Tutorial: Compressive Sensing of Video

Venue: CVPR 2012, Providence, Rhode Island, USA

June 16, 2012

Richard Baraniuk
Mohit Gupta
Aswin Sankaranarayanan
Ashok Veeraraghavan
Tutorial: Compressive Sensing of Video

Unit 3 – Compressive Video Sensing Systems

Ashok Veeraraghavan
RECAP: Compressive Sensing

- Signal recovery via $L_1$ optimization
  
  $\hat{x} = \arg\min_{y=\Phi x} \|x\|_1$

$M \times 1$ random measurements

$y$

$\Phi$

$N \times 1$ sparse signal

$K$ nonzero entries

$M \times N$

[Candes, Romberg, Tao; Donoho]
RECAP: Compressive Sensing

- Signal **recovery** via iterative greedy algorithm
  - (orthogonal) matching pursuit  [Gilbert, Tropp]
  - iterated thresholding  
    [Nowak, Figueiredo; Kingsbury, Reeves; Daubechies, Defrise, De Mol; Blumensath, Davies; ...]
  - CoSaMP  [Needell and Tropp]
The Plan for the next hour

• Model Based Compressive Video Sensing
  – Coded Strobing for Periodic signals
  – CS-Linear Dynamical Systems
  – Linear Motion Model

• Compressive Sensing of Generic Videos
  – Single Pixel Video Camera
  – Voxel Subsampling
  – Voxel Multiplexing
Model Based CS Video

- Periodic or Repetitive Videos
  - CS Strobing

- Linear Dynamical Systems
  - CS – LDS

- Linear Motion Model
  - Multi-Exposure Video
Coded Strobing Camera

Ashok Veeraraghavan
Dikpal Reddy
Ramesh Raskar
PAMI 2010
Periodic Visual Signals
Strobing: Applications

- Strobo-laryngoscopy

- Observing machine parts, motors etc.
Periodic Signals

Periodic Signal \((x(t))\) with period \(P\)

Periodic signal with period \(P\) and band-limited to \(f_{\text{Max}}\).

The Fourier domain contains only terms that are multiples of \(f_p = 1/P\).
Normal Camera and Aliasing

Integrates out the high frequency information and this information is irrevocably lost.

\[ f_{\text{Max}} \] 

Causes aliasing of high frequency information.

\[ T_{\text{Frame}} = \text{Frame Duration} = 40\text{ms} \]

\[ P = 10\text{ms} \]

\[ f_S/2 = \text{Sampling Frequency}/2 \]
High Speed Camera

Nyquist Sampling of $x(t)$ – When each period of $x$ has very fine high frequency variations Nyquist sampling rate is very high.

High speed video camera is inefficient since its bandwidth is wasted on the periodic signal since it is sparse in the Fourier domain.
Traditional Strobing

When sampling rate of camera is low, generate beat frequencies of the high speed periodic signal which now can be easily captured.

The light throughput is also very low in order to avoid blurring of the features during the strobe time.

The period of the signal needs to be known \textit{a priori}.
Coded Strobing

Scene

Coded Shutter
Coded strobing: Time Domain

Proposed Design: In every exposure duration observe different linear combinations of the underlying periodic signal.

Advantage of the design:
Exposure coding is independent of the frequency of the underlying periodic signal.
Further, light throughput is on an average 50% which is far greater than can be achieved via traditional strobing.
Coded strobing: Frequency domain

Measure linear combinations of the Fourier components.

