Multiple-View Object Recognition via Sparse Representation

Allen Y. Yang
Subhrausn Maji, Mario Christoudas, Trevor Darrell, Jitendra Malik, Shankar Sastry

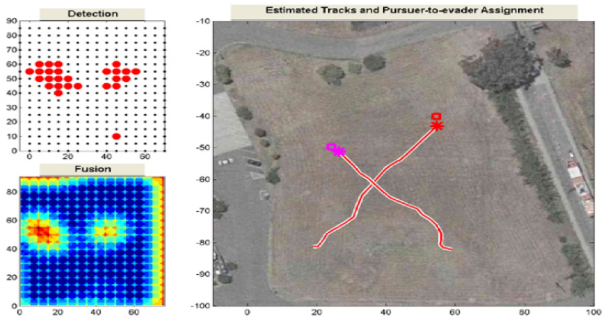
DSP Seminar, University of Illinois, 2009

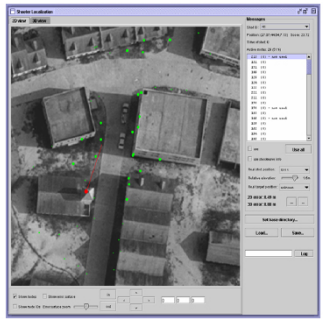
Heterogeneous Sensor Networks

- ARO MURI: **Heterogeneous Sensor Webs** for Automated Target Recognition and Tracking in Urban Terrain

Heterogeneous Sensor Networks

- ARO MURI: **Heterogeneous Sensor Webs** for Automated Target Recognition and Tracking in Urban Terrain





- **Organization**
  - Berkeley: Shankar Sastry (PI)
  - MIT: Alan Willsky (co-PI)
  - Vanderbilt: Xenofon Kousoukos (co-PI)
  - Memphis: Bela Bollobas (co-PI)

Berkeley University of California

VANDERBILT

Heterogeneous Sensor Networks

- ARO MURI: **Heterogeneous Sensor Webs** for Automated Target Recognition and Tracking in Urban Terrain

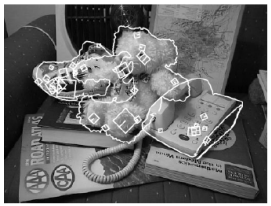
- **Organization**
  - Berkeley: Shankar Sastry (PI)
  - MIT: Alan Willsky (co-PI)
  - Vanderbilt: Xenofon Kousoukos (co-PI)
  - Memphis: Bela Bollobas (co-PI)

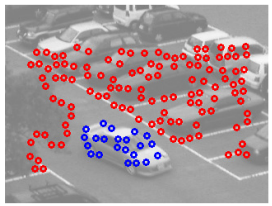
- Collaboration
  - **NSF TRUST Center** at UC Berkeley: Ubiquitous Secure Technology
  - **California CITRIS Center**: Information Technology Research in the Interest of Society.

Key Technical Problems

1. 3-D Scene Analysis in Dense Urban Environments

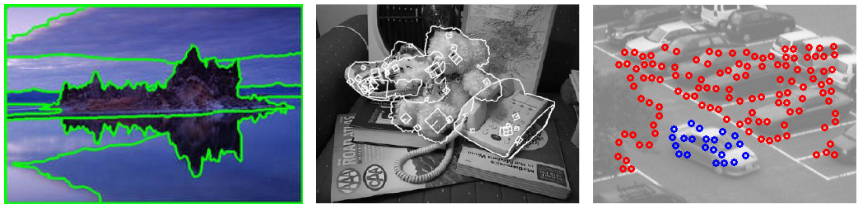

(a) Image


(b) Object


(c) Motion

Key Technical Problems

1. 3-D Scene Analysis in Dense Urban Environments

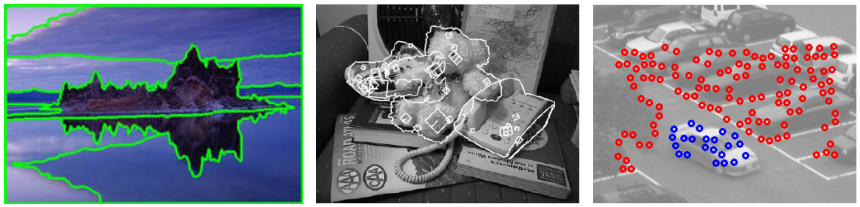


(a) Image (b) Object (c) Motion

2. "Closing the Loop": Sensing on mobile platforms and control

Key Technical Problems

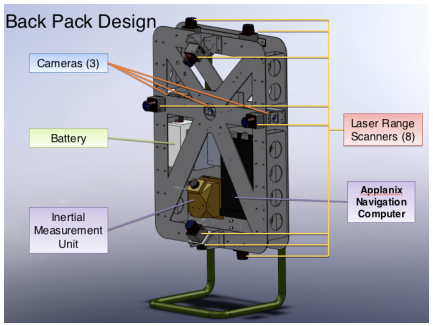
1. 3-D Scene Analysis in Dense Urban Environments

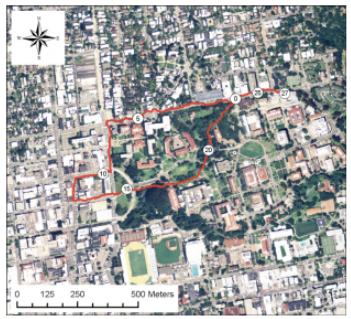


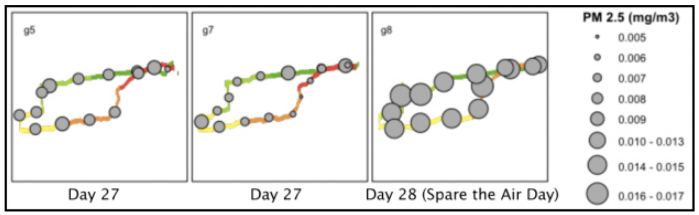
(a) Image (b) Object (c) Motion

2. "Closing the Loop": Sensing on mobile platforms and control

3. The most important mobile platform is **human**: Egocentric sensing, body sensor networks

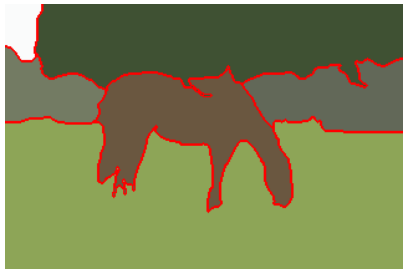






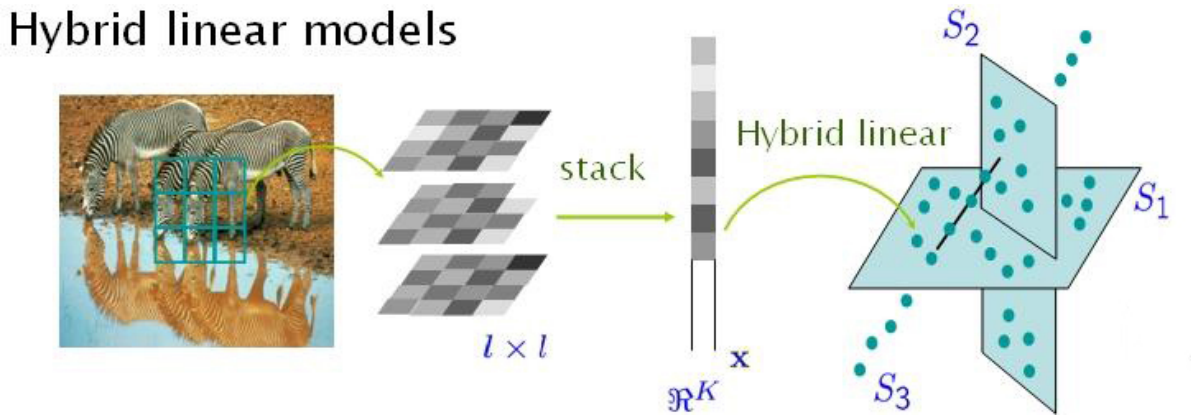
Airborne Particulate Matter Concentrations

