Robots Need Language: 
A computational model for the integration of vision, language and action

Ching L Teo, Yezhou Yang
Yiannis Aloimonos, Hal Daumé III
Dept. of Computer Science, University of Maryland
September 12, 2012
Our Proposal

- Create **Cognitive Robots** of the future:
  - *Interacts* with humans,
  - *Understands* common (complex) situations,
  - *Proposes* reasonable actions.

By exploiting **Language** as a source of world knowledge
A Typical Situation

Key questions for the robot include (but not limited to):

- What is going on? → Scene Understanding
- Who is doing what? → Action Recognition
- What tools are used? → Object Recognition
- ...

Camera system

Laser
The Cognitive Dialog Framework

- A model of a reasoning process that involves the **Visual Executive (VE)** and **Language Executive (LE)**, with **Action** in the middle:

  - **VE**: observations to **LE**, e.g. low-level feature extraction,
  - **LE**: constraints from knowledge, proposing reasonable responses,
  - **Action**: Performs actions, updates **VE**.
Implementation

- Limit ourselves to *Kitchen Scenarios*:
  - Well annotated dataset, with variations.
- Task for robot is to **describe what is going on**

< The human is **stirring the bowl using fork/spoon**>
Key HW & SW Components

- Kinect
- RGB/Depth
- Tracked Skeleton
- Linguistic Resources
  - Large Text Corpus
  - WordNet
- 3D processing
- Features extraction
- Object/Action Classification
- Visual Processes (VE)
  - <Action, Object, Tool>
- Language Reasoner (LE)
Attention-Guided Navigation

Hardware: Kinect, PTU, Laser, Pioneer

Plane + Object Detection

Action Attributes Encoding

Optimization of Visual + Language information

Object-Tools Co-occurrences

Gigawords Corpus

< The human is cutting the bagel with the knife >

Sentence Generation

Demo Video
On-going Work (1)

Attributes-based Recognition

Perception

World Knowledge

Grounding

Generating High-Level Concepts

Action recognition from Activity Tree Grammar
On-going Work (2)

Manipulative action understanding

Attention-based Segmentation+Tracking

Object search using high-level knowledge

Action recognition via cause-effect

Come to our poster session for more details + live demo!
Current Computer Vision techniques are limited when low-level signals are used:

- Introduced *language* as a **key enabler** for *perception* to occur

Formulated the interplay of vision and language as a **Cognitive Dialog**:

- Algorithms developed around this framework
- Suitable for cognitive robots of the future

Beyond integration at the *semantic* (label) level:

- Numerous on-going work on integrating language into all levels of perception
Thank You

- We would like to thank:
  - Mentors Ashwin Sampath and Snehesh Shrestha,
  - Qualcomm for the Innovation Fellowship that has made all these possible.

- Contacts:
  - Yiannis Aloimonos yiannis@cs.umd.edu
  - Hal Daumé III hal@umiacs.umd.edu
  - Ching Lik Teo cteo@cs.umd.edu
  - Yezhou Yang yzyang@cs.umd.edu