Grounding Object Semantics in Multi-modal Interactions

Jivko Sinapov
Assistant Professor
Tufts University
www.cs.tufts.edu/~jsinapov
Motivation: why multi-modal?
Visual Object Perception in Robotics

Sridharan et al. 2008

Collet et al. 2009

Rusu et al. 2009

Lai et al. 2011
Why Exploratory Behaviors?
Why Exploratory Behaviors?
Solution: Lift the Object
Why Exploratory Behaviors?
Why Exploratory Behaviors?

![Image of medicine bottles indicating full and empty status]
Solution: Shake the Object
Solution: Shake the Object

Exploratory behaviors give us information about objects that vision cannot!
<table>
<thead>
<tr>
<th>KNOWLEDGE ABOUT OBJECT</th>
<th>EXPLORATORY PROCEDURE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Substance-related properties</strong></td>
<td><strong>Lateral motion</strong></td>
</tr>
<tr>
<td>Texture</td>
<td>Pressure</td>
</tr>
<tr>
<td>Hardness</td>
<td>Static contact</td>
</tr>
<tr>
<td>Temperature</td>
<td>Unsupported holding</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
</tr>
</tbody>
</table>

| **Structure-related properties** | **Unsupported holding** |
| Weight | |
| Volume | Enclosure, contour following |
| Global shape | Enclosure |
| Exact shape | Contour following |

| **Functional properties** | **Part motion test** |
| Part motion | |
| Specific motion | Function test |

[Power, 2000]  
[Lederman and Klatzky, 1987]
Object Exploration in the Wild

“A young corvide bird, confronted with an object it has never seen, runs through practically all of its behavioral patterns, except social and sexual ones.”

-Konrad Lorenz
Object Exploration in Infancy
Can a robot use auditory perception to recognize objects?

Sinapov et al. “Learning the Acoustic Properties of Household Objects”
2009 IEEE Int’l Conf. on Robotics and Automation
Acoustic Object Recognition

Auditory Data

Dimensionality Reduction

Discrete Auditory Sequence

Object Probability Estimates

Auditory Recognition Model
Object Recognition Results

Fig. 6. Object recognition performance with k-Nearest Neighbor as the number of interactions with the object is varied from 1 (the default, used to generate Table I) to 5 (applying all five behaviors on the object).
Scaling up: more objects + the proprioceptive sensory modality

Sinapov et al. “Interactive Object Recognition using Proprioceptive and Auditory Feedback”
Exploratory Behaviors

Lift:

Drop:

Shake:

Crush:

Push:
The Proprioceptive / Haptic Modality
Feature Extraction

Training a self-organizing map (SOM) using sampled joint torques:

- Set of Joint-Torque records:
- Sampled Vectors:
- Trained SOM:

Training an SOM using sampled frequency distributions:

- Set of Spectrograms:
- Column Vectors:
- Trained SOM:
- GHSOM Toolbox
Feature Extraction

Discretization of joint-torque records using a trained SOM

Discretization of the DFT of a sound using a trained SOM
Sensorimotor Contexts

<table>
<thead>
<tr>
<th>Behaviors</th>
<th>audio</th>
<th>proprioception</th>
</tr>
</thead>
<tbody>
<tr>
<td>lift</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>shake</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>drop</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>press</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>push</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Accuracy vs. Number of Behaviors

![Graph showing accuracy vs. number of behaviors with test object. The graph includes lines for proprioception only, audio only, and proprioception and audio.](image)
1 Behavior

Multiple Behaviors
What can explain this boost?

• Machine Learning Theory:
  – Classification improvement is directly related to classifier diversity [Dietterich, 2000; Kuncheva et al, 2003]

• Our hypothesis:
  – The robot’s collection of behavior-grounded recognition models acts as a classifier ensemble

Sinapov and Stoytchev, “The Boosting Effect of Exploratory Behaviors”
In Proceedings of AAAI, 2010
Take Home Message

- Number of objects $\uparrow$ recognition accuracy $\downarrow$
- Number of behaviors $\uparrow$ recognition accuracy $\uparrow$
- Number of modalities $\uparrow$ recognition accuracy $\uparrow$
- Behaviors are classifiers that can be boosted
Developmental Progression

- Object Individuation
- Object Recognition
- Object Grouping
- Category Acquisition
- Learning Relations
Developmental Progression of Objects Knowledge

**Categorization**: infants can learn object categories based on perceptual object properties.

**Object Permanence**: infants can learn that an object moving behind an occluder remains the same object after it reappears.