Image Segmentation via Lossy Coding Length

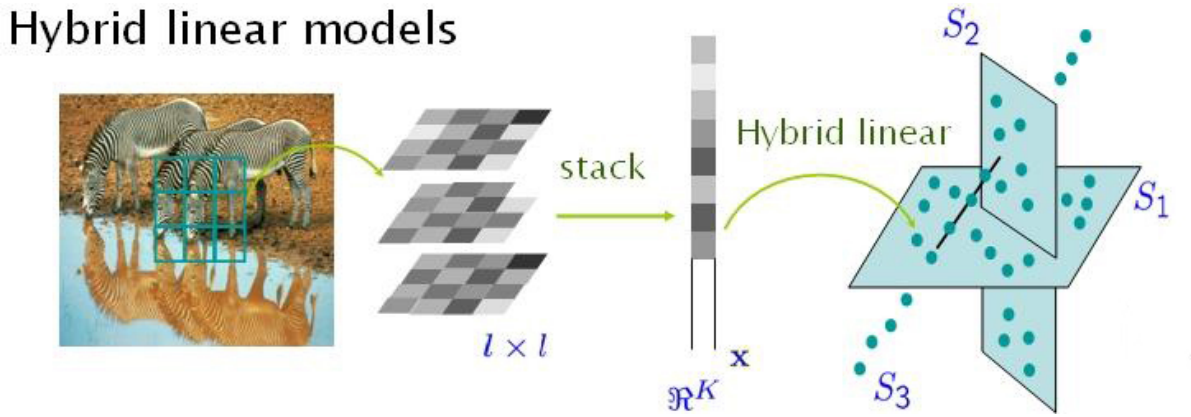





Lossy Minimum Description Length



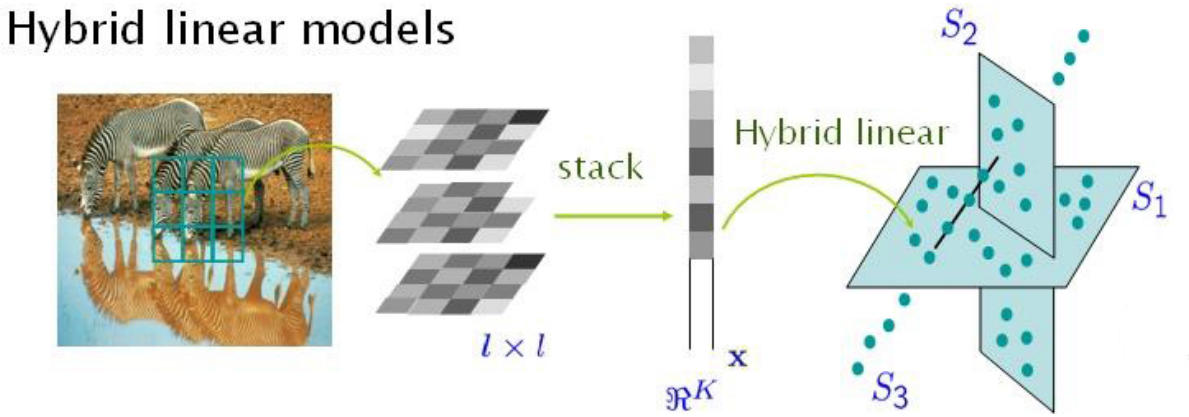
Lossy Minimum Description Length



1. **Lossy coding length** $L_\epsilon(V, \mathcal{A})$:
   Quantize $V = (v_1, \cdots, v_N) \in \mathbb{R}^{D \times N}$ as **a sequence of binary bits** up to a distortion

$$\mathbb{E}[\|v_i - \hat{v}_i\|^2] \leq \epsilon^2.$$

Lossy Minimum Description Length



1. **Lossy coding length** $L_\epsilon(V, \mathcal{A})$:
   Quantize $V = (v_1, \cdots, v_N) \in \mathbb{R}^{D \times N}$ as **a sequence of binary bits** up to a distortion

$$\mathbb{E}[\|v_i - \hat{v}_i\|^2] \leq \epsilon^2.$$

2. **Lossy MDL**

$$\mathcal{A}^*(\epsilon) = \arg\min\{L_\epsilon(V, \mathcal{A}) + \text{Chain Code}(B)\}.$$

Compression-based Image Segmentation

Optimal image segmentation gives rise to the **shortest coding length** to encode images.

Quantitative Comparison on Berkeley Segmentation Dataset

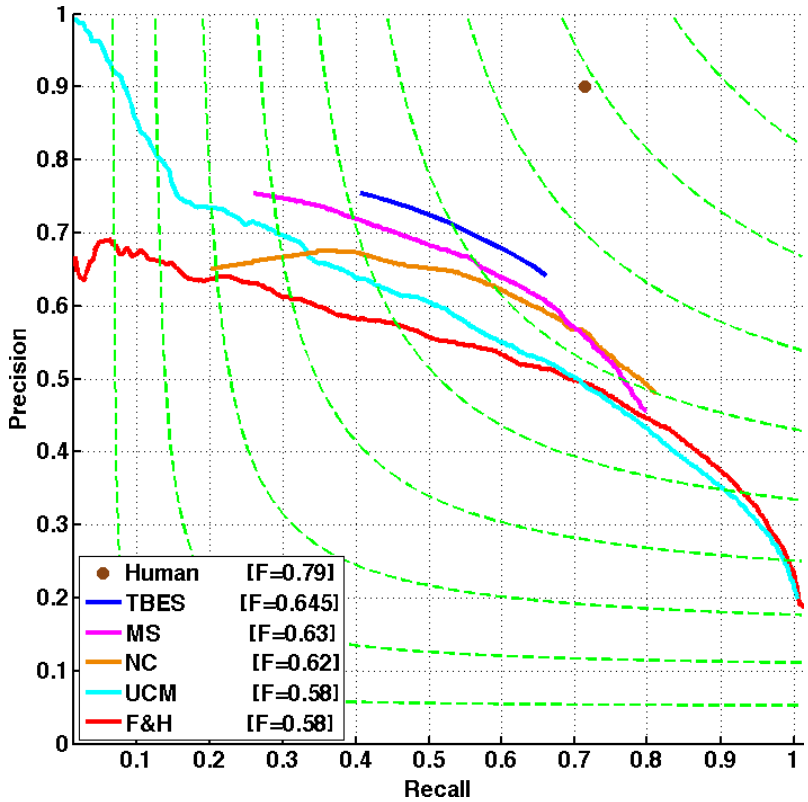


Figure: Precision vs Recall for texture region boundaries.

References:

- Shankar Rao *et al.*, *Natural Image Segmentation with Adaptive Texture and Boundary Encoding*, ACCV, 2009. (best student paper)

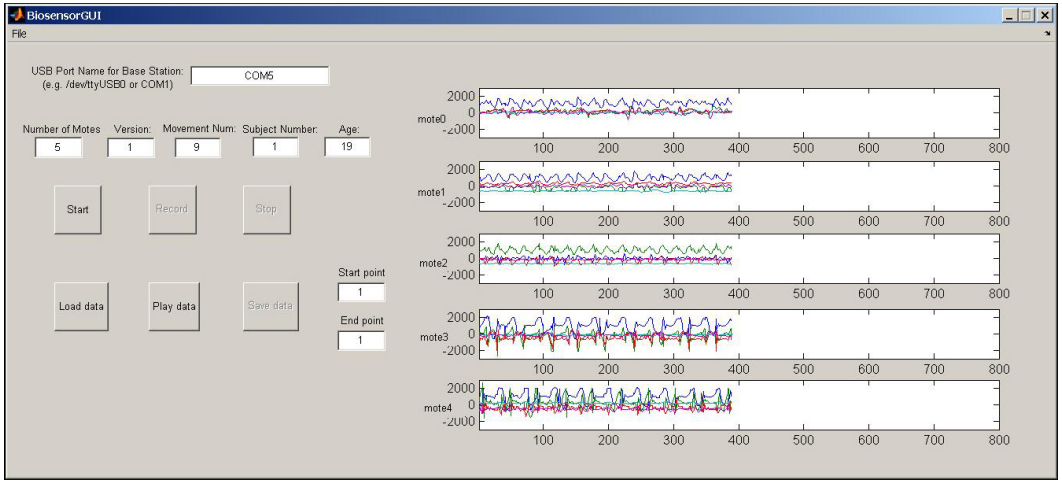
DexterNet: A Wearable Body Sensor Platform

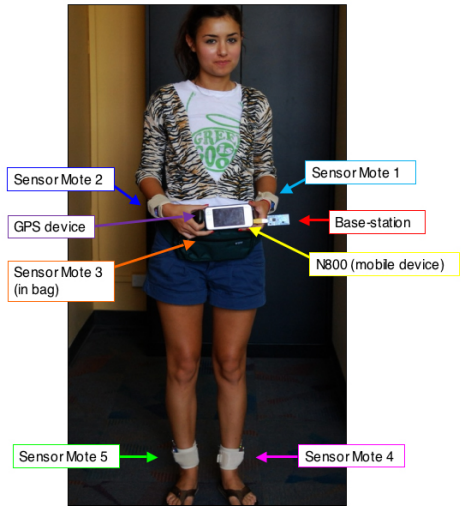
References:

Yang, *et al.*, DexterNet: An open platform for heterogeneous body sensor networks and its applications, BSN 2009.

