Acting and Interacting in Natural Environments

Danica Kragic, Jeannette Bohg, Dan Song, Javier Romero, Matthew Johnson-Roberson and Gabriel Skantze

Centre for Autonomous Systems
Computer Vision and Active Perception Laboratory
KTH, Stockholm, Sweden

IROS 2010 Workshop: Semantic Mapping and Autonomous Knowledge Acquisition, Taipei, Taiwan
A Point Cloud! And Now?

- From Stereo to Object Hypotheses
- Uncertainties

"Scene Representation and Object Grasping Using Active Vision", Gratal et al., IROS Workshop 2010
How do we Plan Grasping and Manipulation under Uncertainty?

- Example Tasks:
  - Prepare the dinner table!
  - Pour me a cup of coffee!
  - Clean the table!
  - Unload the dishwasher!

- Partially unsolved → challenges

- Robot needs to understand the environment (human activities, obstacles, objects and their poses etc.)

- Fill in the gaps in the knowledge e.g. scene model
Recognition of Objects and Pose Estimation
The Necessity of Geometric Scene Understanding

Example Tasks:
- Prepare the dinner table!
- Pour me a cup of coffee!
- Clean the table!
- Unload the dishwasher!

- Collision detection, reachability
- Pre-grasp manipulation, pushing objects in the scene
- Placing things at certain positions
- Free and occupied spaces need to be known
Multi-Modal Scene Exploration

- "Strategies for Multi-Modal Scene Exploration", IROS 2010
- Predict scene structure of unobserved spaces from the observed space
- Confirmation of this prediction through haptic exploration
- Scene representation:
  Occupancy Grid from Initial Stereo Reconstruction
- Scene prediction:
  Gaussian Processes
An Example on Synthetic Data

Figure: Example for the prediction of a 2D map from camera measurements using GPs.

- Prediction through a Gaussian Process
- Sampling of Known Grid Cells
- Squared Exponential Covariance Function
Exploration Strategies Compared

Goal: Minimise the number of explorative actions

- Spanning Tree Coverage
- Each cell gets explored once
- Active Learning Scheme with PRMs
- Minimise the uncertainty in the scene

Figure: Occupancy Grid After 250 Measurements

Figure: Occupancy Grid After 250 Measurements
Demonstration on the Robot

See www.csc.kth.se/~bohg/IROS2010Grasp.mp4
Experimental Results

1. Gaussian Process produces a valid scene prediction
   - **Task**: Classify each grid cell to be empty or occupied
   - Classification Performance in Occupancy Grid: 77%
   - Classification Performance in Predicted Map: 91% = **Increase of 14%**

2. Active Learning scheme produces a better scene prediction early on in the exploration process
**Scenes for Task Planning and Execution**

- **So far:**
  - Scene model suitable for planning manipulation and grasping
  - Free and occupied spaces
  - Representation of known and unknown objects

- **Example Tasks:**
  - Prepare the dinner table!
  - Pour me a cup of coffee!
  - Clean the table!
  - Unload the dishwasher!

- Given these tasks, grasps fulfilling specific constraints required

- One way: Learn from humans → Programming by Demonstration
Learning Task Constraints for Robotic Grasping

- Correspondence problem in imitation learning
  How to map the human grasp to the robot hand?

- **Task constraints:**
  Characterize task requirements
  Can be independent of embodiment

- If *task* can be *recognised from human demonstration*, then this *task* can be *performed by a robot* through its **own means**!
A Graphical Model for Learning Task Constraints

”Learning Task Constraints for Robot Grasping Using Graphical Models”, Song et al., IROS 2010

- Task label $T$
- Object Features $O$
- Action Features $A$
- Constraint features $C$
- Bayesian Network (BN) for modelling joint distribution of these variables
- Training BN with labeled training data
  
  1. What is the task the human is doing?
  2. Given a task, how should this object be grasped?
  3. How to perform for example pouring?
Given a task, how should this object be grasped?

<table>
<thead>
<tr>
<th></th>
<th>$T = \text{hand-over}$</th>
<th>$T = \text{pouring}$</th>
<th>$T = \text{tool-use}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hammer</td>
<td><img src="image" alt="Hammer" /></td>
<td><img src="image" alt="Hammer" /></td>
<td><img src="image" alt="Hammer" /></td>
</tr>
<tr>
<td>Bottle</td>
<td><img src="image" alt="Bottle" /></td>
<td><img src="image" alt="Bottle" /></td>
<td><img src="image" alt="Bottle" /></td>
</tr>
<tr>
<td>Mug</td>
<td><img src="image" alt="Mug" /></td>
<td><img src="image" alt="Mug" /></td>
<td><img src="image" alt="Mug" /></td>
</tr>
</tbody>
</table>
How to perform pouring?

