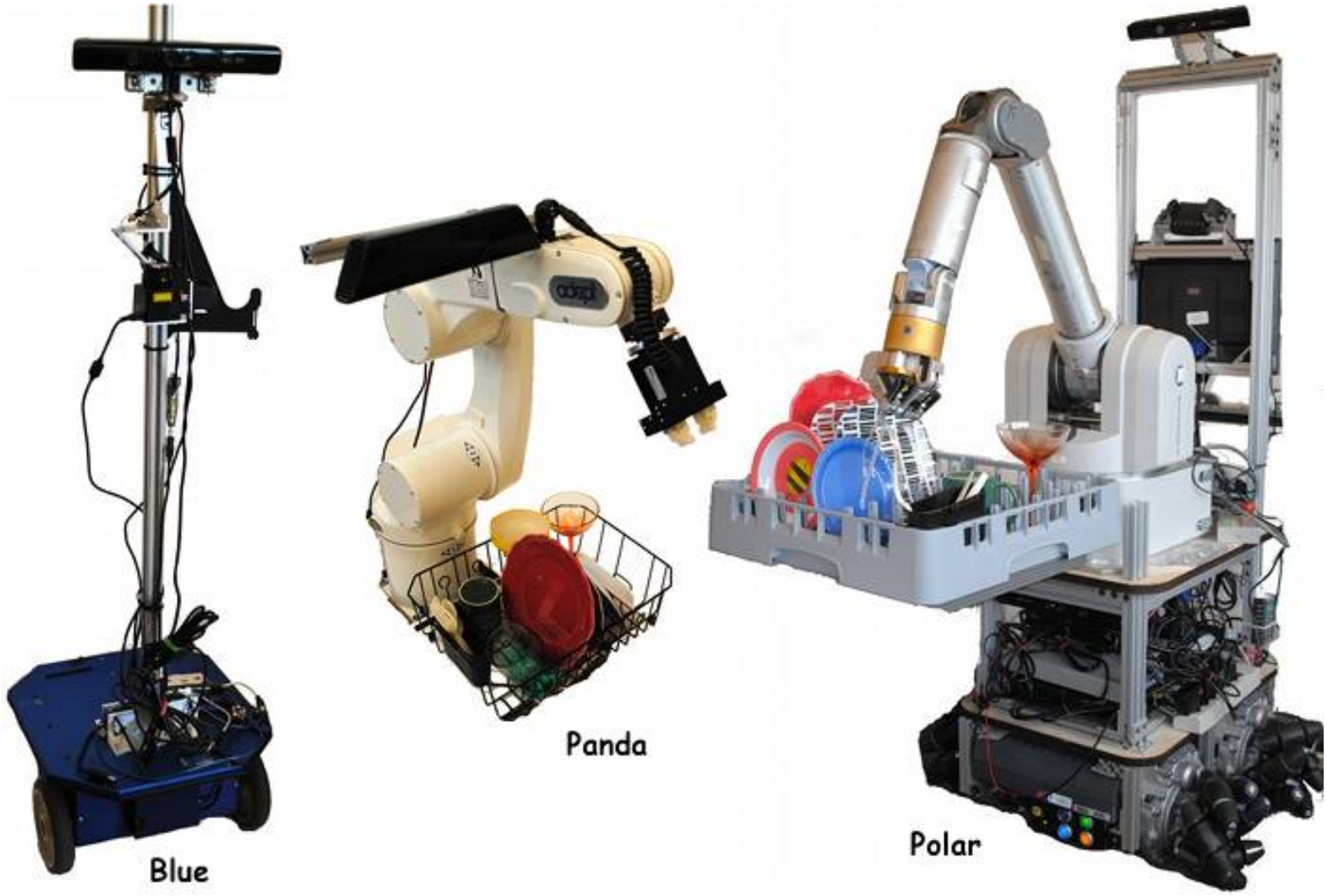
# 3D Perception for Personal Robots

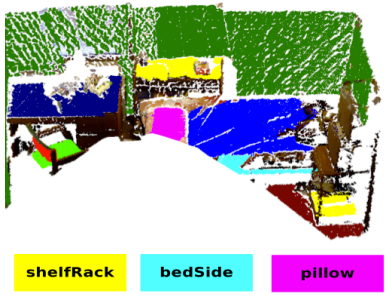
Ashutosh Saxena

Cornell Personal Robotics

http://pr.cs.cornell.edu


Blue
Panda
Polar

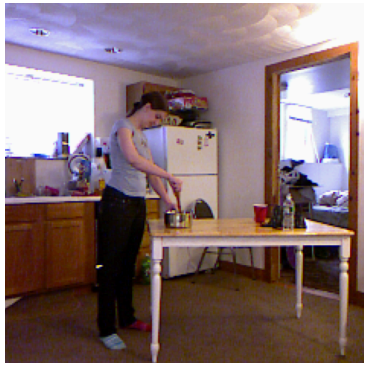
# Cornell Personal Robotics




Labeling 3D Scenes.

Learning Manipulation:

Placing, grasping, 3d articulated objects.



Human Activity Detection, Social Interaction.

More details, videos, data and code at:

