Semantically-Aware Visual Navigation

Han-Pang Chiu

Center for Vision Technologies
SRI International, Princeton, NJ, USA.
Email: han-pang.chiu@sri.com

SRI International
Agenda

- Plug-and-Play Metric Navigation
- Semantically-Aware Visual Navigation
  - Hybrid Navigation, Mapping, and Geo-Registration
  - SAANE: Semantically-Aware Attentive Neural Embeddings
- Computation Hardware – NVIDIA JETSON TX2
About SRI International

• An independent nonprofit research center merging two world-class research institutions
  • Stanford Research Institute founded in 1946
  • RCA/Sarnoff Labs founded in 1941
• $450+ million in annual revenue
• 1,800 employees, 40 labs, 20 locations
• 4000+ patents
• 500 research projects annually
• 100+ commercial projects/year
• 70+ startups spun-out in 10 years
Autonomous Vehicles

Navigation & Mapping

Augmented & Virtual Reality

DOD Operations in GPS-Denied

Cars

Drones

Robots

Appliances

VR Headsets

AR Headsets

Drones

Urban Operations

3D Modeling Outdoors & Indoors
Factor Graphs for Multi-Sensor Navigation and Mapping

- We introduce a plug-and-play factor graph framework to encode sensor measurements with different frequencies, latencies, and noise distributions.

- Unary Factor: \( z = h(x_i) + v \)
- Binary Factor: \( z = h(x_{i-1}, x_i) + v \)
- Extrinsic Factor: \( z = h(x_i, l_j) + v \)
- IMU Factor: Summarize multiple readings between two pose states.

\[ \begin{align*}
\text{Unary Factor:} & \quad z = h(x_i) + v \\
\text{Binary Factor:} & \quad z = h(x_{i-1}, x_i) + v \\
\text{Extrinsic Factor:} & \quad z = h(x_i, l_j) + v \\
\text{IMU Factor:} & \quad \text{Summarize multiple readings between two pose states.}
\end{align*} \]

* Note this formulation covers full-SLAM problems, which keep all past states for optimization.


**GOAL:** Achieve GPS-level performance under all application scenarios (DARPA ASPN)

- Eliminate GPS as single point of failure
- Provide redundant capabilities and architectures with no single point of failure
- Provide Plug-and-Play, adaptable solution for multiple platforms and sensors

<table>
<thead>
<tr>
<th>Shipboard (Jan. 2015)</th>
<th>Leader-Follower Ground Vehicles (December 2015)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demo Lead: Navy: SSC-PAC</td>
<td>Demo Lead: Army CERDEC</td>
</tr>
<tr>
<td>3 week test in Pacific Ocean aboard the USNS Waters</td>
<td>3 day test on two Humvees at Aberdeen</td>
</tr>
<tr>
<td>USNS Waters</td>
<td>Demonstrate GPS denied collaborative navigation.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Demo Lead: Army CERDEC</td>
<td>Demo Lead: AFRL</td>
</tr>
<tr>
<td>3 day outdoor/ indoor tests at Aberdeen</td>
<td>Three-week live DC3 and S3 flights at Ohio.</td>
</tr>
<tr>
<td>Demonstrated scalability to low-SWAP, wearable processor platforms</td>
<td>Demonstrated GPS-denied navigation over different heights and terrains.</td>
</tr>
</tbody>
</table>

GPS-Denied Collaborative Navigation and Mapping

Dynamically created maps allows the system to ensure low drift errors in subsequent runs

- On the fly Creation of Visual Maps
- Match to Dynamic Maps to reset Drift
- Create Maps Across Multiple Platforms or Runs

Navigation sensors: IMU, Cameras, GPS, Magnetometer, Barometer, Ranging Radios

SRI CAMSLAM
Robot Localization in Warehouse Environment
Robotic Applications

NIOSH Visual-Inertial Drone Navigation - Recorded Live Demo Video (8x speed)

Scotty Labs (bought by DoorDash): Autonomous Cars

Area-17: Warehouse Drones
Visual Navigation

- Existing large-scale navigation solutions:
  - High-Precision Differential GPS with High-End IMU
  - Non-Differential GPS
  - Pre-Built Geo-Referenced Map Using LIDAR