Decode by enforcing sparsity of the reconstructed signal in Fourier domain.
Signal model

\[ \mathbf{x} = \mathbf{B} \mathbf{s} \]

Signal

\[ \begin{bmatrix} b_1 \ b_2 \ \vdots \ b_N \end{bmatrix} \]

Fourier Basis

Basis Coeff

Non-zero elements
Observation Model

\[
\begin{align*}
\text{Observed Intensity} & \quad \text{Frame Integration Period } T_s \\
\text{Frame 1} & = \quad \text{Coded Strobing} \\
\text{Frame M} & = \quad \text{Signal}
\end{align*}
\]

\[
y = CX
\]

\(N \text{ unknowns}\)
Signal and Observation model

**Signal Model:**

\[ x = B \, s \]

**Observation Model:**

\[ y = C \, x \]

<table>
<thead>
<tr>
<th>Observed low-frame rate video</th>
<th>Mixing matrix (satisfies Restricted Isometry under certain conditions)</th>
<th>Fourier coefficients of the signal (sparse if periodic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>( C , B , s )</td>
<td>( A , s )</td>
</tr>
</tbody>
</table>

\[ A \text{ of size } M \times N, \quad M \ll N \]
Linear Inversion?

- \( \mathbf{y} = \mathbf{A} \mathbf{s} \) highly underdetermined. \( M << N \)

- For example: Capture at 10 fps for 30s with 1 ms resolution has \( M=300, N=30000 \)

- Highly unstable in noise

- To reconstruct can we use extra information about the signal?
Extra info: Periodic signal

\[ y = A \cdot s \]

Observed low rate frames

Mixing matrix

Sparse Basis Coeff

Very few (K) non-zero elements
Compressed Sensing

Estimate sparse $s$ that satisfies $y = A s$

Solve

$$\min \| s \|_1 \quad s.t. \quad \| y - As \|_2 \leq \varepsilon$$

(P1)

P1 is a linear program. Fast and robust algorithms available for solving P1
Signal model and recovery

- Signal Model
- Observation Model
- Compressed sensing
- Sparsity and location enforced recovery

- $x = B \, s$
- $y = C \, x$

Solve

$$\min \|s\|_1 \quad s.t. \quad \|y - As\|_2 \leq \varepsilon$$
Periodic Signals

Periodic signal with period $P$ and band-limited to $f_{\text{Max}}$.

The Fourier domain contains only terms that are multiples of $f_p=1/P$. 

![Diagram of a periodic signal with period $P$]
Structure of Sparsity

\[ y = A \cdot s \]

- \( y \): Observed low rate frames
- \( A \): Mixing matrix
- \( s \): Sparse Basis Coeff

Fundamental frequency and its harmonics
Sparsity and location

- In addition to sparsity, use the location of the non-zero coefficients as constraint

- All the non-zero coefficients equally spaced at $f_p$

- Solve

\[ \min \| \overline{s} \|_1 \quad s.t. \quad \| y - \overline{A}s \|_2 \leq \varepsilon \]  
(P1t)

with truncated mixing matrix and basis coefficients
Signal model and recovery

- **Signal Model**
  - $x = B s$

- **Observation Model**
  - $y = C x$

- **Compressed sensing**

- **Sparsity and location enforced recovery**
  - $\min \| s \|_1 \text{ s.t. } \| y - As \|_2 \leq \varepsilon$
  - $\min \| \tilde{s} \|_1 \text{ s.t. } \| y - \tilde{A}s \|_2 \leq \varepsilon$
Implementation

Captured at 10fps using PGR Dragonfly2

Can be strobed at 1ms resolution

Can flutter at 250us

FLC Shutter
Example: Fractal simulation

Original Video: Period 25ms

Recovered Video

Captured by 10fps camera

Captured by 10fps coded strobing camera
Rotating Mill Tool (Dragonfly2)

Normal Video: 10fps

Reconstructed Video
3000 RPM

Reconstructed Video
4500 RPM

Reconstructed Video
6000 RPM

Reconstructed Video
12000 RPM
Compressive Sensing: Linear Dynamical Systems

Aswin Sankaranarayanan
Pavan Turaga
Richard Baraniuk
Rama Chellappa
**CSLDS: Video CS under LDS model**

- **Motivation**
  - Extensive use in **dynamic textures**, sequence of LDS for **activity analysis**, etc...
  - LDS $==$ Frames of the video lie on a low dimensional subspace. Linearity applicable over short durations