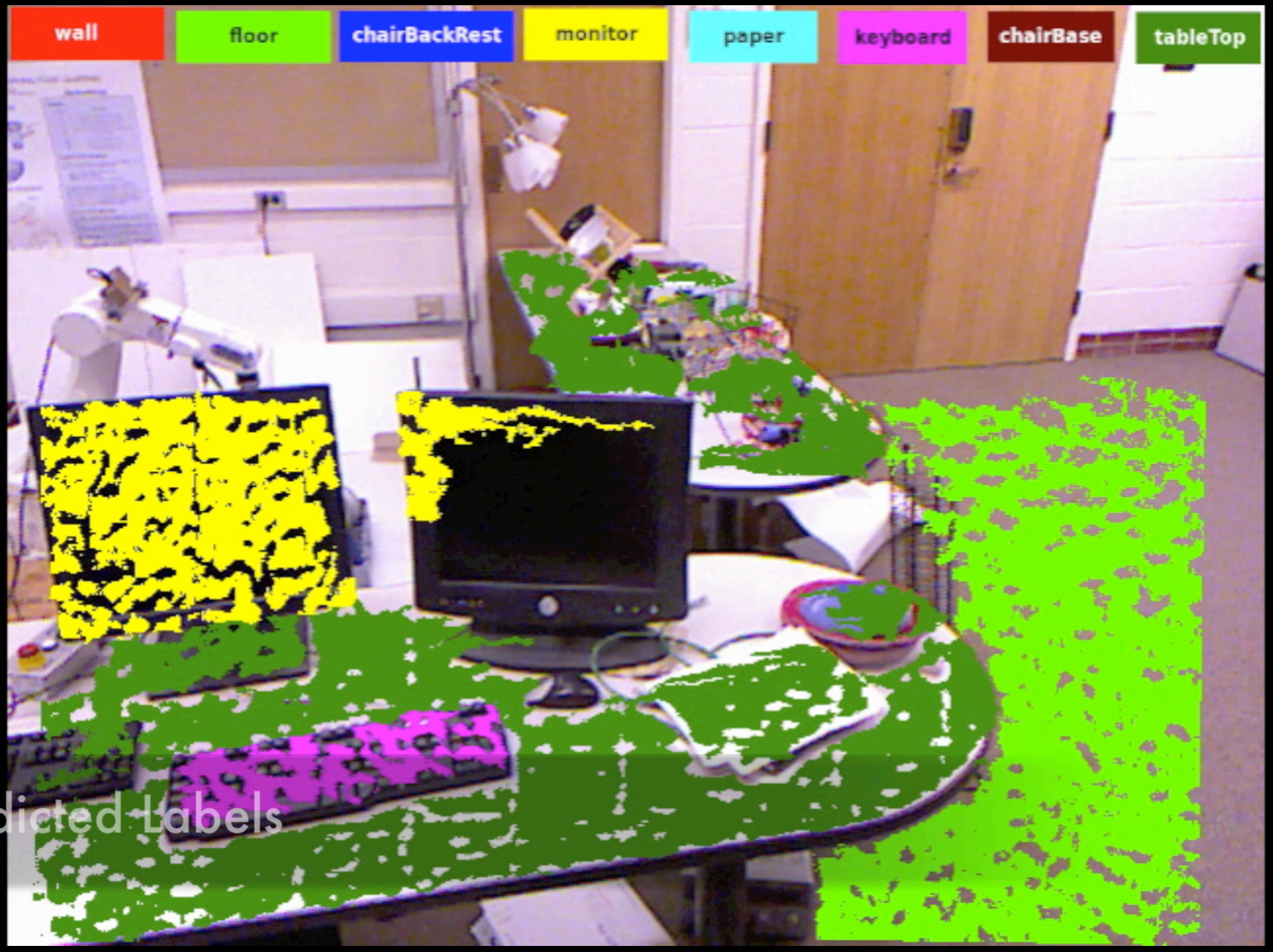
http://pr.cs.cornell.edu

# Labeling 3D Scenes for Personal Assistant Robots

Hema Koppula, Abhishek Anand, Thorsten Joachims, Ashutosh Saxena.

In *R:SS workshop on RGB-D cameras*, 2011.

# Polar Robot Finding Objects: Video



# Problem Statement

- Given a full 3D point-cloud of a scene
  - Each scene obtained from 8-9 views.

- Goal: Label each ***3D segment***:
  - *Labels: wall, floor, tableTop, tableDrawer, tableLeg, chairBackRest, chairBase, chairBack, monitor, printerFront, printerSide keyboard, cpuTop, cpuFront, cpuSide, book, paper, sofaBase, sofaArm, sofaBackRest, bed, bedSide, quilt, pillow, shelfRack, laptop, book.*
  - *Attributes: wall, floor, flat-horizontal-surfaces, furniture, fabric, heavy, seating-areas, small-objects, table-top-objects, electronics.*

# How to label an object?

Several properties should be captured:

- Local
  - Visual Appearance.
  - Shape and geometry.
- Context
  - Visual: similar looking segments have same label.
  - Geometric Context, e.g., "on top of", "in front of", "convexity", etc.

Some of these properties have been explored with RGB images, e.g., Make3D (Saxena et al. 2005) and photo popup (Hoiem et al. 2005).