Yang, *et al.*, Distributed Recognition of Human Actions Using Wearable Motion Sensor Networks, JAISE 2009.

Wearable Action Recognition Database

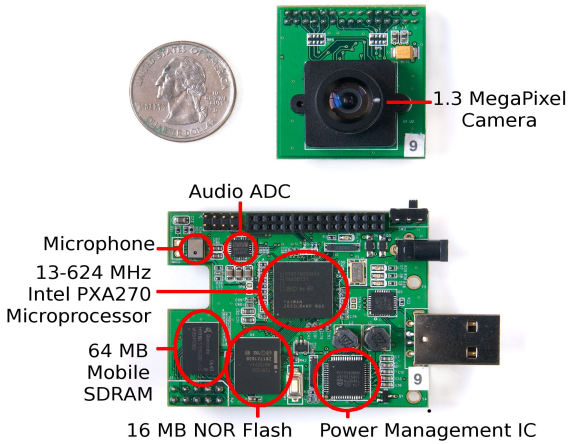




- **Free** for noncommercial users.
- **5 motion sensors**, each carries an accelerometer and gyroscope sampled at 30 Hz.
- **20 test subjects** (13 male & 7 female) ages 19-75.
- **Data processed in Matlab**. Visualization tool is included.

CITRIC: Wireless Smart Camera Platform

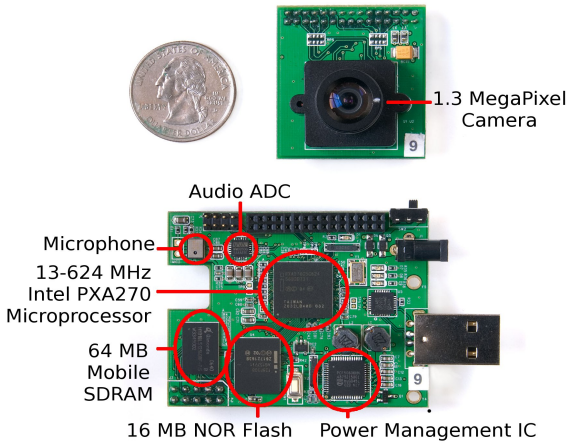
- CITRIC platform



- Available library functions
  1. Full support **Intel IPP Library** and **OpenCV**.
  2. **JPEG compression**: 10 fps.
  3. **Edge detector**: 3 fps.
  4. **Background Subtraction**: 5 fps.
  5. **SIFT detector**: 10 sec per frame.

CITRIC: Wireless Smart Camera Platform

- CITRIC platform



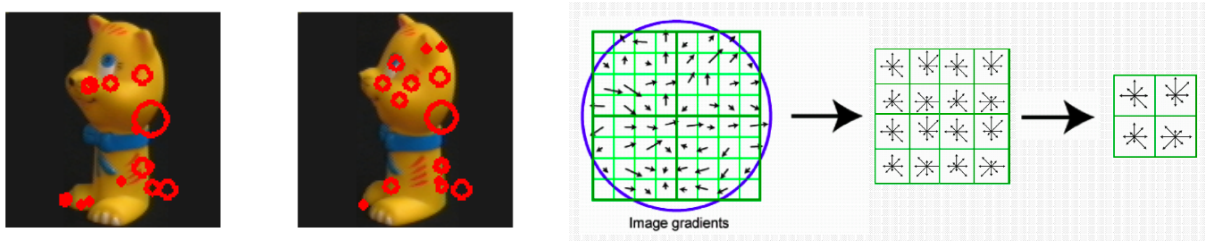
- Available library functions
  1. Full support **Intel IPP Library** and **OpenCV**.
  2. **JPEG compression**: 10 fps.
  3. **Edge detector**: 3 fps.
  4. **Background Subtraction**: 5 fps.
  5. **SIFT detector**: 10 sec per frame.

- Academic users:

Demo: Topological Recovery of a Camera Network

Motivation: Object Recognition

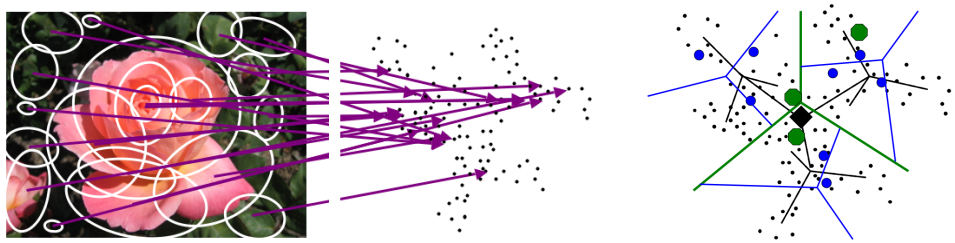
- Affine invariant features, SIFT.



- SIFT Feature Matching [Lowe 1999, van Gool 2004]

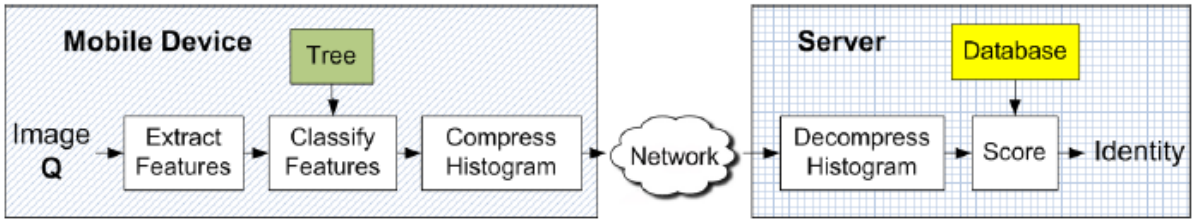
(a) Autostitch (b) Recognition

- Bag of Words [Nister 2006]



Object Recognition in Band-Limited Sensor Networks

1. Compress scalable SIFT tree [Girod et al. 2009]

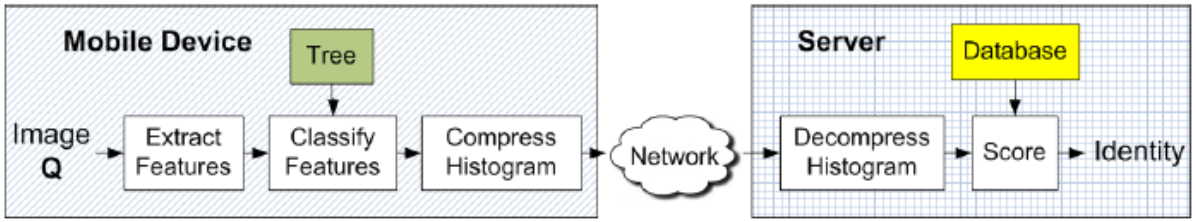


Observation: SIFT histogram is largely sparse (up to 10^6-dim)

R : Sequence of consecutive zero bins.
S : Sequence of nonzero bin values.