**Recognition**: infants can perform visual object recognition. Infants can discriminate textures by touch. In the absence of visual cues, infants use more efficient exploratory strategies.

**Individuation**: infants can individuate objects. Furthermore, naming objects enhances infant object individuation.

**Categorization**: infants begin to form categories based on abstract properties such as the objects’ function and their spatial relations.
Developmental Progression

- Object Individuation
- Object Recognition
- Object Grouping
- Category Acquisition
- Learning Relations
Beyond Recognition: Category Acquisition

pans

pots

cups

silverware

glasses
Categorizing a novel object

- pans
- pots
- cups
- silverware
- glasses

?
The Symbol Grounding Problem

GOFAI was the dominant paradigm of AI research from the middle 1950s until the late 1980s. The Symbol Grounding Problem is related to the problem of how words (symbols) get their meanings, and hence to the problem of what meaning itself really is.

If symbols (words) always are explained with other symbols we get infinite regress. Somewhere symbols must be “grounded”! In what way does that grounding happen?
Progression of dataset sizes

- **Sinapov et al. 2008**: 3 objects, 10 behaviors
- **Sinapov et al. 2009**: 36 objects, 5 behaviors
- **Sinapov et al. 2011**: 100 objects, 10 behaviors

- **Audio and proprioception**
- **Audio**
- **Tactile (and proprioception)**
Progression of dataset sizes

- Audio and proprioception
  - Sinapov et al. 2014
- Audio
  - Sinapov et al. 2011
- Tactile (and proprioception)
  - Sinapov et al. 2011
  - Sinapov et al. 2009
  - Sinapov et al. 2008

# of objects

# of behaviors

Sinapov et al. 2008:
- 3 objects
- 18 behaviors

Sinapov et al. 2009:
- 5 objects
- 36 behaviors

Sinapov et al. 2011:
- 10 objects
- 50 behaviors

Sinapov et al. 2014:
- 100 behaviors
Experimental Setup

Experimental Platform

- ZCam (RGB+D)
- Logitech Webcam
- 3-axis accelerometer
- Microphones in the head
- Torque sensors in the joints
100 objects
Exploratory Behaviors

grasp  lift  hold  shake  drop

tax  poke  push  press
Overview

Interaction with Object

Sensorimotor Feature Extraction

Category Estimates

Category Recognition Model
Coupling Action and Perception

Action: poke

Perception: optical flow

Time
### Sensorimotor Contexts

<table>
<thead>
<tr>
<th>audio (DFT)</th>
<th>proprioception (joint torques)</th>
<th>proprioception (finger pos.)</th>
<th>Color</th>
<th>Optical flow</th>
<th>SURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grasp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>push</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>press</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Sensorimotor Contexts

<table>
<thead>
<tr>
<th></th>
<th>audio (DFT)</th>
<th>proprioception (joint torques)</th>
<th>proprioception (finger pos.)</th>
<th>Color</th>
<th>Optical flow</th>
<th>SURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>grasp</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>lift</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>hold</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>shake</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>drop</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>tap</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>poke</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>push</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>press</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Feature Extraction: Color

Object Segmentation

Color Histogram (4 x 4 x 4 = 64 bins)
Feature Extraction: Optical Flow

Angular bins

Count
Feature Extraction: SURF
Feature Extraction: SURF

Each interest point is described by a 128-dimensional vector
Feature Extraction: SURF