3D data (RGB-D) makes this much more powerful.

# Related Work: 3D Point-clouds.

- Quifley et al., Collet et al. (MOPED)
  - Local features only.

- Anguelov et al., CVPR 2005; Munoz, Vanapel, Hebert, ICRA 2009.
  - LIDAR data of *outdoor* environments.
  - Typically 3-5 geometric classes (ground, trees, building, etc.)
  - Can only model favoring similar labels (associative Markov networks).

- Xiong and Huber, BMVC 2010.
  - Label planar segments from a single view.
  - Label 4 geometric classes: *walls, floors, ceilings, clutter*.
  - Learning method based on pseudo-likelihood.

# Problem Statement

- Given a full 3D point-cloud of a scene
  - Each scene obtained from 8-9 views.

- Goal: Label each ***3D segment***:
  - *Labels: wall, floor, tableTop, tableDrawer, tableLeg, chairBackRest, chairBase, chairBack, monitor, printerFront, printerSide keyboard, cpuTop, cpuFront, cpuSide, book, paper, sofaBase, sofaArm, sofaBackRest, bed, bedSide, quilt, pillow, shelfRack, laptop, book.*
  - *Attributes: wall, floor, flat-horizontal-surfaces, furniture, fabric, heavy, seating-areas, small-objects, table-top-objects, electronics.*

# Learning Algorithm

- Pointcloud of a segment $x_i$
- Label $y_i$
- $y_i^k$ = 1, if $i^{th}$ segment has label k.