Object Recognition in Band-Limited Sensor Networks

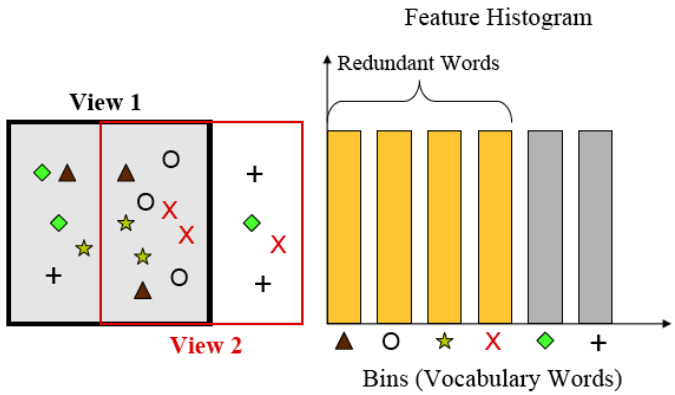
1. Compress scalable SIFT tree [Girod et al. 2009]



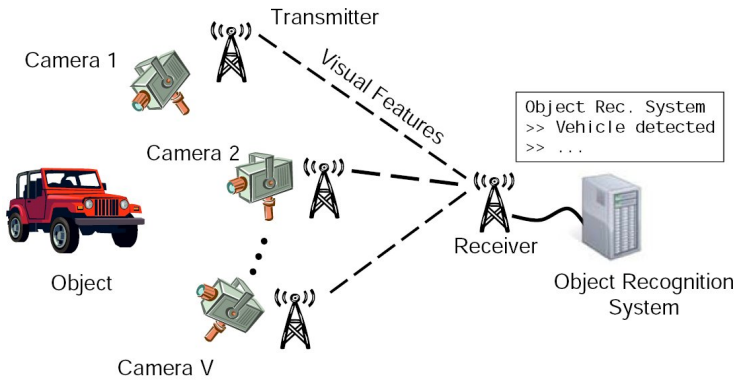
Observation: SIFT histogram is largely sparse (up to 10^6-dim)

R : Sequence of consecutive zero bins.
S : Sequence of nonzero bin values.

2. Multiple-view SIFT feature selection [Darrell et al. 2008]



Problem Statement



1. L camera sensors observe a single object in 3-D.
2. The mutual information between cameras are unknown, cross-sensor communication is prohibited.
3. On each camera, seek an encoding function for a **nonnegative, sparse** histogram $\mathbf{x}_i$

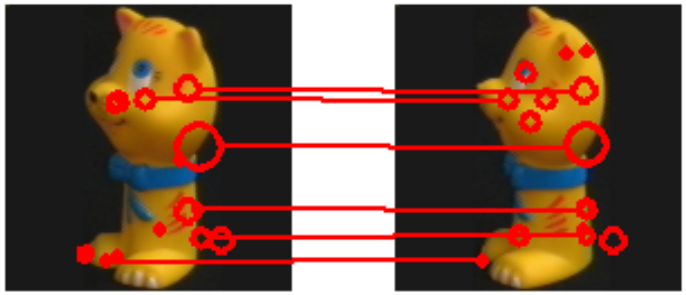
$$f : \mathbf{x}_i \in \mathbb{R}^D \mapsto \mathbf{y}_i \in \mathbb{R}^d$$

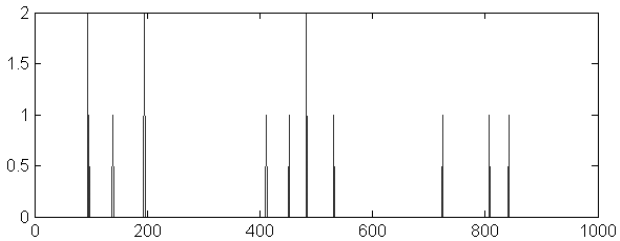
4. On the base station, upon receiving $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_L$, **simultaneously recover**

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L,$$

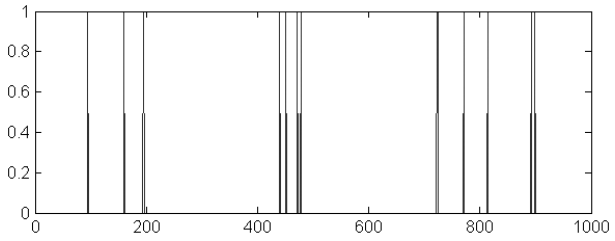
and classify the object class in space.

Key Observations





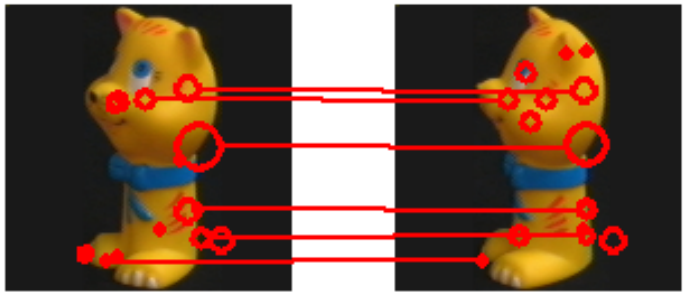
(a) Histogram 1

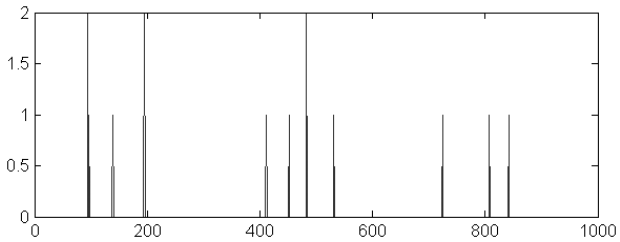


(b) Histogram 2

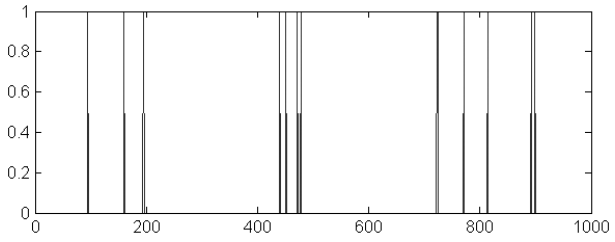
- All histograms are **nonnegative** and **sparse**.

Key Observations





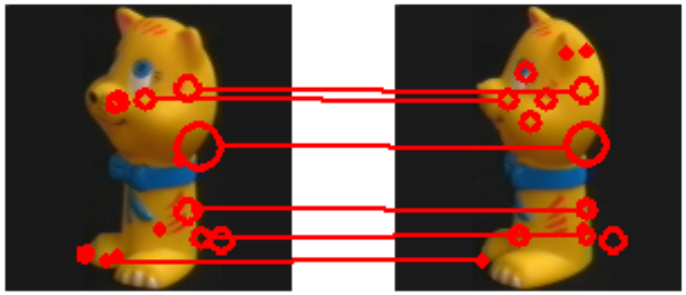
(a) Histogram 1

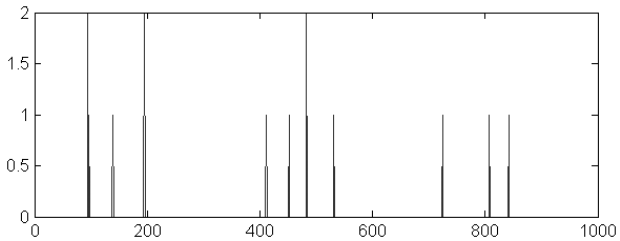


(b) Histogram 2

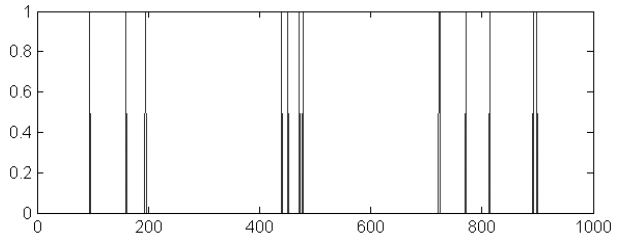
- All histograms are **nonnegative** and **sparse**.
- Multiple-view histograms share **joint sparse patterns**.

Key Observations





(a) Histogram 1



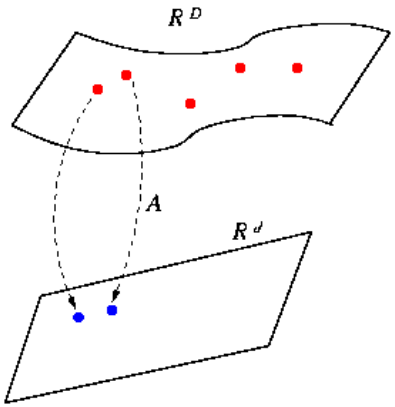
(b) Histogram 2

- All histograms are **nonnegative** and **sparse**.
- Multiple-view histograms share **joint sparse patterns**.
- **Classification** is based on the similarity measure in ℓ^2-norm (linear kernel) or ℓ^1-norm (intersection kernel).