- Limitation to visual navigation:
  - The real world is very dynamic with lots of moving platforms and continuous scene changes.
  - Maps need to be continuously updated.
  - System needs to be adaptive to appearance changes across weather, time, and illumination variations.
Navigation Using Pre-Mapped Visual Landmarks

<table>
<thead>
<tr>
<th>3D RMS Error (meter)</th>
<th>2015 Winter Partly Cloudy Noon</th>
<th>2016 Spring Cloudy Afternoon</th>
<th>2016 Spring Sunny Morning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Navigation Solution</td>
<td>0.5378</td>
<td>0.9440</td>
<td>1.1200</td>
</tr>
</tbody>
</table>

- Note we use INS solution (IMU, high-precision differential GPS) as ground truth.

Princeton City: 10.14km, 25 min

---

A hybrid approach: Develop and integrate both semantic-inference and metric-inference for navigation and mapping

**Accuracy**
High-level objects are more robust to scene changes, and can be matched across time/space/platforms.

**Efficiency**
Sharing semantic information reduces bandwidth required for collaboration/storage.

**Applicability**
It enables semantic scene representation and natural human-machine interaction for more applications.
Semantic Segmentation

- Our framework is designed to incorporate semantic information from any pre-trained networks that generate dense segmentation pixel labels on video frames.
  - We take the popular trained SegNet Encoder-Decoder Network as an example.

- We improve the efficiency of semantic segmentation while still maintaining its accuracy by converting the model into a low rank approximation of itself.
  - We extend the method* to apply both convolution and deconvolution layers.
  - The segmentation time is improved from 160 ms to 89 ms (almost 2x performance for >10Hz speed) to process one image on a single Nvidia K40 GPU.

Semantic Visual Landmarks

- Using mapped landmark from permanent semantic classes is able to **improve overall navigation accuracy between 8%~20% while reducing 30%~50% of storage.**

Improved Feature Tracking for Navigation

- Using consistent feature tracks from static semantic classes reduces 0.1~0.5% drift rate using ORB-SLAM2 over 11 KITTI sequences.
- The following example shows final error is reduced from 12.64 meters to 3.93 meters over 1.71 km.

Real-Time Semantic Mapping and Modeling

Fuse semantic labels for 3D dynamic mapping and modeling in new environments.
Revisit – Search Visual Map Database

Matching query images to geo-referenced databases across appearance changes, viewpoint changes, and scene changes.

$$i^* = \arg \max_{i \in 1 \ldots N} s(Q, V_i)$$

Nearest Neighbor search
Goal: Estimate position for a given monocular query image by matching to a database of images of known locations, in the presence of large appearance changes across weather and illumination variations.

We propose to use embedding for this problem: A deep-learned compact Euclidean space where distances directly correspond to a measure of data similarity.

Semantically-Aware Visual Localization

- Estimate position for a given monocular query image by matching to a database of images of known locations.

- **Semantic-Aware:** Our deep network model incorporates pixel-wise semantic features in learning the image embeddings.

- **Attention-Based:** We train self-attention modules to encourage our model to focus on semantically consistent spatial regions.

- Training data: ~2 million images collected from 2,685 static webcams.

- **Our accuracy outperforms STOA methods by 19%.

SAANE: Semantically-Aware Attentive Neural Embeddings

- **SAANE**: Modeling a joint appearance-semantic image embedding space.
- **Training**: Collect ~2,000,000 images from 2,079 webcams taken at multiple times of year.
- **Test**: Several public navigation datasets, wholly distinct from the training data.
- **Evaluation**: Area Under PR Curve: A curve of precision % across different recall % for all data.

Evaluation Results

- Our baseline achieves comparable results to state-of-the-art methods.
- Adding semantic cue helps challenging tasks, such as 25% at Day->Night.
- Multi-modal attention gives best performance.

Area Under PR Curve: A curve of precision % across different recall % for all data.