- **Benefits**
  - LDS is a parametric model. Video CS is equivalent to parameter estimation
  - Majority of the LDS parameters are **static**. This is extremely helpful in achieving high compression at sensing
Linear dynamical system example

Video sequence

Few frames

Six key images

All frames can be estimated using linear combinations of SIX images
LDS model

Observation model

Data lies on a low dimensional subspace

C is a basis for the subspace. For uncompressed data, C is usually learnt using PCA on training examples.

\[ x_t = C v_t \]

\[ x_t \in \mathbb{R}^N, v_t \in \mathbb{R}^d, d << N \]

Terminology

\( x_t \) are the frames of the video (high dim)

\( v_t \) are the subspace coefficients (low dim)
Can we recover \( C \) and \( v_t \) from \( y_t \)?

Very hard since the unknowns are bilinear.
Challenges and solution

- **Key challenge:** Unknown variables appear as **bilinear relationships**
  - CS theory handles unknowns in a linear relationship
- **Solution:** A novel measurement model that reduces the bilinear problem to a sequence of linear/convex problems
  - Stable reconstruction at a measurement rate proportional to the state space dimension.
Results

Ground truth  Comp 20x  Comp 50x
SNR: 24.9 dB  SNR: 20.4 dB

Ground truth  Comp20x  Comp50x
SNR: 25.3 dB  SNR: 23.8 dB
Linear Motion Model:
Invertible Motion Blur in Video

Amit Agrawal
Yi Xu
Ramesh Raskar
Traditional Exposure Video

Motion PSF (Box Filter)

Fourier Transform

Information is lost

Exposure Time

Slide courtesy Agrawal
Varying Exposure Video

Exposure Time

Fourier Transform
Varying Exposure Video

Exposure Time

Fourier Transform

No common nulls
Varying Exposure Video

Exposure Time

Fourier Transform

No common nulls

Slide courtesy Agrawal
Varying Exposure Video = PSF Null-Filling

Joint Frequency Spectrum
Preserves All Frequencies

Fourier Transform

Slide courtesy Agrawal
Varying Exposure Video
Blurred Photos

Deblurred Result

Slide courtesy Agrawal
Single Image Deblurring (SID)

Multiple Image Deblurring (MID)

Slide courtesy Agrawal
The Story so far

• Periodic or Repetitive Videos
  – CS Strobing

• Linear Dynamical Systems
  – CS – LDS

• Linear Motion Model
  – Multi-Exposure Video
How about general Videos?

- General Multiplexing
  - Single Pixel Camera

- Voxel Subsampling
  - Flexible Voxels
  - Coded per-Pixel Exposure Video

- Voxel Multiplexing
  - Programmable Pixel Compressive Camera
Single pixel camera

- Each configuration of micro-mirrors yield **ONE** compressive measurement

- **Non-visible wavelengths**
  - Sensor material costly in IR/UV bands

- **Light throughput**
  - **Half the light** in the scene is directed to the photo-detector
Single-Pixel Camera

Target

$N = 65536$ pixels

$M = 11000$ measurements (16%)

$M = 1300$ measurements (2%)
From Image to Video Sensing

• SPC works under the assumption of a static scene
  – However, for dynamic scenes changes as we collect measurements

• Need to measure temporal events at their “information rate”
  – Need to borrow richer models for videos and exploit them at sensing and reconstruction
InGaAs Photo-detector (Short-wave IR)
SPC sampling rate: 10,000 sample/s
Number of compressive measurements: \( M = 16,384 \)
Recovered video: \( N = 128 \times 128 \times 61 = 61 \times M \)
CS-MUVI: IR spectrum

initial estimate
Upsampled

Recovered Video

Joint work with Xu and Kelly
CS-MUVI on SPC

Naïve frame-to-frame recovery

CS-MUVI

Joint work with Xu and Kelly
Voxel Subsampling Compressive Video

Flexible Voxels for Content Aware Videography
ECCV 2010
Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, Srinivas Narasimhan