Visual “words”
Feature Extraction: Proprioception

Joint-Torque values for all 7 Joints

Joint-Torque Features
Feature Extraction: Audio

audio spectrogram

Spectro-temporal Features
Dimensionality of Data

<table>
<thead>
<tr>
<th>audio (DFT)</th>
<th>proprioception (joint torques)</th>
<th>proprioception (finger pos.)</th>
<th>Color</th>
<th>Optical flow</th>
<th>SURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>70</td>
<td>6</td>
<td>64</td>
<td>10</td>
<td>200</td>
</tr>
</tbody>
</table>
## Data From a Single Exploratory Trial

<table>
<thead>
<tr>
<th></th>
<th>audio (DFT)</th>
<th>proprioception (joint torques)</th>
<th>proprioception (finger pos.)</th>
<th>Color</th>
<th>Optical flow</th>
<th>SURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>grasp</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>lift</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>hold</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>shake</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>drop</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>tap</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>poke</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>push</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>press</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Data From a Single Exploratory Trial

<table>
<thead>
<tr>
<th></th>
<th>audio (DFT)</th>
<th>proprioception (joint torques)</th>
<th>proprioception (finger pos.)</th>
<th>Color</th>
<th>Optical flow</th>
<th>SURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>grasp</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>lift</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>hold</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>shake</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>drop</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>tap</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>poke</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>push</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>press</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

x 5 per object
Overview

Interaction with Object

Sensorimotor Feature Extraction

Category Estimates

Category Recognition Model
Context-specific Category Recognition

Observation from *poke-audio* context

Recognition model for *poke-audio* context

Distribution over category labels
Combining Model Outputs

\[ M_{\text{look-color}} \quad M_{\text{tap-audio}} \quad M_{\text{lift-SURF}} \quad \ldots \quad M_{\text{press-prop.}} \]

Weighted Combination
Model Evaluation: 5 fold Cross-Validation

Train Set

Test Set
## Recognition Rates (%) with SVM

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Proprioception</th>
<th>Color</th>
<th>Optical Flow</th>
<th>SURF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td>58.8</td>
<td></td>
<td>58.9</td>
<td>67.7</td>
</tr>
<tr>
<td>grasp</td>
<td>45.7</td>
<td>38.7</td>
<td></td>
<td>12.2</td>
<td>57.1</td>
<td>65.2</td>
</tr>
<tr>
<td>lift</td>
<td>48.1</td>
<td>63.7</td>
<td></td>
<td>5.0</td>
<td>65.9</td>
<td>79.0</td>
</tr>
<tr>
<td>hold</td>
<td>30.2</td>
<td>43.9</td>
<td></td>
<td>5.0</td>
<td>58.1</td>
<td>67.0</td>
</tr>
<tr>
<td>shake</td>
<td>49.3</td>
<td>57.7</td>
<td></td>
<td>32.8</td>
<td>75.6</td>
<td>76.8</td>
</tr>
<tr>
<td>drop</td>
<td>47.9</td>
<td>34.9</td>
<td></td>
<td>17.2</td>
<td>57.9</td>
<td>71.0</td>
</tr>
<tr>
<td>tap</td>
<td>63.3</td>
<td>50.7</td>
<td></td>
<td>26.0</td>
<td>77.3</td>
<td>82.4</td>
</tr>
<tr>
<td>push</td>
<td>72.8</td>
<td>69.6</td>
<td></td>
<td>26.4</td>
<td>76.8</td>
<td>88.8</td>
</tr>
<tr>
<td>poke</td>
<td>65.9</td>
<td>63.9</td>
<td></td>
<td>17.8</td>
<td>74.7</td>
<td>85.4</td>
</tr>
<tr>
<td>press</td>
<td>62.7</td>
<td>69.7</td>
<td></td>
<td>32.4</td>
<td>69.7</td>
<td>77.4</td>
</tr>
</tbody>
</table>
## Recognition Rates (%) with SVM

