Cellphone as a Perceptual Platform for Micro UAVs

Nikhil Naikal

Action Webs
Perceptual Capabilities of Cell Phones

- Multiple tightly integrated sensors onboard.
  - Multiple cameras.
  - MEMS gyroscopes, accelerometers and digital compass.
  - Wireless and RF antennas.
  - Proximity and luminous intensity sensors.
  - Touch screen.
- Reasonable fast processing speeds and good memory.
- GPUs for parallel processing.
- Can potentially be used as main processing platform for small UAVs.
UAV Missions and Control

- US Army Description of most common missions performed by UAVs\(^1\):
  - Reconnaissance - Near real-time information about terrain, Search and rescue of friendly units, and disposition of possible enemy elements.
  - Surveillance - Area surveillance in friendly or enemy territory.
  - Situational Awareness - Provide commanders with situational awareness and mission planning information.
  - Security - Reaction time and maneuver space for the main body and area security.
  - Targeting - Target acquisition, target detection and recognition, target designation and illumination.
  - Communication Support - Voice and data communications retransmission.
  - Movement support - Convoy security, mines/IED detection.

- UAV Control
  - Currently 6 human operators for 1 UAV (Predator, Global Hawk, etc.)
  - Expert pilots for remote controller.

Computer Vision Algorithms for UAV Missions and Control

- Computer vision research directions:
  - Object detection/recognition.
  - Image/video segmentation.
  - 3D reconstruction/mosaicing.
  - Object tracking.

- Can computer vision algorithms be used for aiding in UAV missions?
  - Reconnaissance - Object detection/recognition, 3-D reconstruction.
  - Surveillance - Object detection/recognition.
  - etc.

- Can computer vision algorithms aid untrained personnel to control micro UAVs?
  - Control with commands such as, "follow road", "fly until objective reached".
  - Abstracting autonomous back-end from front-end human interface.
Multi-View Object Recognition

- Low cost cameras integrated with mobile platforms easily deployed.
  - Inter camera calibration usually not possible.
- Need to leverage multiple observations of objects from different vantage points.
- Problem Statement: I focus on recognition of common object over band limited communication channel.
Object Recognition - Overview

- Affine invariant features such as SIFT [Lowe 2002], SURF [Bay 2006], CHoG [Chandrasekhar 2009]

- Feature matching robust in harsh environments; popular for variety of applications.

- Scalable recognition with vocabulary tree [Nister 2006]
Visual Histograms

- Vocabulary tree constructed offline.
- All histograms are **nonnegative** and **sparse**.
- Multiple-view histograms share **joint sparse patterns**.
- Classification is based on a similarity measure.
CITRIC: Wireless Smart Camera Platform

- CITRIC platform [Chen 2008]

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. SURF detector: 10 fps.
Berkeley Multiple-view Wireless Database

- 20 landmarks at UC Berkeley.
- 16 different vantage points (large baseline); five images at one location (small baseline).
- Low-quality camera images: resolution, focal length, dusty lenses.
Training Phase

For each object category, $i = 1 \ldots C$, multiple histograms generated for all $j = 1 \ldots M$ training images, $Y_i = \{y_1, y_2, \ldots, y_M\}$.

All $C$ subsets form training set, $Y = \{Y_1, Y_2, \ldots, Y_C\}$. 
System Pipeline

- Invariant features extracted onboard.
- Visual histogram computed for image using stored vocabulary tree, and transmitted wirelessly.
- Functions on sensor largely stabilized, thereby facilitating deployment.
- Computationally heavy functions performed by the server, and can be updated.
Random Projection to Compress Histograms

\[ \mathbf{b} = \mathbf{A} \mathbf{x} \]

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

Advantages of Random Projection

1. Easy to generate and update
2. Does not need training prior; (universal dimensionality reduction).
3. Faster recognition speed.
Decoding via $\ell_1$-Minimization

**Noiseless case**
Assume $x$ is sufficiently $k$-sparse. Given triplet $(D, d, k)$ and random $A$ with $d > \delta(A)$ for some threshold $\delta$, solving

$$(P_1) : \min \|x\|_1 \text{ subject to } b = Ax$$

recovers the unique solution.

