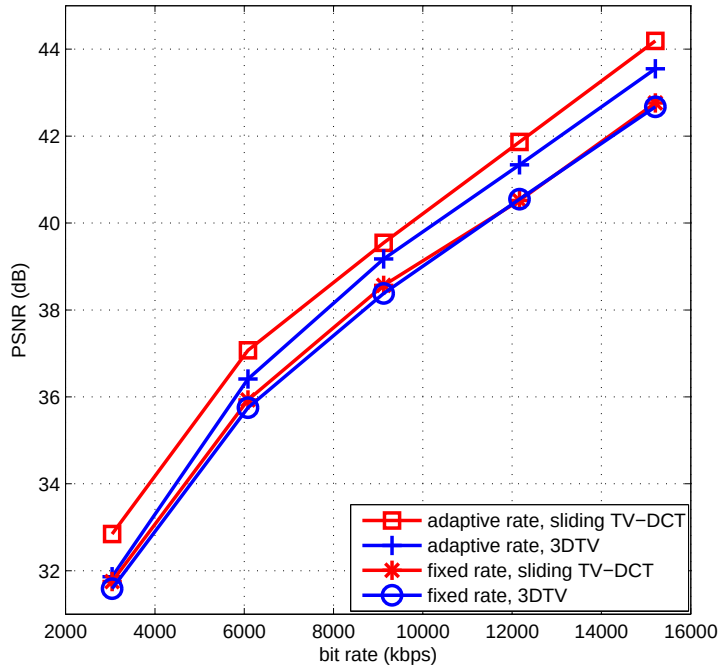

Figure 10. Rate-distortion studies on the Highway sequence.

6. CONCLUSIONS

We proposed a block-level adaptive-rate CS acquisition system for video streaming with inter-frame sliding-window TV minimization decoding. In particular, to exploit the different sparsity levels of different regions of the video sequence, the encoder adaptively allocates the number of CS measurements to different video cubes based on their encoder calculated 3D-TV (open loop system). At the decoder side, to exploit the small differences among successive frames, inter-frame TV minimization is carried out on each cube for video reconstruction. Experimental results demonstrate that the proposed adaptive rate CS video acquisition with sliding window TV-DCT decoding outperforms significantly fixed-rate CS video systems. In terms of future work, to further reduce decoder complexity and improve video reconstruction quality, one may seek other effective and efficient recovery algorithms together with measurement matrices of deterministic design at the encoder to facilitate efficient encoding/decoding.

REFERENCES

[1] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406-5425, Dec. 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289-1306, Apr. 2006.

[3] E. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Proc. Magazine*, vol. 25, pp. 21-30, Mar. 2008.

[4] K. Gao, S. N. Batalama, D. A. Pados, and B. W. Suter, "Compressive sampling with generalized polygons," *IEEE Trans. Signal Proc.*, vol. 59, pp. 4759-4766, Oct. 2011.

[5] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure and Applied Math.*, vol. 59, pp. 1207-1223, Aug. 2006.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, vol. 58, pp. 267-288, 1996.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407-451, Apr. 2004.

[8] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, pp. 4655-4666, Dec. 2007.

[9] 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 Proc. Magazine*, vol. 25, pp. 83-91, Mar. 2008.

[10] S. Pudlewski, T. Melodia, and A. Prasanna, "Compressed-sensing-enabled video streaming for wireless multimedia sensor networks," *IEEE Trans. Mobile Comp.*, vol. 11, pp. 1060-1072, June 2011.

[11] I. E. Richardson, *The H.264 Advanced Video Compression Standard*. New York, NY: Wiley, 2010, 2nd Ed.

[12] G.J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, pp. 1649-1668, Dec. 2012.

[13] V. Stankovic, L. Stankovic, and S. Cheng, "Compressive video sampling," in *Proc. European Signal Proc. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[14] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk, "Compressive imaging for video representation and coding," in *Proc. Picture Coding Symposium (PCS)*, Beijing, China, Apr. 2006.

[15] R. F. Marcia and R. M. Willet, "Compressive coded aperture video reconstruction," in *Proc. European Signal Proc. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[16] J. Prades-Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," in *Proc. Picture Coding Symposium (PCS)*, Chicago, IL, May 2009.

[17] H. W. Chen, L. W. Kang, and C. S. Lu, "Dynamic measurement rate allocation for distributed compressive video sensing," in *Proc. Visual Comm. and Image Proc. (VCIP)*, Huang Shan, China, July 2010.

[18] J. Y. Park and M. B. Wakin, "A multiscale framework for compressive sensing of video," in *Proc. Picture Coding Symposium (PCS)*, Chicago, IL, May 2009.

[19] Y. Liu, M. Li, K. Gao, and D. A. Pados, "Motion compensation as sparsity-aware decoding in compressive video streaming," in *Proc. 17th Intern. Conf. on Digital Signal Processing (DSP 2011)*, Corfu, Greece, July, 2011, pp. 1-5.

[20] Y. Liu, M. Li, and D. A. Pados, "Decoding of purely compressed-sensed video," in *Proc. SPIE, Compressive Sensing Conf., SPIE Defense, Security, and Sensing*, Baltimore, MD, Apr., 2012, vol. 8365.

[21] Y. Liu, M. Li, and D. A. Pados, "Motion-aware decoding of compressed-sensed video," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 23, pp. 438-444, May 2012.

[22] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, pp. 259-268, 1992.

[23] J. Yang, Y. Zhang, and W. Yin, "An efficient TVL1 algorithm for deblurring of multichannel images corrupted by impulsive noise," *SIAM J. Sci. Comput.*, vol. 31, pp. 2842-2865, 2009.

[24] C. Li, H. Jiang, P. Wilford, and Y. Zhang, "Video coding using compressive sensing for wireless communications," in *Proc. IEEE Wireless Communications & Networking Conf. (WCNC)* Cancun, Mexico, Mar. 2011, pp. 2077-2082.

[25] H. Jiang, C. Li, R. Haimi-Cohen, P. Wilford, and Y. Zhang, "Scalable video coding using compressive sensing," *Bell Labs Technical Journal*, vol. 16, pp. 149-169, Mar. 2012.

[26] Y. Liu and D. A. Pados, "Decoding of framewise compressed-sensed video via interframe total variation minimization" *SPIE Journal of Electronic Imaging, Special Issue on Compressive Sensing for Imaging*, Apr.-June 2013.

[27] E. Candès and J. Romberg, "ℓ1-magic: Recovery of sparse signals via convex programming," URL: www.acm.caltech.edu/l1magic/downloads/l1magic.pdf.

[28] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging," *Magn. Reson. Med.*, vol. 6, pp. 1182-95, Dec. 2007.

[29] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed MR imaging using total variation and wavelets," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1-8.

[30] C. Li, "An efficient algorithm for total variation regularization with applications to the sigle pixel camera and compressive sensing," *Master Thesis*, Dept. of Computational and Applied Mathematics, Rice University, 2009.

[31] M. R. Dadkhah, S. Shirani, and M. J. Deen, "Compressive sensing with modified total variation minimization algorithm," in *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 1310-1313.

[32] H. Ganapathy, D. A. Pados, and G. N. Karystinos, "New bounds and optimal binary signature sets - Part I: Periodic total squared correlation," *IEEE Trans. Comm.*, vol. 59, pp. 1123-1132, Apr. 2011.

[33] H. Ganapathy, D. A. Pados, and G. N. Karystinos, "New bounds and optimal binary signature sets - Part II: Aperiodic total squared correlation," *IEEE Trans. Comm.*, vol. 59, pp. 1411-1420, May 2011.