<table>
<thead>
<tr>
<th>Action</th>
<th>Audio</th>
<th>Proprioception</th>
<th>Color</th>
<th>Optical Flow</th>
<th>SURF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td>58.8</td>
<td></td>
<td>58.9</td>
<td>67.7</td>
</tr>
<tr>
<td>grasp</td>
<td>45.7</td>
<td>38.7</td>
<td></td>
<td>12.2</td>
<td>57.1</td>
<td>65.2</td>
</tr>
<tr>
<td>lift</td>
<td>48.1</td>
<td>63.7</td>
<td></td>
<td>5.0</td>
<td>65.9</td>
<td>79.0</td>
</tr>
<tr>
<td>hold</td>
<td>30.2</td>
<td>43.9</td>
<td></td>
<td>5.0</td>
<td>58.1</td>
<td>67.0</td>
</tr>
<tr>
<td>shake</td>
<td>49.3</td>
<td>57.7</td>
<td></td>
<td>32.8</td>
<td>75.6</td>
<td>76.8</td>
</tr>
<tr>
<td>drop</td>
<td>47.9</td>
<td>34.9</td>
<td></td>
<td>17.2</td>
<td>57.9</td>
<td>71.0</td>
</tr>
<tr>
<td>tap</td>
<td>63.3</td>
<td>50.7</td>
<td></td>
<td>26.0</td>
<td>77.3</td>
<td>82.4</td>
</tr>
<tr>
<td>push</td>
<td>72.8</td>
<td>69.6</td>
<td></td>
<td>26.4</td>
<td>76.8</td>
<td>88.8</td>
</tr>
<tr>
<td>poke</td>
<td>65.9</td>
<td>63.9</td>
<td></td>
<td>17.8</td>
<td>74.7</td>
<td>85.4</td>
</tr>
<tr>
<td>press</td>
<td>62.7</td>
<td>69.7</td>
<td></td>
<td>32.4</td>
<td>69.7</td>
<td>77.4</td>
</tr>
</tbody>
</table>
## Recognition Rates (%) with SVM

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Proprioception</th>
<th>Color</th>
<th>Optical Flow</th>
<th>SURF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td>58.8</td>
<td></td>
<td>58.9</td>
<td>67.7</td>
</tr>
<tr>
<td>grasp</td>
<td>45.7</td>
<td>38.7</td>
<td></td>
<td>12.2</td>
<td>57.1</td>
<td>65.2</td>
</tr>
<tr>
<td>lift</td>
<td>48.1</td>
<td>63.7</td>
<td></td>
<td>5.0</td>
<td>65.9</td>
<td>79.0</td>
</tr>
<tr>
<td>hold</td>
<td>30.2</td>
<td>43.9</td>
<td></td>
<td>5.0</td>
<td>58.1</td>
<td>67.0</td>
</tr>
<tr>
<td>shake</td>
<td>49.3</td>
<td>57.7</td>
<td></td>
<td>32.8</td>
<td>75.6</td>
<td>76.8</td>
</tr>
<tr>
<td>drop</td>
<td>47.9</td>
<td>34.9</td>
<td></td>
<td>17.2</td>
<td>57.9</td>
<td>71.0</td>
</tr>
<tr>
<td>tap</td>
<td>63.3</td>
<td>50.7</td>
<td></td>
<td>26.0</td>
<td>77.3</td>
<td>82.4</td>
</tr>
<tr>
<td>push</td>
<td>72.8</td>
<td>69.6</td>
<td></td>
<td>26.4</td>
<td>76.8</td>
<td>88.8</td>
</tr>
<tr>
<td>poke</td>
<td>65.9</td>
<td>63.9</td>
<td></td>
<td>17.8</td>
<td>74.7</td>
<td>85.4</td>
</tr>
<tr>
<td>press</td>
<td>62.7</td>
<td>69.7</td>
<td></td>
<td>32.4</td>
<td>69.7</td>
<td>77.4</td>
</tr>
</tbody>
</table>
## Recognition Rates (%) with SVM