Video from a Single Exposure Coded Photograph
ICCV 2011
Yasunobu Hitomi, Jinwei Gu, Mohit Gupta, Tomoo Mitsunaga, Shree K. Nayar
Fundamental Resolution Trade-off

- **Spatial resolution (pixels)**
  - 64K: 256x256
  - 1M: 1024x1024

- **Temporal resolution (fps)**
  - High-speed Cameras: 500
  - Conventional video camera: 30

*Slide adapted from Ben-Ezra et al.*
Imaging Dynamic Scenes

Low-Frame Rate (30 fps) Camera
Large Motion Blur

High-Speed (480 fps) Camera
Low Spatial Resolution
‘Motion Aware’ Camera

Variable Spatio-Temporal Resolution
Spatio-Temporal Sampling

High SR, Low TR
- Fast Moving
- Slow Moving
- Static

Low SR, High TR
- Fat and short voxels

Motion Aware Video
- Flexible voxels

Thin and long voxels
Sampling of the Space-Time Volume

Conventional Sampling Scheme:

Our Sampling Scheme:
Multi-resolution Sampling

Captured Data

Desired Spatio-temporal Interpretations

Time

Space

TR = Gain in Temporal Resolution  SR = Gain in Spatial Resolution
Multi-resolution Sampling

Captured Data

Different Spatio-temporal Interpretations

Re-binning

TR = 1X
SR = 1X

TR = 2X
SR = ½ X

TR = 4X
SR = ¼ X

What is the correct firing order?

Unsampled Bins

TR = 1X
SR = 1X

TR = 2X
SR = ½ X

TR = 4X
SR = ¼ X
Computing the Firing Order

Captured Data

Desired Spatio-temporal Interpretations

\[ x_p \in \{0,1\} \]

\[ p = \text{location index} \quad K = \text{Number of pixels in a group} \]

\[ B_{ij} = \text{\(i^{th}\) block in \(j^{th}\) interpretation} \]

Integer Linear Program:

Each Block contains exactly one measurement

Number of measurements = Number of pixels

\[ \sum_{p \in B_{ij}} x_p = 1 \quad \forall B_{ij} \]

\[ \sum_{p=1}^{K} x_p = K \]

\[ K^2 \]
Firing Order and Reconstructions for a 4x4 Block

TR = 1, SR = 1/1
1 16 2 15
8 9 7 10
4 13 3 14
5 12 6 11

TR = 2, SR = 1/2
1 2
8 7
4 3
5 6

TR = 4, SR = 1/4
1 16 15
9
13 14
11

TR = 8, SR = 1/8
1 16 15
9
13 14
11

TR = 16, SR = 1/16
1 16 15
9
13 14
11
Co-located Projector-Camera Setup

Scene

Beam Splitter

Image Plane

Projector

Image Plane

Camera

Camera Integration Time

Projector Pattern

Pixel 1

Pixel 2

Pixel K

Time
Co-located Projector-Camera Setup

Scene

Beam Splitter

Image Plane

Projector

Image Plane

Camera

Beam-splitter  Projector

Camera
LCOS Hardware Prototype

Image Sensor

Polarizing beam splitter

Virtual image plane

LCoS

Relay lens 2

Relay lens 1

Objective lens

Image Sensor

LCoS

C-Mount

Polarizing beam splitter

Objective lens

C-Mount

Relay lens
Multiple Balls Bouncing and Colliding (15 FPS)

Large Motion Blur

Close-up
Reconstruction at 4X Temporal Resolution (60FPS)

Aliasing Artifacts on the Background
Reconstruction at 16X Temporal Resolution (240FPS)

Aliasing Artifacts on the Background
Motion-aware Video

Captured Frame

Increasing Temporal Resolution

Different Spatio-temporal Interpretations

Motion Analysis

Optical Flow Magnitudes

Motion-Aware Video
Motion-aware Video

Simple Rebinning

Motion-Aware Video
Multiple Balls Bouncing

Input Sequence

Motion-Aware Video
Comparison

Input Sequence

Motion-Aware Video
Voxel Subsampling with Dictionary Learning?