Random Projection as Encoding Function

$$\mathbf{y} = A\mathbf{x}$$

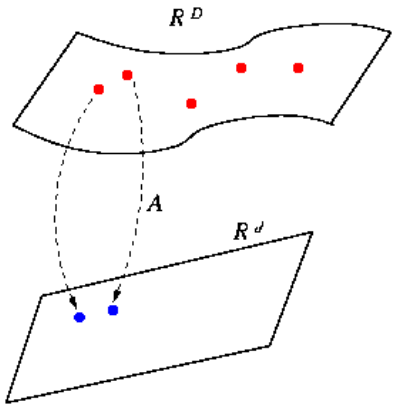
Coefficients of $A \in \mathbb{R}^{d \times D}$ are drawn from zero-mean Gaussian distribution.



Random Projection as Encoding Function

$$\mathbf{y} = A\mathbf{x}$$

Coefficients of $A \in \mathbb{R}^{d \times D}$ are drawn from zero-mean Gaussian distribution.



Johnson-Lindenstrauss Lemma [Johnson & Lindenstrauss 1984, Frankl 1988]

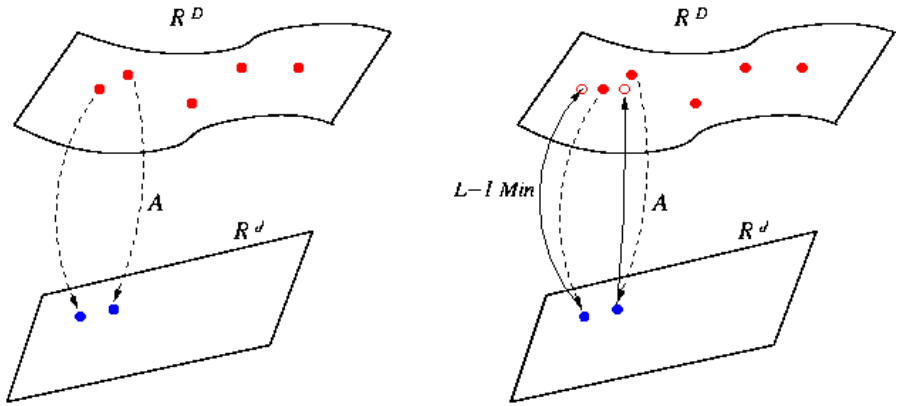
For n number of point cloud in $\mathbb{R}^D$, given distortion threshold ϵ, for any

$$d > O(\epsilon^2 \log n),$$

a Gaussian random projection $f(\mathbf{x}) = A\mathbf{x} \in \mathbb{R}^d$ preserves pairwise ℓ^2-distance

$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 \leq (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2.$$

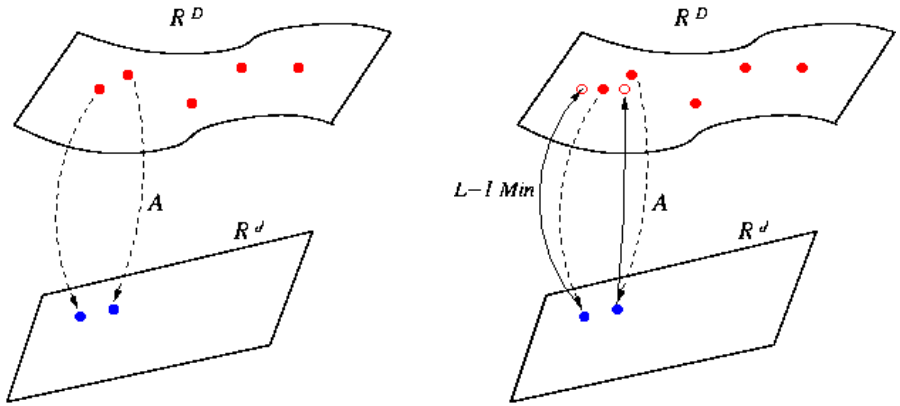
From J-L Lemma to Compressive Sensing



(a) J-L lemma (b) Compressive sensing

1. **Problem I:** J-L lemma does not provide means to reconstruct **histogram hierarchy**.

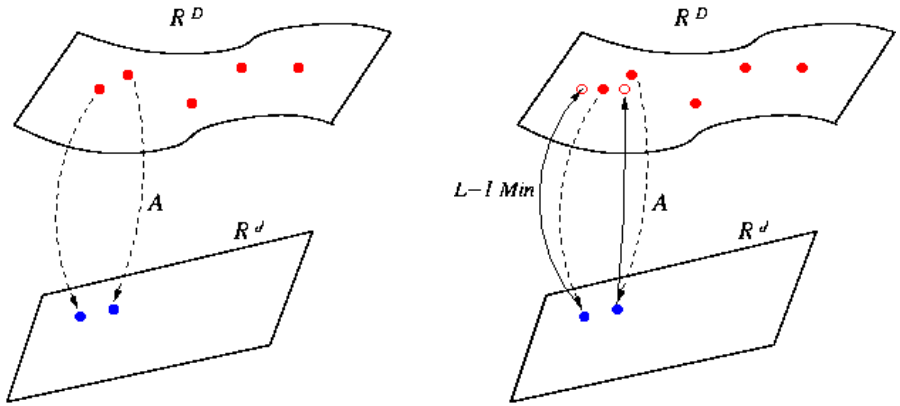
From J-L Lemma to Compressive Sensing



(a) J-L lemma (b) Compressive sensing

1. **Problem I:** J-L lemma does not provide means to reconstruct **histogram hierarchy**.
2. **Problem II:** Gaussian projection **does not preserve ℓ^1-distance** (for intersection kernels).

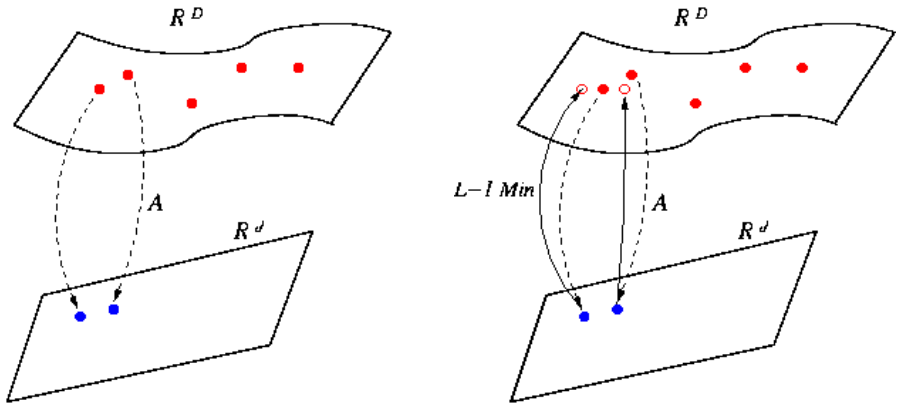
From J-L Lemma to Compressive Sensing



(a) J-L lemma (b) Compressive sensing

1. **Problem I:** J-L lemma does not provide means to reconstruct **histogram hierarchy**.
2. **Problem II:** Gaussian projection **does not preserve ℓ^1-distance** (for intersection kernels).
3. **Problem III:** Difficult (if not possible) to incorporate multiple-view information.

From J-L Lemma to Compressive Sensing



(a) J-L lemma (b) Compressive sensing

1. **Problem I:** J-L lemma does not provide means to reconstruct **histogram hierarchy**.
2. **Problem II:** Gaussian projection **does not preserve ℓ^1-distance** (for intersection kernels).
3. **Problem III:** Difficult (if not possible) to incorporate multiple-view information.

Compressive sensing provides principled solutions to the above problems.

Compressive Sensing

- **Noise-free case:** Assume $\mathbf{x}_0$ is sufficiently k-sparse and mild condition on A,

$$(P_1): \quad \min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = A\mathbf{x}$$

recovers the exact solution.

