Robots, Objects and Space

• **Goal:**
  classify objects and build 3D object spatial models.

• **Robots will need to understand and predict the 3D object layout of human houses.**

• **Applications:**
  – Object recognition
  – Fetching and cleaning tasks
  – Identifying 3D relationships from 2D data
  – Interpreting human commands
Maps, objects, activities

A home robot should know about homes:
  kitchen, living room, playroom, bathroom, bedroom, patio

These are *places*: sites for activities:
  cooking, talking, xbox, showering, sleeping, partying

With assumptions (priors) over the locations of objects
  – e.g., it’s unlikely the barbecue is in the bedroom.

Our challenge:
  to connect object’s names with appearances and shapes
to link the robot’s maps to activities of our world
to enable the robot to work in our world
The robot will know about a prototypical home:

- kitchen,
- living room,
- playroom,
- bathroom,
- bedroom,
- patio

These are *places*: sites for activities that accomplish tasks.

With assumptions (priors) over the properties — e.g., it’s unlikely the barbecue is in the bedroom.

Our challenge:

- to connect names with appearances and shapes
- to link the robot’s maps to activities of our world
- to enable the robot to work in our world
Motivation

- Semantic concepts for interaction with impaired users of technology and assistive technology applications in general
- High-level task planning based on human concepts
Overview

- Objects structure space in the context of tasks
- The Semantic Robot Visual Challenge
- Objects and Places
- Cluttered Scenes

He served sushi at a reception. He:
- found people
- asked if they wanted sushi
- moved on to find other people
- when out of food,
  - he found his way back to home base to load more sushi
  - without bumping into people!
From stereo to maps
Local map generated by autonomous exploration
Simultaneous Localization and Mapping (SLAM): determining camera viewpoint and landmark position, providing a map that supports local and global localization.

In order to collaborate with other robots and humans – its partners – a robot needs to determine its location in a map.

SLAM provides a geometric map built from observed, distinguishable landmarks.

We use visual features (SIFT) from a stereo camera to build re-usable maps.
Scale Invariant Feature Transform (SIFT)

- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters.
SIFT-based localization

SIFT features: scale, orientation

SIFT stereo: distance indicated by size
Occupancy grids and landmarks
Building a map
Overview

- Structuring space
- The Semantic Robot Visual Challenge
- Objects and Places
- Cluttered Scenes
- Summary
A robot is given a list of names of objects, both particular and generic, which it must find in a small test room.

The robot can use its broad knowledge about object classes to find the objects.

Or it could then download images of the named objects from the Web and construct classifiers to recognize those objects.

The robot enters the test area and searches for the objects in a small test room. The robot returns a list of the objects' names and the best image matching the name. In each image it must outline the object with a tight-fitting rectangle.
Phases of the SRVC

Web Search

Model formation

Search and discovery
SRVC Phases: Web search


- shoe
- umbrella
- robosapien
The banana problem
Google images are re-ordered by a ranking algorithm using visual and URL cues.

Features for obtained images are computed.
**Model Formation**

SIFT-based Image Matching

- SIFT features (Scale Invariant Feature Transform)
- Feature consistent cliques in training set are found.
- Training set, and exploration phase photos are matched.
- Geometric consistency checking through voting
Exploration:
collect images of objects
Return best candidate image for each category.
SRVC Phases

- **Training phase**: Web-crawling and classifier learning.
- **Exploration phase**: Collect photos of potential objects.
- **Classification phase**: Return best candidate image obtained for each category.
Embodied Visual Search

- State-of-the-art object recognition performs well on static databases
- Object recognition faces numerous challenges:
  - Scene clutter
  - Drastically different viewing angles
  - Large variation in scale
  - Poor optical sensing
- An embodied system faced numerous additional challenges:
  - Time constraints on operation
  - Navigation and coverage
Hardware Platform

Fast saliency computation (0.1 sec/frame) based on spectral saliency (Hou et al. CVPR07) and MSER (Matas et al. BMVC02)
Multi-Scale Saliency
Spatial Semantic Maps
Coverage and viewpoint planning are achieved using map information.
SIFT-based Image Matching
CD “Hey Eugene” by Pink Martini
DVD “Gladiator”
pepsi bottle
red bell pepper
red plastic cup
banana
Drive System Damage
Detection Results
Detection Results
Curious George Generations
Lest those without robots despair, there is a software competition where the organizers act as “robot image capture devices” and grab random images of objects:
Histogram of oriented gradients
Deformable Parts Model
Learned Models of bowls from LabelMe images
Results 2010

Scoring:
- 8 of 12 specific instances
- 4 of 8 generic categories

Observations:
- Simple environment makes 3D segmentation very effective
- Most instances missed due to bad pictures
- Categories missed since learning generic mode from web data is hard!
Currently, Curious George uses a hierarchical planner that adaptively select between exploring new areas and acquiring more views of previously seen objects.