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Proprioception</th>
<th>Color</th>
<th>Optical Flow</th>
<th>SURF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td></td>
<td></td>
<td>58.8</td>
<td></td>
<td>58.9</td>
<td>67.7</td>
</tr>
<tr>
<td>grasp</td>
<td>45.7</td>
<td>38.7</td>
<td>12.2</td>
<td></td>
<td>57.1</td>
<td>65.2</td>
</tr>
<tr>
<td>lift</td>
<td>48.1</td>
<td>63.7</td>
<td>5.0</td>
<td></td>
<td>65.9</td>
<td>79.0</td>
</tr>
<tr>
<td>hold</td>
<td>30.2</td>
<td>43.9</td>
<td>5.0</td>
<td></td>
<td>58.1</td>
<td>67.0</td>
</tr>
<tr>
<td>shake</td>
<td>49.3</td>
<td>57.7</td>
<td>32.8</td>
<td></td>
<td>75.6</td>
<td>76.8</td>
</tr>
<tr>
<td>drop</td>
<td>47.9</td>
<td>34.9</td>
<td>17.2</td>
<td></td>
<td>57.9</td>
<td>71.0</td>
</tr>
<tr>
<td>tap</td>
<td>63.3</td>
<td>50.7</td>
<td>26.0</td>
<td></td>
<td>77.3</td>
<td>82.4</td>
</tr>
<tr>
<td>push</td>
<td>72.8</td>
<td>69.6</td>
<td>26.4</td>
<td></td>
<td>76.8</td>
<td>88.8</td>
</tr>
<tr>
<td>poke</td>
<td>65.9</td>
<td>63.9</td>
<td>17.8</td>
<td></td>
<td>74.7</td>
<td>85.4</td>
</tr>
<tr>
<td>press</td>
<td>62.7</td>
<td>69.7</td>
<td>32.4</td>
<td></td>
<td>69.7</td>
<td>77.4</td>
</tr>
</tbody>
</table>
Can behaviors be selected actively to minimize exploration time?
Active Behavior Selection

- For each behavior $b \in \mathcal{B}$, estimate $C_{i,j}^b$ such that

$$C_{i,j}^b = \frac{Pr(\hat{y} = y_i | y = y_j) + Pr(\hat{y} = y_j | y = y_i)}{2}$$

- Let $\mathbf{p} \in \mathbb{R}^{|\mathcal{Y}|}$ be the vector encoding the robot’s current estimates over the category labels and let $\mathcal{B}_r$ be the remaining set of behaviors available to the robot
Active Behavior Selection

1) Compute the set $\mathcal{Y}_K \subset \mathcal{Y}$ such that it contains the $K$ most likely object categories according to $\hat{p}$.

2) Pick the next behavior $b_{next}$ with an associated confusion matrix that is least likely to confuse the categories within the set $\mathcal{Y}_K$, i.e.,

$$b_{next} = \arg\min_{b \in \mathcal{B}_r} \sum_{y_i \in \mathcal{Y}_K} \sum_{y_j \in \mathcal{Y}_K / y_i} C_{ij}^b$$

3) Update the estimate $\hat{p}$ using the classifiers associated with the sensorimotor contexts of $b_{next}$.

4) Remove $b_{next}$ from $\mathcal{B}_r$. If $|\mathcal{B}_r| \geq 1$, go back to step 1).
Active vs. Random Behavior Selection

![Graph showing active vs. random behavior selection accuracy over the number of behaviors performed on a test object. The graph illustrates the difference in accuracy between active selection and random selection as the number of behaviors increases.]
Active vs. Random Behavior Selection

- **Active Selection**
- **Random Selection**

Category Recognition Accuracy (%) vs. Number of Behaviors Performed on Test Object

![Graph showing comparison between Active and Random Behavior Selection]
Summary of Experiment

• Using visual, proprioceptive and auditory sensory modalities, coupled with 10 different behaviors, the robot achieved high category recognition rates

• Sensorimotor contexts can be linked to specific categories enabling even better rates when looking for a specific object type

• Active behavior selection when classifying a novel object reduced exploration time by half
Developmental Progression

- Object Individuation
- Object Recognition
- Object Grouping
- Category Acquisition
- Learning Relations
Developmental Progression

- Object Individuation
- Object Recognition
- Object Grouping
- Category Acquisition
- Learning Relations
Object Individuation

• Object individuation is the problem of inferring how many unique objects have been observed

• To solve it, a robot has to partition its sensorimotor experience according to the identity of the object that is present
Object Individuation

Observations

Unique Objects

Object 1
Object 3
Object 2
Object 4

Sinapov and Stoytchev, “Grounded Object Individuation by a Humanoid Robot”
Proceedings of the 2013 IEEE ICRA
How many unique objects were there?
How many unique objects were there?