(1) Coded sampling

- Unknown scene: \( E \)
- Coded exposure sampling function: \( S \)
- Coded image: \( I \)

(2) Dictionary learning (offline)

- Training data set
- K-SVD
- Overcomplete dictionary: \( D \)

(3) Sparse reconstruction

- Sparse representation: \( E = D \alpha \)
- Estimation of \( \hat{\alpha} \)
- Recovered space-time volume: \( \hat{E} \)

Slide Courtesy Hitomi et. al.
Pixel-wise Coded Exposure Sampling

Original scene \[\rightarrow\] Exposure pattern (white=on, black=off) \[\rightarrow\] Input Coded image

Slide Courtesy Hitomi et. al.
Dictionary learning

Learning video set
20 scenes, 8 spatial rotations,
Play forward and backward

Overcomplete dictionary bases

Slide Courtesy Hitomi et. al
Simulations
Road Traffic

Original scene
(450x300x36)

Coded image
(450x300)

Slide Courtesy Hitomi et. al.
Simulation Results

Thin-out movie
(13x9x36 → 450x300x36)

Grid sampling + interpolation
(450x300x36)

3D DCT representation + sparse reconstruction
(450x300x36)

Our method
(450x300x36)

Slide Courtesy Hitomi et. al
Dog’s Tongue

Original scene
(384x216x36)

Coded image
(384x216)

Slide Courtesy Hitomi et. al
Simulation Results

Thin-out movie
(11x6x36 $\rightarrow$ 384x216x36)

Grid sampling + interpolation
(384x216x36)

3D DCT representation + sparse reconstruction
(384x216x36)

Our method
(384x216x36)
Experimental Results
Falling Ball

Coded image (320x320, 18ms)

Reconstructed Movie (320x320x18)

Slide Courtesy Hitomi et. al.
Fluid (milk) pouring

Coded image
(320x320, 18ms)

Reconstructed Movie
(320x320x18)

Slide Courtesy Hitomi et. al.
Blinking Eye

Coded image (320x320, 27ms)

Reconstructed Movie (320x320x9)

Slide Courtesy Hitomi et. al.
Voxel Multiplexing Compressive Video

Programmable Pixel Compressive Camera
CVPR 2011

Dikpal Reddy, Ashok Veeraraghavan, Rama Chellappa
P2C2: Programmable Pixel Compressive Camera

Capture

Fast moving scene

Camera operating at 25 fps

Spatio-temporal mask modulates every pixel independently.
Mask has 1 megapixel and modulates at 200 fps.

0.25 megapixel Sensor array

Captured low-res coded frames from a 25fps camera

Recovery

Optimization using Brightness constancy & Wavelet domain sparsity

Captured low-res coded frames from a 25fps camera

Video credit: TechImaging
P2C2: Prototype

Objective Lens
F=80 mm
D=40 mm

Polarizing beamsplitter

LCOS Mirror
SXGA-3DM

Camera Lens
Pentax 35mm
P2C2: Programmable pixel compressive camera

Camera operating at 25 fps
Modulation at 200 fps

Integration

Coded motion blur
Subsampling spatially by 2
Frame captured by P2C2

$X(:, :, t)$ $X(:, :, t) \cdot M(:, :, t)$

Reconstruction Algorithm

Temporal superres $L_t = 8$
Spatial superres $L_s = 2$

Frame captured by P2C2
P2C2: Programmable pixel compressive camera

Observations and unknowns

\[ L_s = 1, \quad L_t = 4 \]

\[ \begin{array}{c}
\text{P2C2 captured frame} \quad y_2 \\
\text{mask} \quad m_6 \\
\text{high speed frame} \quad x_6 
\end{array} \]
P2C2: Programmable pixel compressive camera

Observations and unknowns

\[ y = \Phi x \]

\[ L_S = 1, \quad L_t = 4 \]