**Noisy case**
Assuming Gaussian measurement errors in $b$ with bound $\epsilon$, the solution to the convex program,

$$(P_{1,2}) : \min \|x\|_1 \text{ subject to } \|e\| = \|b - Ax\|_2 < \epsilon$$

recovers the sparsest solution.

Compressive sensing theory shows that under broad conditions, the estimates from $P_1$ and $P_{1,2}$ are the sparsest solution.
Why $\ell_1$-Minimization is still a difficult problem?

- General toolboxes do exist: cvx, SparseLab. However, interior-point methods are very expensive in HD space.

- Data Noise and Corruption

\[ \mathbf{b} = \mathbf{Ax} + \mathbf{e}, \quad \text{where } \|\mathbf{e}\|_2 \text{ may not be bounded!} \]

- Special structure of the data (from domain-specific knowledge)
**ℓ¹-Minimization using Iterative Soft-Thresholding (IST) [Donoho 1995]**

**Lagrangian method**

\[
x^* = \arg \min F(x) = \arg \min \frac{1}{2} \|b - Ax\|^2 + \lambda \|x\|_1
\]

\[
= \arg \min f(x) + \lambda g(x)
\]

- IST iteratively approximates the composite objective function

\[
x^{(k+1)} \approx \arg \min \{ f(x^{(k)}) + (x - x^{(k)})^T \nabla f(x^{(k)}) + \frac{\nabla^2 f(x^{(k)})}{2} \|x - x^{(k)}\|^2 + \lambda g(x) \}
\]

where the hessian \( \nabla^2 f(x) \) is approximated by a diagonal matrix \( \alpha I \).

- Denote \( u^{(k)} = x^{(k)} - \frac{1}{\alpha^{(k)}} \nabla f(x^{(k)}) \), then

\[
x^{(k+1)} \approx \arg \min \{ \frac{1}{2} \|x - u^{(k)}\|^2 + \frac{\lambda}{\alpha^{(k)}} g(x) \}.
\]

- When \( g(x) = \|x\|_1 \), a closed-form solution exists *element-wise*

\[
x_i^{(k+1)} = \arg \min_{x_i} \left\{ \frac{(x_i - u_i^{(k)})^2}{2} + \frac{\lambda |x_i|}{\alpha^{(k)}} \right\} = \text{soft}(u_i^{(k)}, \frac{\lambda}{\alpha^{(k)}})
\]
More References

1. **Primal-Dual Interior-Point** Methods

2. **Homotopy** Methods:
   - Polytope Faces Pursuit (PFP) [Plumbley 2006]
   - Least Angle Regression (LARS) [Efron-Hastie-Johnstone-Tibshirani 2004]

3. **Gradient Projection** Methods
   - Gradient Projection Sparse Representation (GPSR) [Figueiredo-Nowak-Wright 2007]
   - Truncated Newton Interior-Point Method (TNIPM) [Kim-Koh-Lustig-Boyd-Gorinevsky 2007]

4. **Iterative Thresholding** Methods
   - Soft Thresholding [Donoho 1995]
   - Sparse Reconstruction by Separable Approximation (SpaRSA) [Wright-Nowak-Figueiredo 2008]

5. **Proximal Gradient** Methods [Nesterov 1983, Nesterov 2007]
   - FISTA [Beck-Teboulle 2009]
   - Nesterov’s Method (NESTA) [Becker-Bobin-Candés 2009]

   - YALL1 [Yang-Zhang 2009]

References:
http://www.eecs.berkeley.edu/~yang/software/l1benchmark/
Joint Decoding

- Multi-view scenario gives rise to Sparse Innovation Model (SIM):

\[
\begin{align*}
  x_1 &= \tilde{x} + z_1, \\
  \vdots \\
  x_L &= \tilde{x} + z_L.
\end{align*}
\]

\(\tilde{x}\) is called the joint sparse component, and \(z_i\) is called an innovation.