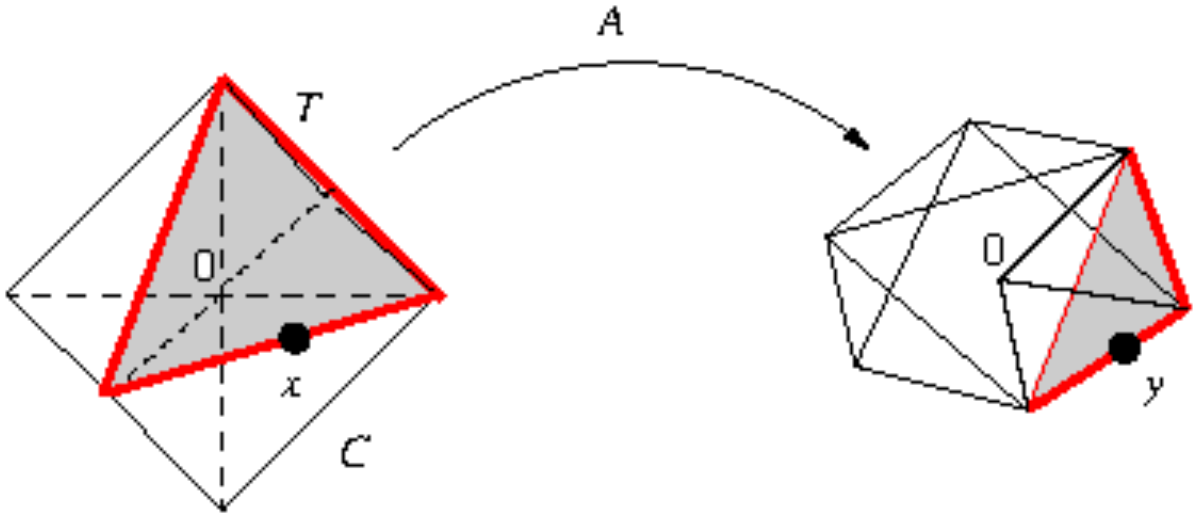
Compressive Sensing

- **Noise-free case:** Assume $\mathbf{x}_0$ is sufficiently k-sparse and mild condition on A,

$$(P_1): \quad \min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = A\mathbf{x}$$

  recovers the exact solution.

- k-Neighborliness



  - Define **cross polytope** C and **quotient polytope** P such that $P = AC$.
  - $\mathbf{x}$ is k-sparse $\Leftrightarrow$ $\mathbf{x}$ lie in a unique $(k-1)$-face of C.

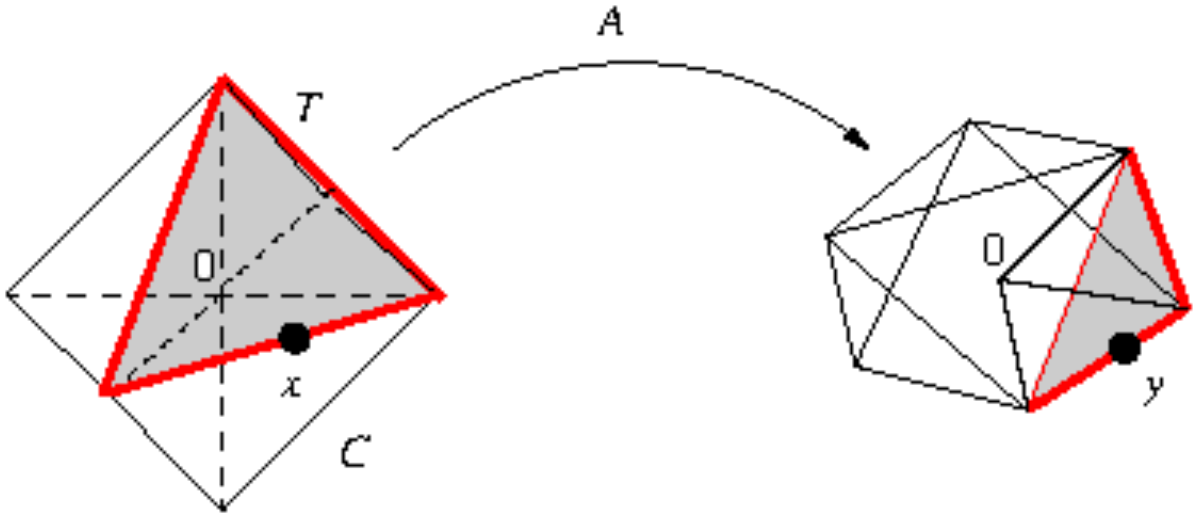
Compressive Sensing

- **Noise-free case:** Assume $\mathbf{x}_0$ is sufficiently k-sparse and mild condition on A,

$$(P_1): \quad \min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = A\mathbf{x}$$

  recovers the exact solution.

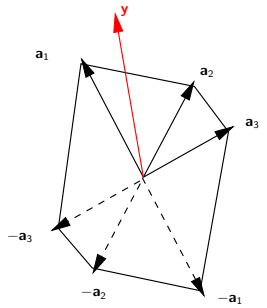
- k-Neighborliness



  - Define **cross polytope** C and **quotient polytope** P such that $P = AC$.
  - $\mathbf{x}$ is k-sparse $\Leftrightarrow$ $\mathbf{x}$ lie in a unique $(k-1)$-face of C.
  - **Necessary and Sufficient:**
    1. If the $(k-1)$-face where $\mathbf{x}$ lies maps to a face of P, then ℓ^1/ℓ^0 holds for this specific $\mathbf{x}$.
    2. If all $(k-1)$-faces of C map to the faces of P on the boundary, ℓ^1/ℓ^0 holds for all k-sparse $\mathbf{x}$.

Matching Pursuit [Mallat & Zhang, 1993]

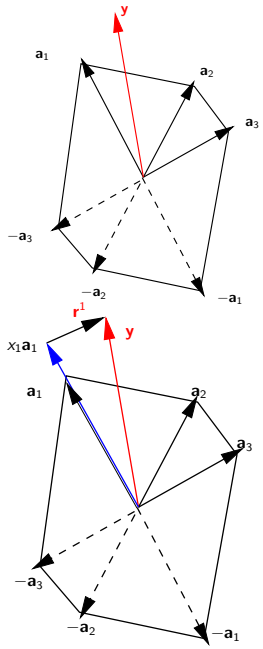
1. **Initialization:**
   - $\mathbf{y} = [A; -A]\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} \geq 0$
   - $k \leftarrow 0$; $\tilde{\mathbf{x}} \leftarrow 0$; $\mathbf{r}^0 \leftarrow \mathbf{y}$; Sparse support $\mathcal{I} = \emptyset$



Matching Pursuit [Mallat & Zhang, 1993]

1. **Initialization:**
   - $\mathbf{y} = [A; -A]\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} \geq 0$
   - $k \leftarrow 0$; $\tilde{\mathbf{x}} \leftarrow 0$; $\mathbf{r}^0 \leftarrow \mathbf{y}$; Sparse support $\mathcal{I} = \emptyset$

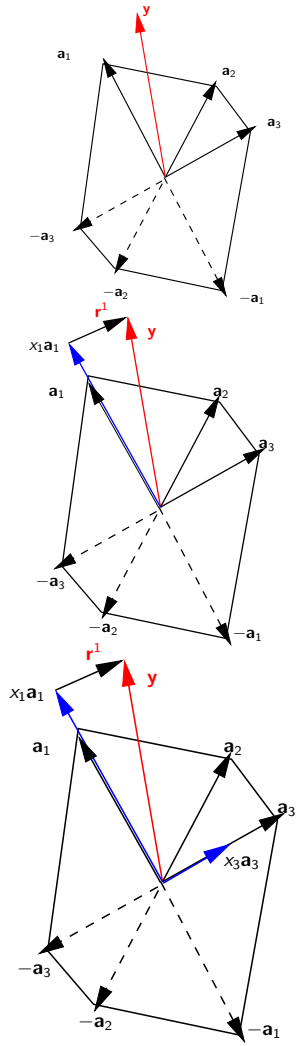
2. $k \leftarrow k + 1$:
   - $i = \arg\max_{j \notin \mathcal{I}} \{\mathbf{a}_j^T \mathbf{r}^{k-1}\}$
   - **Update**: $\mathcal{I} = \mathcal{I} \cup \{i\}$; $x_i = \mathbf{a}_i^T \mathbf{r}^{k-1}$; $\mathbf{r}^k = \mathbf{r}^{k-1} - x_i \mathbf{a}_i$



Matching Pursuit [Mallat & Zhang, 1993]

1. **Initialization:**
   - $\mathbf{y} = [A; -A]\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} \geq 0$
   - $k \leftarrow 0$; $\tilde{\mathbf{x}} \leftarrow 0$; $\mathbf{r}^0 \leftarrow \mathbf{y}$; Sparse support $\mathcal{I} = \emptyset$
2. $k \leftarrow k + 1$:
   - $i = \arg\max_{j \notin \mathcal{I}} \{\mathbf{a}_j^T \mathbf{r}^{k-1}\}$
   - **Update**: $\mathcal{I} = \mathcal{I} \cup \{i\}$; $x_i = \mathbf{a}_i^T \mathbf{r}^{k-1}$; $\mathbf{r}^k = \mathbf{r}^{k-1} - x_i \mathbf{a}_i$
3. **If**: $\|\mathbf{r}^k\|_2 > \epsilon$, **go to** STEP 2; **Else**: **output** $\tilde{\mathbf{x}}$

Fail to search sparse solution on the **boundary** of the quotient polytope.