- Does not use any semantic information.
- Robot ends up spending a lot of time exploring areas that are unlikely to contain the query object.
Overview

- Structuring space
- The Semantic Robot Visual Challenge
- Objects and Places
- Cluttered Scenes
- Summary
Spatial-Semantic Model

uses spatial and semantic information about objects (places they usually occur in and their observed locations) to determine their cluster and place labels.

employs semantic information about objects (places they usually occur in) to determine their corresponding place labels.

determines likely locations of objects by exploiting their semantic information as well as information about spatial-semantic clusters on the map.

Automated Spatial-Semantic Modeling with Applications to Place Labeling and Informed Search (Computer and Robot Vision)
Place Classification

‘Kitchen’
- drawer, oven, pot,
- stool, stove, table top

‘Unknown’
- bathtub, armchair, bed, ceiling,
- chair, door, lamp, molding, phone,
- pillow, vent, wardrobe and
- window
Cluster Model

Group objects based on their place types and spatial locations
Place Classification

- Infer place type based on the objects annotated in a LabelMe image
- For each test image, compute the most likely place type conditioned on the object annotations
- Four scenes: Kitchen (176), Bedroom (37), Bathroom (31), Office (824)

### Confusion Matrix

```
       kitchen  bathroom  bedroom  office  unknown
kitchen    0.88      0.00     0.00     0.01     0.00
bathroom  0.13      0.84     0.00     0.00     0.00
bedroom  0.03      0.00     0.94     0.04     0.00
office   0.00      0.00     0.00     1.00     0.00
```

<table>
<thead>
<tr>
<th>Room Type</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Bathroom</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td>Office</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Rows: Ground Truth
Col: Prediction
Location Model

- Find the towel
Realistic robot simulator (based partly on Player/Stage) developed during preparation for the SRVC competition, has basic collision avoidance and path planning capabilities.
Realistic robot simulator (based partly on Player/Stage) developed during preparation for the SRVC competition, has basic collision avoidance and path planning capabilities.
Informed Search

- Simulate robot's camera and record frequency with which planned paths capture a view of the query object.
- Above information is averaged over 50 trials of 50 planning steps each, for 2500 total robot poses.
- Between each trial, initial robot location and query object are selected at random and each of the two planning methods is evaluated.
Informed Search

![Graph showing the cumulative average number of views of query objects for different planning strategies. The graph compares Coverage Based Planning and Informed Search Planning.](image-url)
Scene Understanding with a Robot
Active Object Recognition for a Mobile Platform  IROS
Scene Understanding in Clutter
Challenges of Clutter
Task 2 – Scene Understanding with a Robot
Multi-View Motivation
Finding Bounding Volumes

Single view Result

Multi-view Result

precision

recall

Single view

Multiview
Multi-view Background

- **Specific Objects:**
  - [David Lowe CVPR 2001]
  - [Fred Rothganger et al. IJCV 2006]

- **Category Objects:**
  - [Bastian Liebe CVPR 2005 Multi-view ISM]
  - [Silvio Savarese et al. ICCV 2007 multi-view model and dataset]
  - [Alexander Thomas et al. IJRR 2009]
  - [Min Sun et al. ICCV 2009]

Image from [Savarese et al. 2007]
Mobile 3D Object Detection in Clutter
Mobile robot with fiducial
Depth and inference

We handle each pixel using the sensed depth information:

1) The sensed depth is closer indicating the object is occluded.
2) The sensed depth falls within volume, so the object is foreground.
3) The depth is farther than the volume, indicating the laser has passed through the volume and it is unoccupied.
Objects and occlusions
Learned model of effect of occlusion
Inference from multiple views
Test Time Multi-view
Finding Bounding Volumes

Single view Result  Multi-view Result

Precision-recall curve for Single view and Multi-view results.
Results
Improvements
Multiple views
Explicit Occlusion Reasoning for 3D Object Detection  BMVC
Partial detectors
Figure 4: Performance of full and partial detectors for (a) mugs and (b) bowls.
Unlike previous, we use data-driven sampling, conditioned by the detections in the images, then scaled to a bounding volume by the sensor depth.

The responses of the full and partial detectors are combined by a mixture of experts.

Location of the bounding volume is improved by gradient descent.
Results

[Graph showing recall vs. 1-precision with three lines representing different models: DPM bowl (0.68), Bowl multiview w/occlusion (0.74), Bowl multiview w/occlusion w/partials (0.84).]
Detections in 3D
Willow Garage PR2
Questions?

- More information on the “Semantic Robot Vision Challenge” and “Curious George Robot” accessible via Google

- Recent work at British Machine Vision Conference: Explicit Occlusion Reasoning for 3D Object Detection