- Joint recovery of SIM

\[
\begin{bmatrix}
  b_1 \\
  \vdots \\
  b_L
\end{bmatrix} =
\begin{bmatrix}
  A_1 & A_1 & 0 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \ddots & \ddots \\
  A_L & 0 & \cdots & 0 & A_L
\end{bmatrix}
\begin{bmatrix}
  \tilde{x} \\
  z_1 \\
  \vdots \\
  z_L
\end{bmatrix}
\]

\[\iff b' = A'x' \in \mathbb{R}^{dL}.\]

- Joint sparsity \(\tilde{x}\) is automatically determined by \(\ell^1\)-Minimization.
Multi-View Classification

- Multi-view relevance score assigned to query category as,

\[ m(X, Y_i) = \text{median}_{x_k \in X} s(x_k, Y_i), \]

where,

\[ s(x_k, Y_i) = \min_{y_j \in Y_i} \frac{x_k}{\|x_k\|_1} - \frac{y_{ij}}{\|y_{ij}\|_1} \].

Nikhil Naikal
Cellphone for UAV perception
Small Baseline Experiments

Table: Small-baseline recognition rates without histogram compression. The best rates are marked in bold face.

<table>
<thead>
<tr>
<th>Expt.</th>
<th># Train Images</th>
<th># Test Images</th>
<th>SIFT Rate(%)</th>
<th>SURF Rate(%)</th>
<th>CHoG Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Cam</td>
<td>160</td>
<td>160</td>
<td>71.25</td>
<td>80.62</td>
<td>81.88</td>
</tr>
<tr>
<td>2 Cam</td>
<td>160</td>
<td>320</td>
<td>72.5</td>
<td>81.25</td>
<td>84.38</td>
</tr>
<tr>
<td>3 Cam</td>
<td>160</td>
<td>480</td>
<td>73.75</td>
<td>81.88</td>
<td>86.25</td>
</tr>
</tbody>
</table>
Large Baseline Experiments

![Graph showing recognition rates for different projection dimensions](image)

**Table:** Large-baseline recognition rates without histogram compression. The best rates are marked in bold face.

<table>
<thead>
<tr>
<th>Expt.</th>
<th># Train Images</th>
<th># Test Images</th>
<th>SIFT Rate(%)</th>
<th>SURF Rate(%)</th>
<th>CHoG Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Cam</td>
<td>160</td>
<td>160</td>
<td>71.25</td>
<td>80.62</td>
<td>81.88</td>
</tr>
<tr>
<td>2 Cam</td>
<td>160</td>
<td>320</td>
<td>76.88</td>
<td>88.13</td>
<td>93.75</td>
</tr>
<tr>
<td>3 Cam</td>
<td>160</td>
<td>480</td>
<td>83.13</td>
<td>90.00</td>
<td>94.88</td>
</tr>
</tbody>
</table>

Nikhil Naikal
Cellphone for UAV perception
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. Multiple view information boosts recognition rates.

3. Drawn from Compressive Sensing theory to formulate distributed codec scheme.

4. Wireless cameras need not be calibrated. Further, system flexible to addition/omission of cameras and mobile platforms.
Future work (near future)

- Extending to video sequences.
  - Scenario: Car broken down in mountains, needs to be found. UAV on "detection" mode, to find car.
  - Multiple images obtained from video stream.
  - Signal has slowly varying sparse support.
  - Developing mathematical methods to speed up $\ell^1$-minimization for time varying sparse signal.

- Multiple camera images to recover 2.5-D or 3-D maps.
  - Sparse support represents common features between multiple images.
  - Using structure from motion methods to recover 3-D representations.

- Identifying "good" features in the training process using geometric relationships between training images.
  - Developing methods to identify strong visual features in the training process.
  - This will potentially make visual histograms more sparse.