Fast ℓ^1-Min Routines

1. **Homotopy** Methods:
   - Polytope Faces Pursuit (PFP) [Plumbley 2006]
   - Least Angle Regression (LARS) [Efron-Hastie-Johnstone-Tibshirani 2004]
2. **Gradient Projection** Methods
   - Gradient Projection Sparse Representation (GPSR) [Figueiredo-Nowak-Wright 2007]
   - Truncated Newton Interior-Point Method (TNIPM) [Kim-Koh-Lustig-Boyd-Gorinevsky 2007]
3. **Iterative Thresholding** Methods
   - Soft Thresholding [Donoho 1995]
   - Sparse Reconstruction by Separable Approximation (SpaRSA) [Wright-Nowak-Figueiredo 2008]
4. **Proximal Gradient** Methods [Nesterov 1983, Nesterov 2007]
   - FISTA [Beck-Teboulle 2009]
   - Nesterov's Method (NESTA) [Becker-Bobin-Candés 2009]

MATLAB Toolboxes

- SparseLab: `http://sparselab.stanford.edu/`
- ℓ^1 Homotopy: `http://users.ece.gatech.edu/~sasif/homotopy/index.html`
- SpaRSA: `http://www.lx.it.pt/~mtf/SpaRSA/`

Berkeley University of California

Distributed Object Recognition in Smart Camera Networks

Outlines:

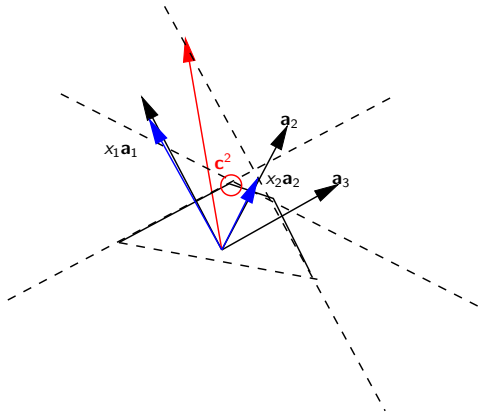
1. How to enforce nonnegativity in decoding SIFT histograms?
2. How to enforce joint sparsity across multiple camera views?

Enforcing Nonnegativity

- **Polytope Pursuit** Algorithms (MP, PFP, LARS):
  1. **Algebraically:** Do not add antipodal vertexes

$$\mathbf{y} = [A; \boxed{\text{-A}}]\tilde{\mathbf{x}}$$

  2. **Geometrically:** Pursuit on positive faces



- **Interior-Point** Algorithms (Lasso, Homotopy, SpaRSA):
  Remove any sparse support that have negative coefficients.

Sparse Innovation Model

- Definition (SIM):

$$\begin{aligned} \mathbf{x}_1 &= \tilde{\mathbf{x}} + \mathbf{z}_1, \\ &\vdots \\ \mathbf{x}_L &= \tilde{\mathbf{x}} + \mathbf{z}_L. \end{aligned}$$

$\tilde{\mathbf{x}}$ is called the **joint sparse** component, and $\mathbf{z}_i$ is called an **innovation**.

Sparse Innovation Model

- Definition (SIM):

$$\begin{aligned} \mathbf{x}_1 &= \tilde{\mathbf{x}} + \mathbf{z}_1, \\ &\vdots \\ \mathbf{x}_L &= \tilde{\mathbf{x}} + \mathbf{z}_L. \end{aligned}$$

  $\tilde{\mathbf{x}}$ is called the **joint sparse** component, and $\mathbf{z}_i$ is called an **innovation**.

- Joint recovery of SIM

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_L \end{bmatrix} &= \begin{bmatrix} A_1 & A_1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \\ A_L & 0 & \cdots & 0 & A_L \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}} \\ \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_L \end{bmatrix} \\ \Leftrightarrow \quad \mathbf{y}' &= A'\mathbf{x}' \in \mathbb{R}^{dL}. \end{aligned}$$

1. New histogram vector is **nonnegative** and **sparse**.
2. Joint sparsity $\tilde{\mathbf{x}}$ is automatically determined by ℓ^1-min: No prior training, no assumption about fixing camera positions.
3. Worst case scenario ($\tilde{\mathbf{x}} = 0$) has the same computational condition as solving individual projections.

Experiment I: Simulation

- Comparison between **matching pursuit**, **polytope faces pursuit**, and **sparse innovation model**:

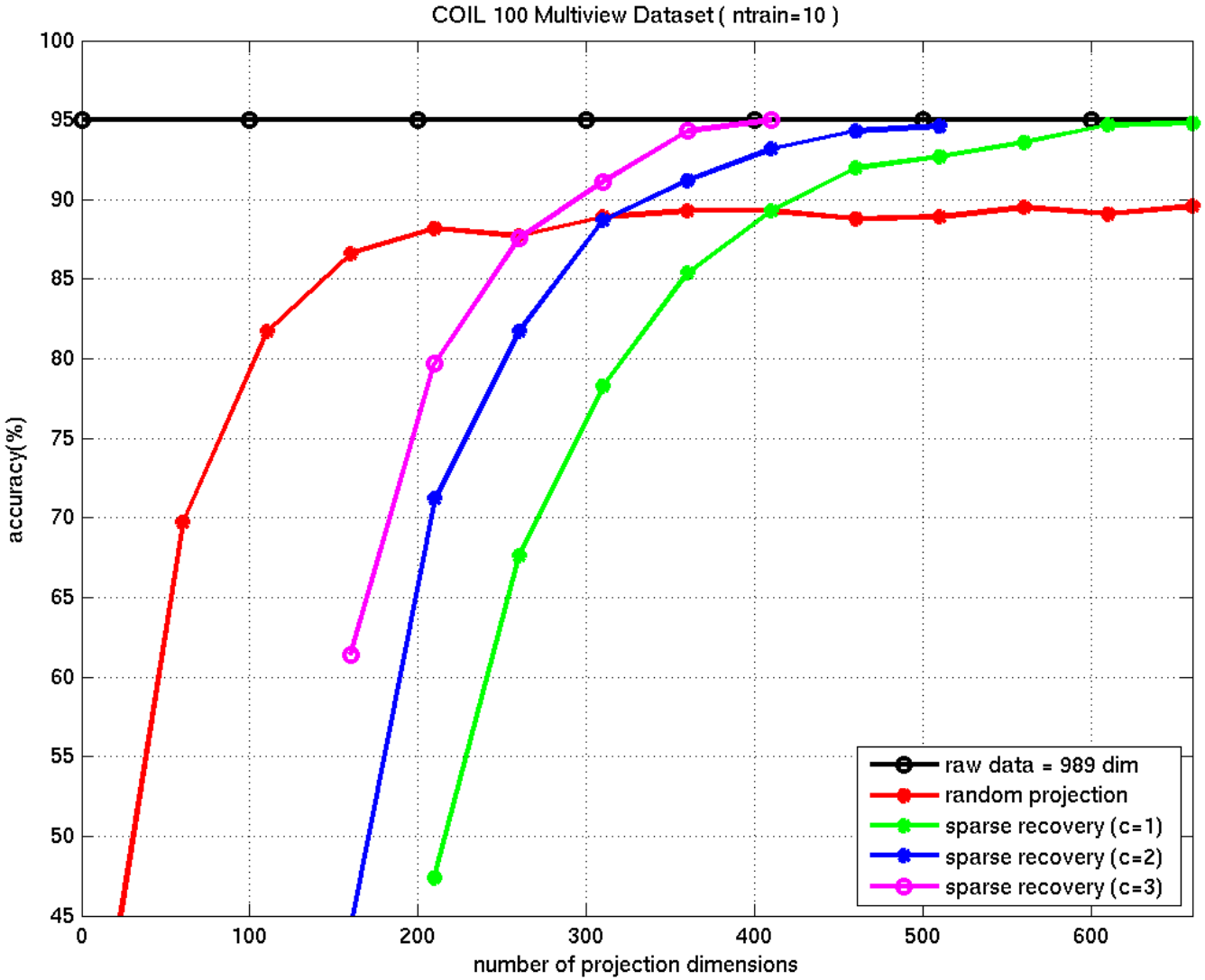
Table: Simulation of solving 1000-D sparse histograms with $d = 200$, $k = 60$, and $L = 3$.