• Severely ill-conditioned system of equations \((2^{16} \times 2^{20})\)

\[ E_{data} = \| y - \Phi x \|_2^2 \]
P2C2: Brightness constancy constraints

\[ x_2 (\cdot) - x_1 (\cdot) = 0 \]

Forward flow

Brightness constancy
P2C2: Brightness constancy constraints

\[ \mathbf{E}_{BC} = \lambda \left\| \mathbf{\Omega} \mathbf{x} \right\|_2^2 \]

\[ \mathbf{\Omega} \mathbf{x} = 0 \]

Forward flow

Backward flow

Subpixel flow

Bilinear interpolation

Occluded pixel

Pixel occluded in t=4

Row removed
P2C2: Transform sparsity

\[ E_{spatial} = \beta \left\| \Psi^{-1} x_t \right\|_1 \]
P2C2: Dancers video

Original high speed scene

P2C2 capture

Normal camera capture

\[
\begin{align*}
\text{Temporal superres} & : \quad L_t = 8 \\
\text{Spatial superres} & : \quad L_s = 1
\end{align*}
\]

Video credit: Photron
If brightness constancy constraints available

• Solve

$$\min \quad E_{data} + E_{BC} + E_{spatial}$$

$$\min \quad \|y - \Phi x\|^2_2 + \lambda \|\Omega x\|^2_2 + \sum_{t=1}^{T} \beta \|\Psi^{-1} x_t\|_1$$

• But optical flow not available apriori

• Optical flow can be extracted from smooth, low spatial resolution images

Estimate sub-frames without brightness constancy constraints
Recover sub-frames by solving

- Solve

\[
\min E_{\text{data}} + E_{\text{spatial}}
\]

\[
x^0 = \arg\min \left\| y - \Phi x \right\|_2^2 + \sum_{t=1}^{T} \beta \left\| \Psi^{-1} x_t \right\|_1
\]

- Significantly ill conditioned
- Set high-frequency coefficients to zero

smooth, high spatial frequency is lost

preserves motion allowing OF to be estimated
P2C2: Estimate Optical Flow

Recover flow from estimated sub-frames

- Perform consistency check using backward and forward flow
- Perform occlusion reasoning to prune the brightness constancy constraints

\[
\text{abs}(u + iv)
\]

\[
\text{abs}(u^0 + iv^0)
\]
If brightness constancy constraints available

- Solve

\[
\min E_{data} + E_{BC} + E_{spatial}
\]

\[
\min \| y - \Phi x \|_2^2 + \lambda \| \Omega x \|_2^2 + \sum_{t=1}^{T} \beta \| \Psi^{-1} x_t \|_1
\]

- But optical flow not available a priori

- Optical flow can be extracted from smooth, low spatial resolution images

Estimate sub-frames without brightness constancy constraints
P2C2: Algorithm

• Estimate sub-frames $\mathcal{X}^0$ without using BC constraints

Begin iteration (iteration index k)

• Estimate OF $(u^{k-1}, v^{k-1})$ from $\mathcal{X}^{k-1}$

• Build BC constraint matrix $\Omega^{k-1}$

• Estimate sub-frames $\mathcal{X}^k$ by minimizing $E_{total}$

End iteration
P2C2: Recovered Examples

Original high speed scene

Reconstructed scene

Normal camera capture

Row-Time slice

occlusion
P2C2: Recovered Examples

Reconstructed scene

Normal camera capture

Original high speed scene

Row-Time slice

occlusion
How about general Videos?

• General Multiplexing
  – Single Pixel Camera

• Voxel Subsampling
  – Flexible Voxels
  – Coded per-Pixel Exposure Video

• Voxel Multiplexing
  – Programmable Pixel Compressive Camera
Whats next?

• Motion Blur
  – Flutter Shutter
  – Parabolic Exposure

• Rolling Shutter Cameras
  – Coded Rolling Shutter

• High Dimensional Visual Compressive Sensing
  – Compressive Light Transport
  – Compressive volumetric Scattering
  – Reinterpretable Imaging