<table>
<thead>
<tr>
<th>Model</th>
<th>NordLand: Sum.-&gt;Win. 51 km</th>
<th>St. Lucia 17 km</th>
<th>RC: Sum.-&gt;Win. 2 km</th>
<th>RC: Day-&gt;Night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance (Our Baseline)</td>
<td>58.0</td>
<td>71.4</td>
<td>79.6</td>
<td>36.6</td>
</tr>
<tr>
<td>DenseVLAD*</td>
<td>19.2</td>
<td>64.5</td>
<td>78.6</td>
<td>22.9</td>
</tr>
<tr>
<td>AMOSNet*</td>
<td>45.6</td>
<td>75.3</td>
<td>75.1</td>
<td>45.3</td>
</tr>
<tr>
<td>Appearance + Semantic</td>
<td>74.6</td>
<td>74.9</td>
<td>87.9</td>
<td>61.4</td>
</tr>
<tr>
<td>Appearance w/ individual Attention</td>
<td>74.2</td>
<td>74.0</td>
<td>86.5</td>
<td>60.9</td>
</tr>
<tr>
<td>Appearance + Semantic w/ Individual Attention</td>
<td>68.5</td>
<td>73.9</td>
<td>83.6</td>
<td>62.9</td>
</tr>
<tr>
<td>SAANE (Appearance + Semantic w/ Multi-modal Attention)</td>
<td><strong>77.3</strong></td>
<td><strong>78.3</strong></td>
<td><strong>88.5</strong></td>
<td><strong>67.3</strong></td>
</tr>
</tbody>
</table>

*DenseVLAD and AMOSNet are state-of-the-art 2D image-based methods for geo-localization and database retrieval.*
Visualization of Localization Results - NordLand
Localization Across Seasons

- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.
Localization Across Day & Night

- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.
Refining with Location Priors

- **Recall at K (R@K)** - Percentage of test samples for which correct result is ranked within top-K retrieved results to the query. **Median Rank (MedR)** - Median of ground-truth matches in the ranking.

<table>
<thead>
<tr>
<th>Location</th>
<th>No prior</th>
<th>5 km</th>
<th>1 km</th>
<th>500 m</th>
<th>100 m</th>
<th>50 m</th>
<th>10 m</th>
<th>5 m</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
<td>MedR</td>
<td>R@1</td>
<td>R@10</td>
<td>MedR</td>
<td>R@1</td>
<td>R@10</td>
</tr>
<tr>
<td>Nordland (51 km)</td>
<td>47.08</td>
<td>85.16</td>
<td>11</td>
<td>61.84</td>
<td>81.27</td>
<td>25</td>
<td>62.38</td>
<td>81.52</td>
</tr>
<tr>
<td>RobotCar (2 km)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Recall at K (R@K)** - Percentage of test samples for which correct result is ranked within top-K retrieved results to the query. **Median Rank (MedR)** - Median of ground-truth matches in the ranking.
Semantic Geo-Registration

We perform 2D-3D geo-registration continuously between the input video frame and the matched geo-registration 2D-3D database (from cameras or LIDAR).

3D Object Hypothesis Verification

- 3D ground plane can be fitted based on depth map by propagating road labels.
- 3D information of objects are estimated based on verified 3D class hypotheses from segmented regions on 3D ground plane.

\[ p(t = \text{true}|m, G) \propto p(t = \text{true}|d) = N(d; 0, \sigma_d), \quad r \approx \frac{c \cdot f}{I}, \quad o \approx \frac{r}{\sqrt{(\frac{u}{f})^2 + (\frac{v}{f})^2 + 1}} \left( \frac{\frac{u}{f}}{\frac{v}{f}} \right), \quad d = O^T n + h, \]

d: 3D dist. of obj. to ground, I: 2d obj. height, O: 3d obj. position, n: ground normal, h: cam. height
Depth Estimation and Occlusion Reasoning for Distant Objects

- Depth information for distant objects can be estimated (up to 120 meters based on minimum detection). This method can be used by propagating on-the-fly nearby stereo depth maps or prior depth maps.
Semantic Depth Reasoning for Static Background and Dynamic Objects

Input: Monocular Video
Output: Overlaid Depth Video

Demonstration: Augmented Reality on Training Field

Computation Hardware: NVIDIA JETSON TX2

- Demonstrated GPS-denied navigation
- Past Performance:
  - DARPA ASPN, Army AR-Stryker:
    - plug-and-play sensing
    - visual tracking and dynamic mapping
  - ONR AITT & WAR, DARPA Squad-X:
    - semantic geo-registration
  - MCWL, DARPA ASPN, DARPA Squad-X
    - collaborative navigation through ranging radios