Towards Robust Human-Robot Interaction using Multimodal Cues

HART Workshop 2012

Ranjini Swaminathan and Mohan Sridharan

Texas Tech University

March 5, 2012
Robust and Natural HRI

Challenges:
- Real world characterized by dynamic environments.
- Noisy information - sensitivity to environmental changes.

Proposed Solutions:
- Multimodal cues - visual, verbal, laser.
- High level non-expert human feedback.
Robust and Natural HRI

- **Challenges**:  
  - Real world characterized by dynamic environments.  
  - Noisy information - sensitivity to environmental changes.

- **Proposed Solutions**:  
  - Multimodal cues - visual, verbal, laser.  
  - High level non-expert human feedback.
Broader Scope

- Richer more complex representation of objects and scenes.
- Learning associations between modalities.
- Semantic representation over modalities.
- Improved cognition and HRI.
Architecture

Multimodal Association

Human–robot Interaction

Bootstrap Learning

Hierarchical Planning
Cup of Features - Multimodal Learning

Visual Features
Verbal Features
Laser cues
Object models

Ranjini Swaminathan and Mohan Sridharan
HART 2012
Contributions

- Probabilistic bootstrap learning algorithm - visual vocabulary.

- Probabilistic graphical models and lexical tools - verbal vocabulary.

- Association between visual and verbal vocabularies.
Approach

- **Multimodal object descriptions**: 
  - Object properties - color, shape, size.
  - Values for properties - red, blue, square, triangle, small, large.

- **Natural language HRI**: 
  - Object categories - normal, questionable - query human.
  - Natural language querying.
Multimodal learning framework

- Multimodal Learning
  - Visual Learning
    - Bootstrap learning + Visual operators
  - Verbal Learning
    - Graphical models + Lexical tools

Visual–Verbal mapping
Query generation
Tabletop Block Object Scenario
Visual properties

- Identify ROI.
- Color $\rightarrow$ RGB channels - 3D histogram.
- Shape $\rightarrow$ External contour using Hu invariant moments.
- Size $\rightarrow$ Relative number of pixels.
Bootstrap Learning

- Define $N$ classes for each property: $C_i, i \in [1, N]$.
- Classes correspond to possible values.
- Find distribution over classes - assign a class.
- For new objects:
  - Compute visual features.
  - Find distribution over classes based on distance metric.
  - If necessary, add a new class before assignment (bootstrapping).
- Repeat for all object properties.
Verbal features

- Part of Speech tags.
- Property labels - SIZ, COL, SHA, COM.
- CRF - POS tags and words (observed variables) and property tags (target variables).

<table>
<thead>
<tr>
<th>Part of Speech Tag</th>
<th>DT</th>
<th>JJ</th>
<th>JJ</th>
<th>NN</th>
<th>VBZ</th>
<th>RB</th>
<th>JJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>This</td>
<td></td>
<td>small</td>
<td>red</td>
<td>triangle</td>
<td>looks</td>
<td>quite</td>
<td>standard</td>
</tr>
<tr>
<td>O</td>
<td>B_SIZ</td>
<td>B_COL</td>
<td>B_SHA</td>
<td>B_COM</td>
<td>L_COM</td>
<td>L_COM</td>
<td></td>
</tr>
</tbody>
</table>
Lexical - Semantic Information using WordNet

- **gigantic**
- **huge**
- **big**
- **ample**
- **enormous**
- **massive**

- **is-a (hyponym)**

- **chromatic_color#n#1**
  - **red#n#1**
  - **blue#n#1**
  - **green#n#1**
  - **yellow#n#1**
Feature Cup

Multimodal Associations and Object Models

- Color histogram
- Shape Features
- Size of ROI
- POS Tags
- Property Tags
- Lexical Features
Associating visual and verbal properties

- **Visual** - classes for each object property and distribution over classes.

- **Verbal** - same as visual + semantic grounding.

- **Object characterization** - Multimodal model for an object.

- **Mapping** - Distance measure between visual and verbal class distributions.
Multimodal associations

- Green objects: 0.691
- Yellow objects: 0.347
- Blue objects: 0.691
- Red objects: 0.032

New object
Semantic Grounding

- Color
  - Blue
    - blue
    - NN
    - COL
  - Green
    - green
    - NN
    - COL
Task: Categorizing objects

- Objects grouped randomly - normal and questionable.
- SVM classification with multimodal object properties as features.
- Classification - with category, visual+verbal, visual, verbal.
- Verbal templates for natural language querying.
Sample annotations:

Normal: *The small yellow circle looks normal.*

Suspicious: *This large red rectangle is atypical.*
Results

<table>
<thead>
<tr>
<th>Features used</th>
<th>Classification accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual + Verbal + Category</td>
<td>97.5</td>
</tr>
<tr>
<td>Visual + Verbal</td>
<td>77.5</td>
</tr>
<tr>
<td>Visual</td>
<td>75</td>
</tr>
<tr>
<td>Verbal</td>
<td>72.5</td>
</tr>
</tbody>
</table>

Table: Accuracy results for object category classification (cross-validation results).

Sample query: This large red rectangle looks suspicious. How should I proceed?
Future Work

- Rich environment descriptions (objects and scenes).
- Multi object descriptions and relationships.
- Natural and more descriptive verbal annotations.
- Uncertainty in human feedback.