- Discriminant Function:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, f_{\mathbf{w}}(\mathbf{x}, \mathbf{y})$$

$$f_{\mathbf{w}}(\mathbf{y}, \mathbf{x}) = \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} y_i^k \left[ w_n^k \cdot \phi_n(i) \right]$$

# Our Model: Markov Random Field

$$f_{\mathbf{w}}(\mathbf{y}, \mathbf{x}) = \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} y_i^k \left[ w_n^k \cdot \phi_n(i) \right] +$$

$$\sum_{(i,j) \in \mathcal{E}} \sum_{T_t \in \mathcal{T}} \sum_{(l,k) \in T_t} z_{ij}^{lk} \left[ w_t^{lk} \cdot \phi_t(i,j) \right]$$

Segment i, j

Label l, k

# Our Model: Markov Random Field

$$f_{\mathbf{w}}(\mathbf{y},\mathbf{x}) = \sum_{i\in\mathcal{V}}\sum_{k=1}^{K} y_i^k \left[w_n^k \cdot \phi_n(i)\right] +$$

$$\sum_{(i,j)\in\mathcal{E}} \sum_{T_t\in\mathcal{T}} \sum_{(l,k)\in T_t} z_{ij}^{lk} \left[w_t^{lk} \cdot \phi_t(i,j)\right]$$

- Parsimonious model: Avoid learning weight vectors for relationships which do not exist.
  - "object-associative" features: used between classes that are parts of the same object (e.g.,"chair base","chair back" and "chair back rest").
  - "non-associative" features: used between any pair of classes.
  - "associative" features: that prefer same label, e.g., similar visual appearance.

# Features

Node features for segment $i$.

| Description | Count |
|---|---|
| **Visual Appearance** | **48** |
| N1. Histogram of HSV color values | 14 |
| N2. Average HSV color values | 3 |
| N3. Average of HOG features of the blocks in image spanned by the points of a segment | 31 |
| **Local Shape and Geometry** | **8** |
| N4. linearness ($\lambda_{i0}$ - $\lambda_{i1}$), planarness ($\lambda_{i1}$ - $\lambda_{i2}$) | 2 |
| N5. Scatter: $\lambda_{i0}$ | 1 |
| N6. Vertical component of the normal: $\hat{n}_{iz}$ | 1 |
| N7. Vertical position of centroid: $c_{iz}$ | 1 |
| N8. Vert. and Hor. extent of bounding box | 2 |
| N9. Dist. from the scene boundary (Fig. 2) | 1 |

Features for edge (segment $i$, segment $j$).

| Description | Count |
|---|---|
| **Visual Appearance (associative)** | **3** |
| E1. Difference of avg HSV color values | 3 |
| **Local Shape and Geometry (associative)** | **2** |
| E2. Coplanarity and convexity (Fig. 2) | 2 |
| **Geometric context (non-associative)** | **6** |
| E3. Horizontal distance b/w centroids. | 1 |
| E4. Vertical Displacement b/w centroids: $(c_{iz} - c_{jz})$ | 1 |
| E5. Angle between normals (Dot product): $\hat{n}_i \cdot \hat{n}_j$ | 1 |
| E6. Diff. in angle with vert.: $\cos^{-1}(n_{iz})$ - $\cos^{-1}(n_{jz})$ | 1 |
| E8. Dist. between closest points: $\min_{u \in s_i, v \in s_j} d(u, v)$ (Fig. 2) | 1 |
| E8. rel. position from camera (in front of/behind). (Fig. 2) | 1 |

# Learning and Inference

- Learning
  - Large-margin approach using structural SVMs.