**Step 1** Human demonstration: recognize task $t^*$

- $t_1 = \text{hand-over}$
- $t_2 = \text{pouriing}$
- $t_3 = \text{tool-use}$

**Step 2** Select object $o^*$: matching $t^*$, or also similar to $o^H$

**Step 3** Select action $a^*$: matching $t^*$, or also similar to $a^H$

**Reward Functions**

- $P^H(t \mid o^H, a^H, c^H)$
- $P^R(t_2 \mid o)$
- $P^R(t_2 \mid o) \cdot 0.2$
- $+ S(o, o^H \mid t_2) \cdot 0.8$
- $P^R(t_2 \mid o^*, a)$
- $P^R(t_2 \mid o^*, a) \cdot 0.2$
- $+ S(a, a^H \mid t_2) \cdot 0.8$

**Goal-directed imitation:**

- Achieving same task based on robot’s own motor capabilities.
From Synthetic to Real Data

- System on learning task constraints has been shown to work on synthetic data
- **Future Goal**: Apply it to Real Data
- Needed:
  1. Object features e.g. 2D/3D visual representation
  2. Action features → observation of human hands
Real-Time Hand Pose Estimation

See www.csc.kth.se/~jrgn/2010_ICRA_rkk.mpg
Database Composition

- Synthetic images generated with Poser™
- 5 timesteps of 31 different grasp types
- 648 viewpoints
- The images include a prototypical object in order to include typical occlusions
Hand tracking system

- **Appearance Likelihood**
  1. Skin-color hand segmentation
  2. HOG computation
  3. Database Nearest Neighbor search based on HOG
  4. Appearance Likelihood: Gaussian weight based on HOG distance for NN

- **Temporal Likelihood**: Kernel density estimation based on previous frame
- The likelihood of each pose is the product of temporal likelihood and appearance likelihood

J. Romero et al., Hands in Action: Real-Time 3D Reconstruction of Hands in Interaction with Objects, ICRA10
Real-Time Hand Pose Estimation

Improving temporal likelihood

1. The temporal likelihood should encapsulate human dynamics

2. Human demonstrations of the grasps in the database were recorded with a magnetic tracker

3. The mapping of those demonstrations to a lower dimensional space can be used to predict the next frame pose

"Spatial-Temporal Modelling of Grasping Actions" Romero et al., IROS 2010
A Short Re-Cap of the Talk

- So far:
  - Scene model suitable for planning manipulation and grasping
    - Free and occupied spaces
    - Representation of known and unknown objects
  - Task model taught by a human demonstrator
- Vision cannot give us everything! → wrong scene segmentation, wrong labels
- Can we bootstrap scene understanding by human input?
Human in the Loop

Enhanced Visual Scene Understanding through Human-Robot Dialog

See www.csc.kth.se/~bohg/Enhanced.mp4
How is the scene segmentation refined?

- **Initial Scene Segmentation**
- **Questions:**
  1. I can see \( n \) objects. Is this correct?
  2. Which segment is incorrect?
  3. How are the objects in the wrong segment positioned?

"Enhanced Visual Scene Understanding through Human-Robot Dialog", Johnson-Roberson et al, AAAI Fall Symposium 2010
ICRA 2011 Submission
Which segment is incorrect?

- **Segment analysis**: point and colour distribution

- **Observation**: Single objects are homogenous in their attributes → Undersegmented Regions are not → Captured by Entropy

- **SVM** to classify incorrect segments based on Feature Vector with Entropy Values
  - 264 segments in the database (127 incorrect, 137 correct)
  - Training on 25 incorrect and correct examples; Testing on 214 examples

- **Area under ROC Curve**: 98%
How are the objects in the wrong segment positioned?

- Query the user
- Three options:
  1. On top of each other
  2. Next to each other
  3. In front of one another
- Split the bounding box along the user specified axis
- Re-label initial segmented points and re-segment in an energy minimisation framework

"Attention-based Active 3D Point Cloud Segmentation", Johnson-Roberson et al., IROS 2010

"Mechanical Support as a Spatial Abstraction for Mobile Robots", Sjöö et al., IROS 2010
How much does the Initial Segmentation improve?
Conclusion

- Example Tasks:
  - Prepare the dinner table!
  - Pour me a cup of coffee!
  - Clean the table!
  - Unload the dishwasher!

- Vision is hard!
- Grasping is hard!
- Scene understanding through
  - Segmentation, Recognition and Classification
  - Multi-Modal Interaction (Speech, Haptic, Vision)

- Markerless understanding human actions
- Bayesian Learning for Modelling of Complex Tasks