• Correct answer: 4
Related Work in 3D Object Discovery

Collet et al. 2013
Problem Formulation

**Exploratory Trials**
- Interaction 1
- Interaction 2
- ...
- Interaction N

**Individuated Objects**
- Interactions with object 1
- Interactions with object 2
- Interactions with object 3
- ...
- Interactions with object M
Approach

Learn Individuation Model

Trials with known number of objects (labeled)

Use Individuation Model

Trials with unknown number of objects (unlabeled)

Individuated Objects

- Trials with object 1
- Trials with object 2
- ...
- Trials with object M
Learning an Individuation Model

• Learn a model that classifies a trial pair \((T_i, T_j)\) as either “same object” or “different object”
Learning an Individuation Model

$T_i$

- look
- color
- look
- SURF
- grasp
- audio
- grasp
- torques
- grasp
- SURF
- .
- .
- .
- press
- Flow

$T_j$

- look
- color
- look
- SURF
- grasp
- audio
- grasp
- torques
- grasp
- SURF
- .
- .
- .
- press
- Flow
Learning an Individuation Model

$T_i$

- look color
- look SURF
- grasp audio
- grasp torques
- grasp SURF

... 

$T_j$

- look color
- look SURF
- grasp audio
- grasp torques
- grasp SURF

... 

press Flow
Learning an Individuation Model

\[ T_i \]
- Look
- Color
- Look
- SURF
- Grasp
- Audio
- Grasp
- Torques
- Grasp
- SURF
- Press
- Flow

... Compute Euclidean Distances ...

\[ T_j \]
- Look
- Color
- Look
- SURF
- Grasp
- Audio
- Grasp
- Torques
- Grasp
- SURF
- Press
- Flow
Learning an Individuation Model

\[ T_i \]
- look
color
- look
SURF
- grasp
audio
- grasp
torques
- grasp
SURF
- press
Flow

\[ f_{ij} \]
- 5.73
- 0.34
- 2.76
- 10.7
- 67.3
- 23.8

\[ T_j \]
- look
color
- look
SURF
- grasp
audio
- grasp
torques
- grasp
SURF
- press
Flow
Example Training Set
(2 objects and 4 trials)

Objects:

Trials:

\[ T_1 \]

\[ T_2 \]

\[ T_3 \]

\[ T_4 \]
Learning an Individuation Model

Objects:

Trials:

\( f_{12} \in \mathbb{R}^{39} \), “same”
Learning an Individuation Model

Objects:

Trials:

\[ T_1 \quad T_2 \]

\[ T_3 \quad T_4 \]

\[ f_{12} \in \mathbb{R}^{39}, \text{ "same"} \]

\[ f_{13} \in \mathbb{R}^{39}, \text{ "different"} \]
Learning an Individuation Model

Objects:

Trials:

\[ f_{12} \in \mathbb{R}^{39}, \text{ “same”} \]
\[ f_{13} \in \mathbb{R}^{39}, \text{ “different”} \]
\[ f_{14} \in \mathbb{R}^{39}, \text{ “different”} \]
\[ f_{23} \in \mathbb{R}^{39}, \text{ “different”} \]
\[ f_{24} \in \mathbb{R}^{39}, \text{ “different”} \]
\[ f_{34} \in \mathbb{R}^{39}, \text{ “same”} \]
Individuating Novel Objects

- Let $T_{test}$ be a test set of 25 trials with 5 different objects (5 trials per object)
Output of Individuation Model
Output of Individuation Model

“same”