Sparsity	(60,0)	(40,20)	(30,30)
ℓ^0_{MP}	56.14	56.14	56.14
ℓ^2_{MP}	1.76	1.76	1.76
ℓ^0_{PFP}	3.48	3.48	3.48
ℓ^2_{PFP}	0.05	0.05	0.05
ℓ^0_{SIM}	**1.85**	**1.65**	**1.95**
ℓ^2_{SIM}	**0.02**	**0.02**	**0.02**

Experiment II: COIL-100 object database

- **Database:** 100 objects, each provides 72 images captured with 5 degree difference.

- **Setup:**
  - Dense sampling of overlapping 8×8 grids. Standard SIFT descriptor.
  - 4-level hierarchical k-means ($k = 10$): Leaf-node histogram is 1000-D.
  - Classifier via intersection-kernel SVM: 10 random training images per class.


COIL 100 Multiview Dataset (ntrain=10)
accuracy(%)
number of projection dimensions
100
95
90
85
80
75
70
65
60
55
50
45
0
100
200
300
400
500
600
raw data = 989 dim
random projection
sparse recovery (c=1)
sparse recovery (c=2)
sparse recovery (c=3)

Distributed Object Recognition in Band-Limited Smart Camera Networks

1. To harness the smart camera capacity, the system is separated in two components: **distributed feature extraction** and **centralized recognition**.

Distributed Object Recognition in Band-Limited Smart Camera Networks

1. To harness the smart camera capacity, the system is separated in two components: **distributed feature extraction** and **centralized recognition**.
2. **Gaussian random projection** as universal dimensionality reduction function: J-L lemma.

Distributed Object Recognition in Band-Limited Smart Camera Networks

1. To harness the smart camera capacity, the system is separated in two components: **distributed feature extraction** and **centralized recognition**.
2. **Gaussian random projection** as universal dimensionality reduction function: J-L lemma.
3. ℓ^1**-minimization** exploits two properties of SIFT histograms:
   - Sparsity.
   - Nonnegativity.

Distributed Object Recognition in Band-Limited Smart Camera Networks

1. To harness the smart camera capacity, the system is separated in two components: **distributed feature extraction** and **centralized recognition**.
2. **Gaussian random projection** as universal dimensionality reduction function: J-L lemma.
3. ℓ^1**-minimization** exploits two properties of SIFT histograms:
   - Sparsity.
   - Nonnegativity.
4. **Sparse innovation model** exploits joint sparsity of multiple-view histograms.

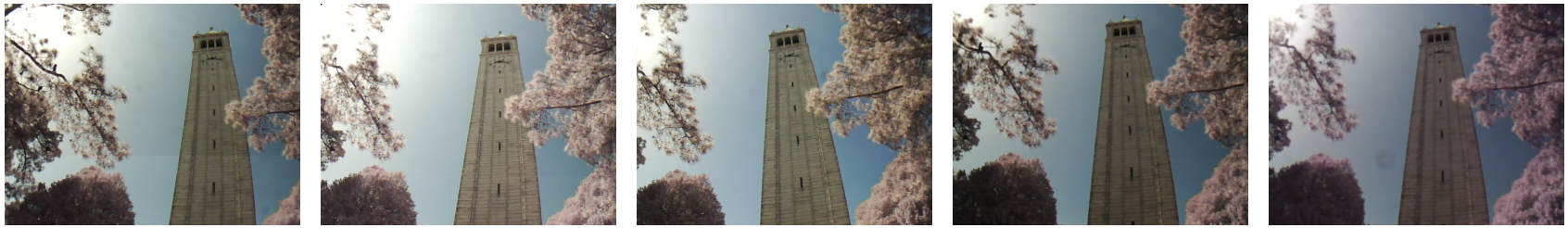
Distributed Object Recognition in Band-Limited Smart Camera Networks

1. To harness the smart camera capacity, the system is separated in two components: **distributed feature extraction** and **centralized recognition**.
2. **Gaussian random projection** as universal dimensionality reduction function: J-L lemma.
3. ℓ^1**-minimization** exploits two properties of SIFT histograms:
   - Sparsity.
   - Nonnegativity.
4. **Sparse innovation model** exploits joint sparsity of multiple-view histograms.
5. Complete system implemented on Berkeley CITRIC sensors.

References

- *Distributed Compression and Fusion of Nonnegative Sparse Signals for Multiple-View Object Recognition*. Information Fusion, 2009. (best paper award)
- *Multiple-View Object Recognition in Band-Limited Distributed Camera Networks*. ICDSC, 2009.

Berkeley Multiple-view Wireless Database



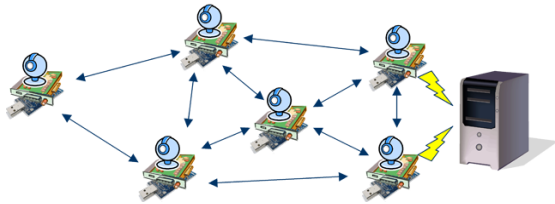
(a) Campanile

(b) Bowles

(c) Sather Gate

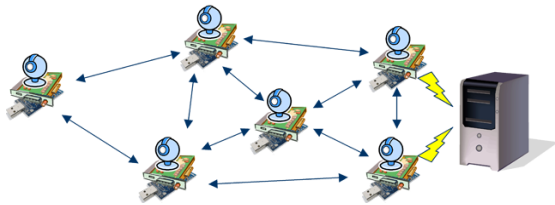
Sensing and Perception in Resource-Constrained Distributed Networks

Centralized Perception	
	
Up:	powerful processors
Up:	unlimited memory
Up:	unlimited bandwidth
Down:	single modality

Distributed Perception	
	
Down:	mobile processors
Down:	limited onboard memory
Down:	band-limited communications
Up:	distributed, multi-modality

Sensing and Perception in Resource-Constrained Distributed Networks

Centralized Perception	
	
Up:	powerful processors
Up:	unlimited memory
Up:	unlimited bandwidth
Down:	single modality

Distributed Perception	
	
Down:	mobile processors
Down:	limited onboard memory
Down:	band-limited communications
Up:	distributed, multi-modality

Whether the total network is greater than the sum of its parts?

Sensing ⇒ Perception ⇒ Action

1. Perception on Smart-Phone Architecture

2. Active Sensing

Sensing ⇒ Perception ⇒ Action

1. Perception on Smart-Phone Architecture

2. Active Sensing

3. Parallel Computing and Networked Computing Services

Figure: NVidia Tesla Solution: 240 core per GPU, up to 4 GPUs per server = 4 teraflops computing power.

Acknowledgments

Collaborators

- **Berkeley**: Dr. Shankar Sastry, Dr. Ruzena Bajcsy, Dr. Trevor Darrell, Dr. Jitendra Malik, Subhransu Maji, Mario Christoudas
- **UIUC**: Dr. Yi Ma, Dr. Shankar Rao, Hossein Mobahi
- **Cornell**: Dr. Philip Kuryloski
- **Tampere University of Technology:** Ville-Pekka Seppa
- **Telecom Italia**: Dr. Marco Sgnoi, Roberta Giannantonio, Raffaele Gravina

Funding Support

- ARO MURI: Heterogeneous Sensor Networks in Urban Terrains
- BAE Systems: Micro Autonomous Systems and Technology
- NSF TRUST Center at UC Berkeley