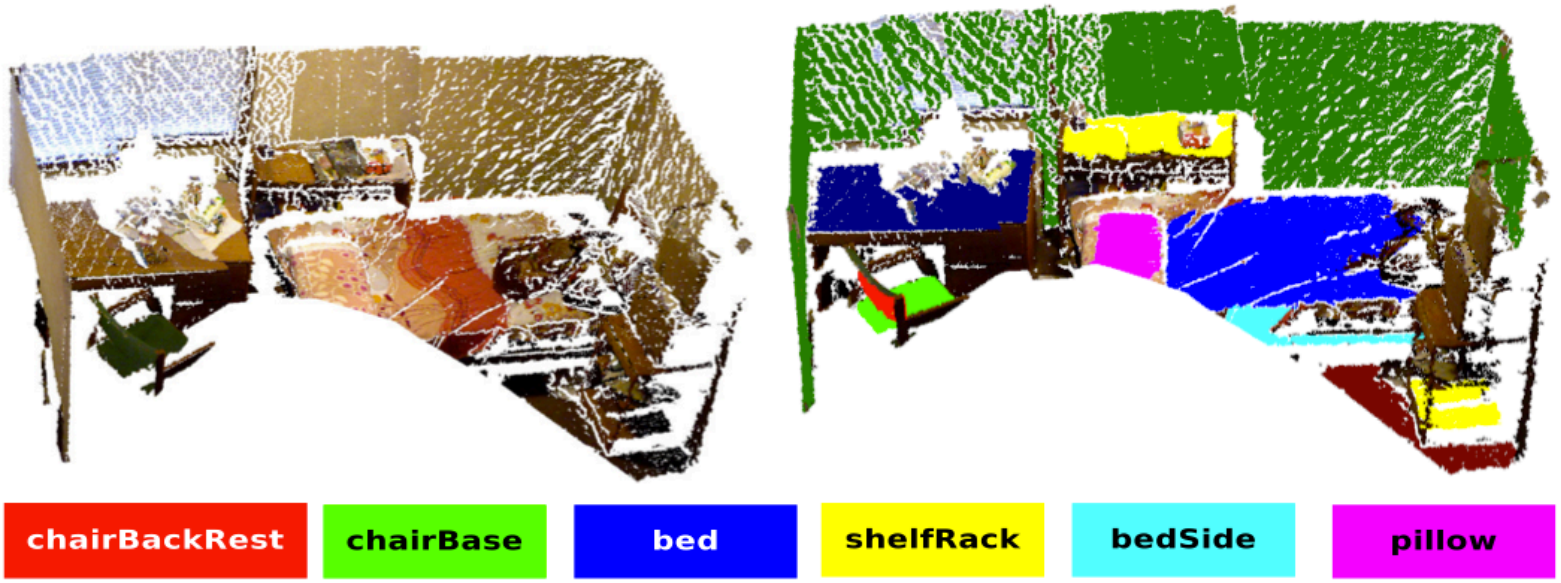
- Inference takes 2 seconds on a full scene, and 0.2 seconds on a single view.

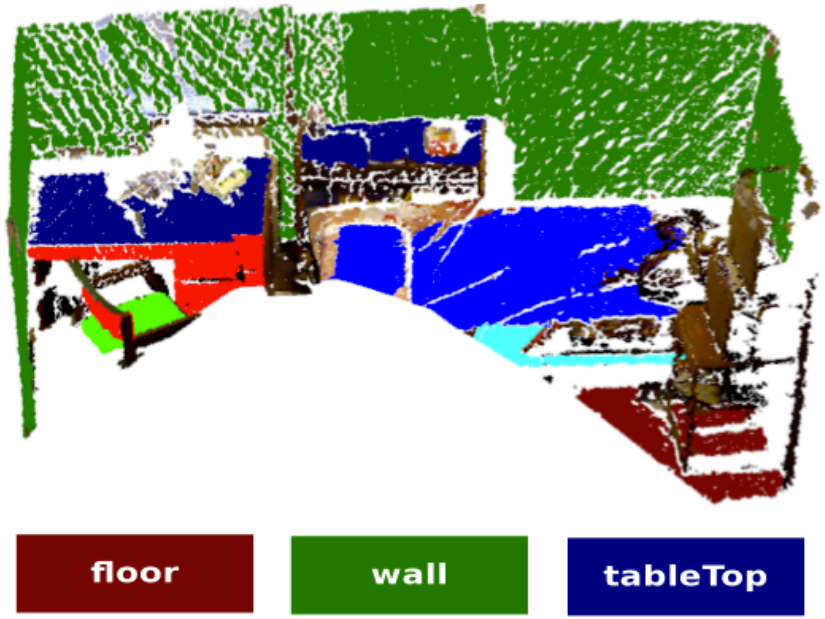
$$\text{Inference: } \hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \max_{\mathbf{z}} f_{\mathbf{w}}(\mathbf{x}, \mathbf{y})$$

$$\forall i, j, l, k: \quad z_{ij}^{lk} \leq y_i^l, \quad z_{ij}^{lk} \leq y_j^k,$$
$$y_i^l + y_j^k \leq z_{ij}^{lk} + 1, \quad z_{ij}^{lk}, y_i^l \in \{0, 1\}$$