“different”
Individuating Novel Objects

Recursive Spectral Clustering
Individuating Objects

$C_1$

$C_2$

$C_3$

$C_4$

$C_5$
Evaluation Metrics

- Normalized Mutual Information:
  - captures the similarity between two different clusterings over the same dataset

\[
\phi^{NMI}(\omega^a, \omega^b) = \frac{\sum_{h=1}^{k^a} \sum_{\ell=1}^{k^b} n_{h,\ell} \log \left( \frac{n \cdot n_{h,\ell}}{n_h^a \cdot n_{\ell}^b} \right)}{\sqrt{\left( \sum_{h=1}^{k^a} n_h^a \log \left( \frac{n_h^a}{n} \right) \right) \left( \sum_{\ell=1}^{k^b} n_{\ell}^b \log \left( \frac{n_{\ell}^b}{n} \right) \right)}}
\]
How many training objects do we need?

![Normalized Mutual Information vs. Number of Objects in the Training Set](image)

- **Normalized Mutual Information**: The y-axis represents the normalized mutual information, which is a measure of how much information one random variable contains about another. It is normalized to range from 0 to 1, where 1 indicates perfect mutual information.
- **Number of Objects in the Training Set**: The x-axis shows the number of objects in the training set, ranging from 5 to 40.

The graph illustrates the trend of normalized mutual information as the number of objects in the training set increases. The data suggests that the mutual information stabilizes as the number of objects reaches a certain threshold, indicating that additional training objects beyond this point do not significantly improve mutual information.
How many training objects do we need?

The graph shows the normalized mutual information as a function of the number of objects in the training set. The graph indicates that the mutual information reaches a stable value around 0.95 when the number of objects is approximately 15. This suggests that 15 training objects may be sufficient to achieve a high level of mutual information in the training set.
Baseline Comparison

- The approach is compared with unsupervised clustering over a matrix encoding raw perceptual similarity.
Baseline Comparison
How many training objects do we need?

![Graph showing normalized mutual information vs. number of objects in the training set]

- **Normalized Mutual Information**
- **Number of Objects in the Training Set**

The graph illustrates the normalized mutual information for different numbers of objects in the training set. It appears that an unsupervised baseline model reaches a high normalized mutual information value after a certain number of objects are included in the training set.
Summary

• The robot discovered novel objects using exploratory behaviors coupled with multiple sensory modalities

• Unsupervised clustering alone is insufficient for solving the object individuation problem

• Object discovery can be improved using small amounts of training data
Developmental Progression

- Object Individuation
- Object Recognition
- Object Grouping
- Category Acquisition
- Learning Relations
Claims

• Behaviors allow robots not only to affect the world, but also to perceive it
• Non-visual sensory feedback improves object classification and perception tasks that are typically solved using vision alone
• A diverse sensorimotor repertoire is necessary for scaling up object recognition, categorization, and individuation to a large number of objects
What's next?
Learning 'in the wild'

Gori et al, “Robot-Centric Activity Recognition 'in the Wild'”
In proceedings of the 7th International Conference on Social Robotics (ICSR 2015)
Learning 'in the wild'

Gori et al., “Robot-Centric Activity Recognition 'in the Wild'”
In proceedings of the 7th International Conference on Social Robotics (ICSR 2015)
Language Learning

In proceedings of the 2016 International Joint Conference on Artificial Intelligence
Language Learning

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

In proceedings of the 2016 International Joint Conference on Artificial Intelligence
Acknowledgements

Peter Stone  J. K. Aggrawal  Ray Mooney  Ilaria Gori  Matteo Leonetti  Shiqi Zhang  Alex Stoytchev

Vasant Honavar  Heiko Hoffmann  Cornelia Caragea  Sanmit Narvekar  Proiyanka Khante  Maxwel Svetlik  Jesse Thomason  Piyush Khandelwal

Nick Walker  Vlad Sukhoy  Shane Griffith  Connor Schenck  Ritika Sahai  Taylor Bergquist  Adrian Silvescu
Acknowledgements
...and a big thanks to my undergrads students
Thanks! Questions?
Thank You