Navigation

Saurabh Gupta
Semantic Reasoning for Visual Navigation

Robot with a first person camera
Dropped into a novel environment
Discover paths or Explore

“Go 300 feet North, 400 feet East”
“Go Find a Chair”
“Explore the Environment”
Classical Solution

Observed Images → Mapping → Planning → Geometric Reconstruction → Path Plan

Hartley and Zisserman. 2000. Multiple View Geometry in Computer Vision


Video Credits: Mur-Artal et al., Palmieri et al.
Geometric 3D Reconstruction of the World

Do we need to tediously reconstruct everything on this table?

**Video Credit:** Mur-Artal and Tardos, TRobotics 2016. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras.
Spatial Reasoning in Classical SLAM +

Common Sense or Semantics via Learning
Explicit Semantics

Going to a Image Goal

Figure 1: Semantic Priors and Landmarks.

When asked to go to target image of an oven most humans would use the path number 2 since it allows access to kitchen. Humans use semantic priors and common-sense to explore and navigate everyday yet most navigation algorithms struggle to do so.
Implicit Semantics

Finding a bathroom in a new restaurant
Implicit Semantics

Speculating about space not directly observed.
Implicit Semantics

Exploiting patterns in layout of indoor spaces.
Action to Execute Goal (300, 400)

Mapper

Spatial Representation of the World

Planner

Neural Network

Action to Execute

Mapper

Planner

Goal (300, 400)

Egomotion

Action to Execute
- Mapper and planner are differentiable functions
- Mapper and planner are learned for end task
- Hand-crafted obstacle maps to task-driven semantic maps
Spatial Representation

Feature $f_t$

Confidence $c_t$

Egocentric Bird’s Eye Coordinate Frame
Differentiable Mapper

\[ f_t \rightarrow \text{Differentiable Warping} \rightarrow f_{t+1} \]

\[ C_t \]
Goal
Action to
Execute

Differentiable
Mapper

Differentiable
Planner

Egomotion
Differentiable Planner

Differentiable Planner

Differentiable Planner

Local computation that can be done using **Convolutions** and **Channel-wise Max-Pooling**.

Tamar et al. NIPS 2016. *Value Iteration Networks.*
Differentiable Planner

Local neighborhood tells about optimal action

Policy Training

Simulator based on scans of Real World Environments

Simulate robot views and motion

Compute ground truth traversability

Armeni et al. CVPR 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces
Policy Training by Expert Imitation

Train with back-propagation
Point Goal Task

Goal: Straight 5m
Results (Novel Env., Go To Relative Offset)

**Depth Input**

<table>
<thead>
<tr>
<th></th>
<th>Reactive</th>
<th>LSTM</th>
<th>Classical</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**RGB Input**

<table>
<thead>
<tr>
<th></th>
<th>Reactive</th>
<th>LSTM</th>
<th>Classical</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Goal \((chair, table, door)\)

Goal \((300, 200)\)
Semantic Tasks (Go to a chair)

Successful Navigations by CMP (Semantic Task)
Results (Novel Env, Go To Object)

- **Depth Input**
  - Reactive
  - LSTM
  - Classical
  - Our

- **RGB Input**
  - Reactive
  - LSTM
  - Classical
  - Our
Agent can make predictions about its surroundings

Free Space
Agent can make predictions about its surroundings.
Agent can make predictions about its surroundings
Agent can make predictions about its surroundings
Agent can make predictions about its surroundings
Representation for Spaces

- Semantic reasoning
- Spatial reasoning
- Robustness to actuation noise
- Easy to acquire

More recently,
- Topological representations (robustness to noise)
- Ease of training
  - Modular approaches
  - Training via supervised learning
  - Training without interaction, using Internet Videos
$F_G$ = Geometric Prediction: Free space in different directions

“Ghost Nodes” 0.8

$F_S$ = Semantic Prediction: Closeness to target

$F_L$ = Localization

$F_R$ = Relative Pose

4 Functions

\[ F_S(I_1, I_2) = \text{Semantic Prediction: Closeness to target} \]
\[ F_G(I_1) = \text{Geometric Prediction: Free directions} \]
\[ F_R(I_1, I_2) = \text{Relative Pose} \]
\[ F_L(I_1, I_2) = \text{Localization} \]
High-Level Policy

Value Predictions via $F_S(I, c)$.

Short-term Goal

Past Nodes

Current Node

ΔPose

Low-Level Policy

Occupancy Map

FMM Cost Map

Hierarchical Policy

Forw.

Left

Right

Stop

Execution (Image Goal)

Observation

Goal Image

Topological Map and Pose

Source Image ($I_s$)

|   |   | 0.05 | 0.05 |   | 0.91 | 0.87 |   | 0.21 | 0.16 |   |   |
## Results

<table>
<thead>
<tr>
<th>End-to-end Learning</th>
<th>RGB</th>
<th>RGBD</th>
<th>RGBD (No Noise)</th>
<th>RGBD (No Stop)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + Imitation</td>
<td>0.10</td>
<td>0.14</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>LSTM + RL</td>
<td>0.10</td>
<td>0.13</td>
<td>0.14</td>
<td>0.17</td>
</tr>
<tr>
<td>Occupancy Maps + FBE + RL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANS</td>
<td>0.23</td>
<td>0.29</td>
<td>0.35</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>NTS (Our)</strong></td>
<td>0.38</td>
<td>0.43</td>
<td>0.45</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Map based methods are better than vanilla learning methods even in presence of noise.

Results

Semantic score function improves efficiency when no prior experience with environment is available.

As experience in environment increases, utility of semantic function decreases.

But, at the same time, importance of the topological representation increases.

Thanks!