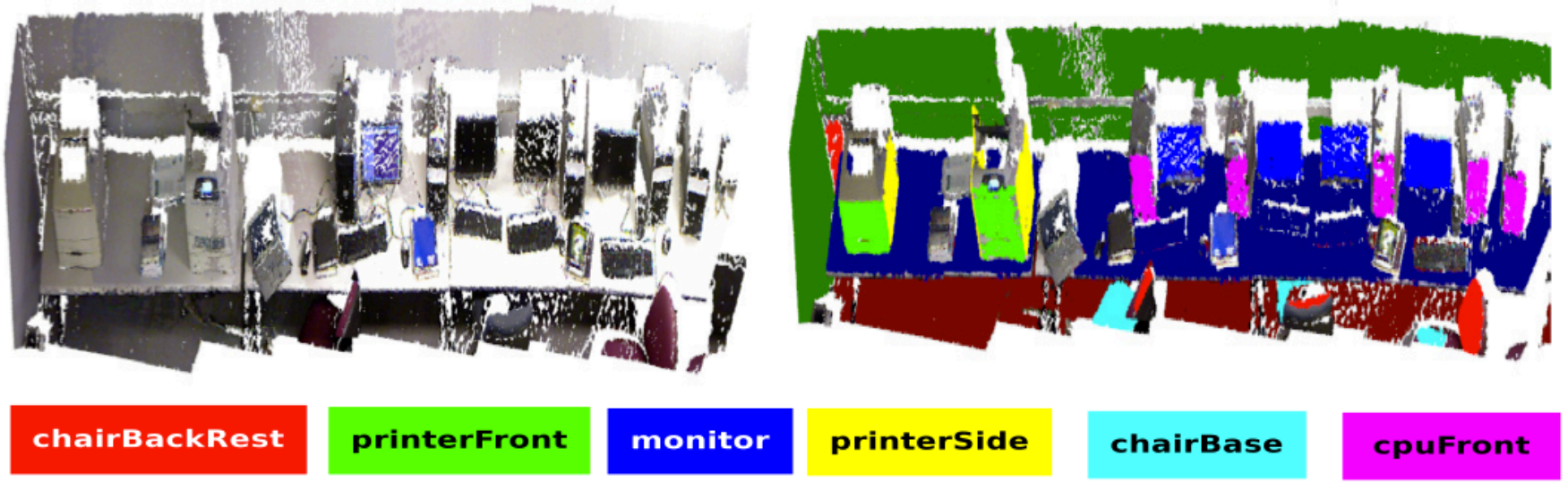
Download code at:

`http://pr.cs.cornell.edu/sceneunderstanding`


chairBackRest
chairBase
bed
shelfRack
bedSide
pillow


floor
wall
tableTop

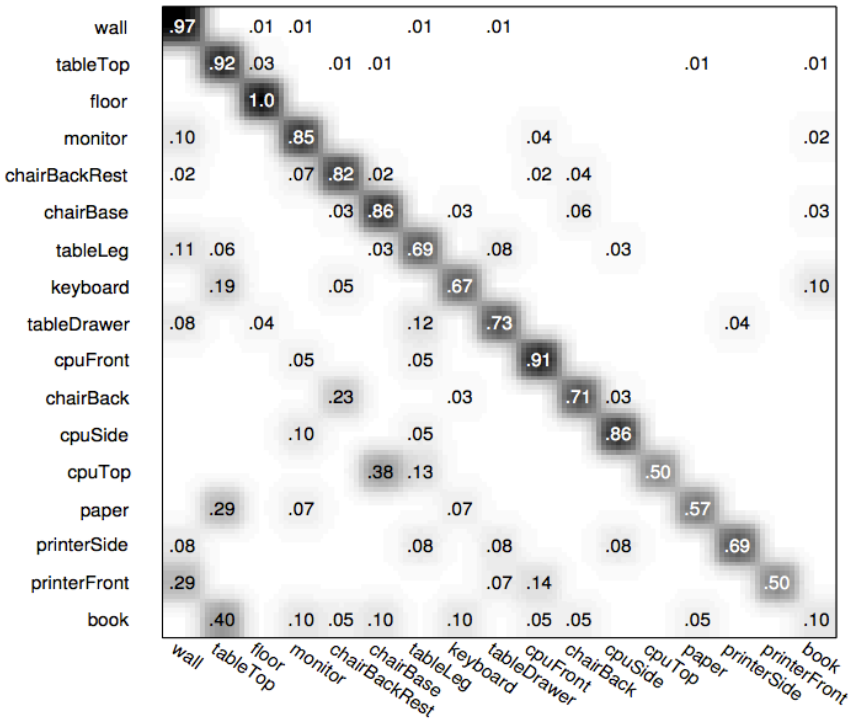
[[See Mesh.]]


chairBackRest
printerFront
monitor
printerSide
chairBase
cpuFront
floor
wall
tableTop

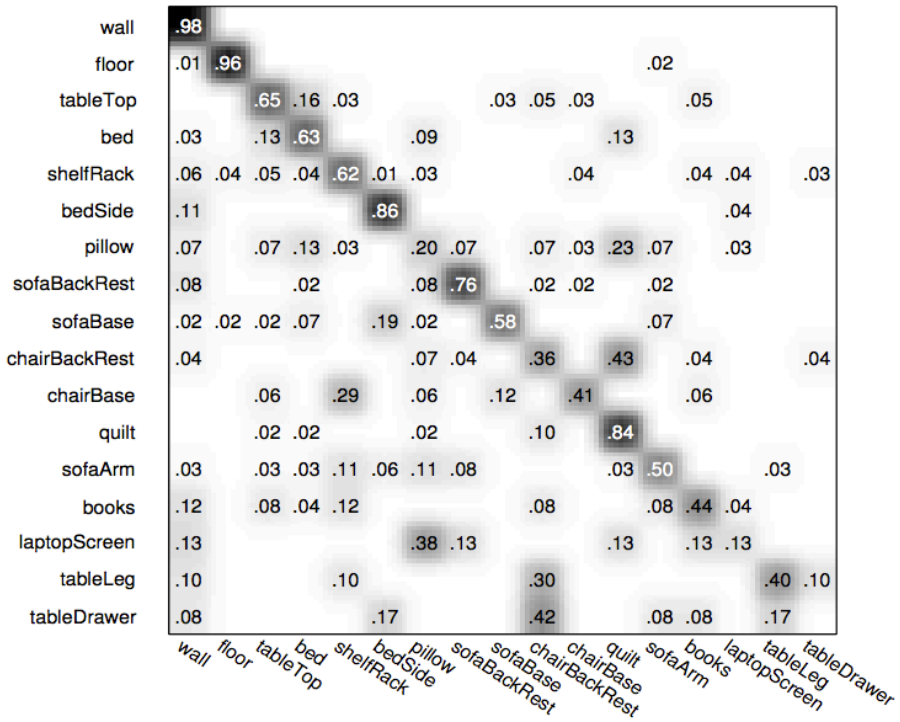
# Experiments

- 24 office scenes & 28 home scenes consisting of 550 views.

Confusion matrix for office scenes




Confusion matrix for home scenes



# Quantitative Results

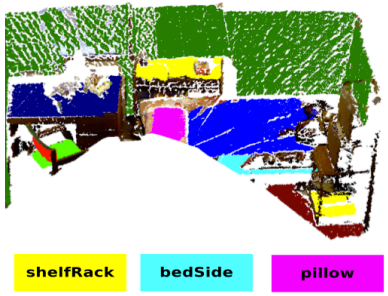
| | | Office Scenes | | | Home Scenes | | |
|---|---|---|---|---|---|---|---|
| | | micro | macro | | micro | macro | |
| features | algorithm | $P/R$ | Precision | Recall | $P/R$ | Precision | Recall |
| None | chance | 26.23 | 5.88 | 5.88 | 29.38 | 5.88 | 5.88 |
| Image Only | svm_node_only | 46.67 | 35.73 | 31.67 | 38.00 | 15.03 | 14.50 |
| Shape Only | svm_node_only | 75.36 | 64.56 | 60.88 | 56.25 | 35.90 | 36.52 |
| Image+Shape | svm_node_only | 77.97 | 69.44 | 66.23 | 56.50 | 37.18 | 34.73 |
| Image+Shape & context | single_frames | 84.32 | 77.84 | 68.12 | 69.13 | 47.84 | 43.62 |
| Image+Shape & context | svm_mrf_assoc | 75.94 | 63.89 | 61.79 | 62.50 | 44.65 | 38.34 |
| Image+Shape & context | svm_mrf_nonassoc | 81.45 | 76.79 | 70.07 | 72.38 | 57.82 | 53.62 |
| Image+Shape & context | svm_mrf_parsimon | 84.06 | 80.52 | 72.64 | 73.38 | 56.81 | 54.80 |

# Data and Code

For more details, data and code, visit:

http://pr.cs.cornell.edu/sceneunderstanding

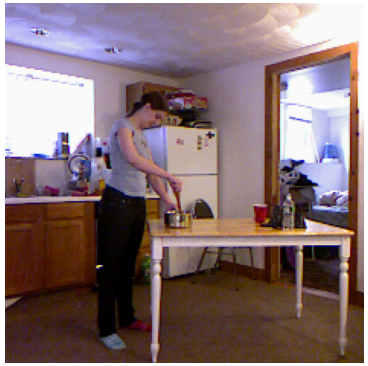
# Cornell Personal Robotics




Labeling 3D Scenes.

Learning Manipulation:

Placing, grasping, 3d articulated objects.



Human Activity Detection, Social Interaction.

More details, videos, data and code at:

http://pr.cs.cornell.edu

# Learning Manipulation Skills: Placing, Grasping, Articulated Objects.

Yun Jiang, Changxi Zheng, Marcus Lim, Paul Yang, Tiffany Low, Matthew Cong, Stephen Moseson, Ashutosh Saxena.

Cornell University.

# Placing Objects

- Daily tasks:
  - Setting a dinner table.
  - Arranging grocery in a fridge.
  - Packing items in boxes.

*Grasping and placing objects.*

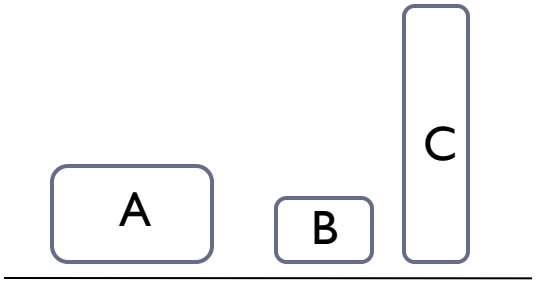


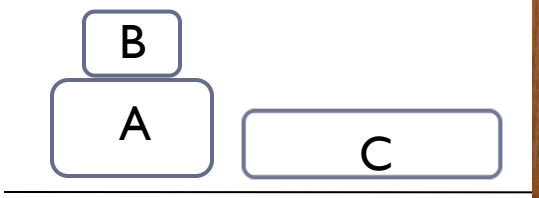
- Grasping: Significant improvements with RGB-D data:
  - Efficient Grasping from RGBD images: Learning using a new Rectangle Representation, Yun Jiang, Stephen Moseson, Ashutosh Saxena. In *ICRA*, 2011.

- Placing is an important skill for a personal robot.
  - However, it has heretofore been little-studied.

# Placing Objects: Challenges

- Different configurations in different placing situations.
  - Plate: vertical in a dish-rack; slanted on a support.
  - Glasses: upright on table, upside-down on stemware holder,
- Stacking objects.



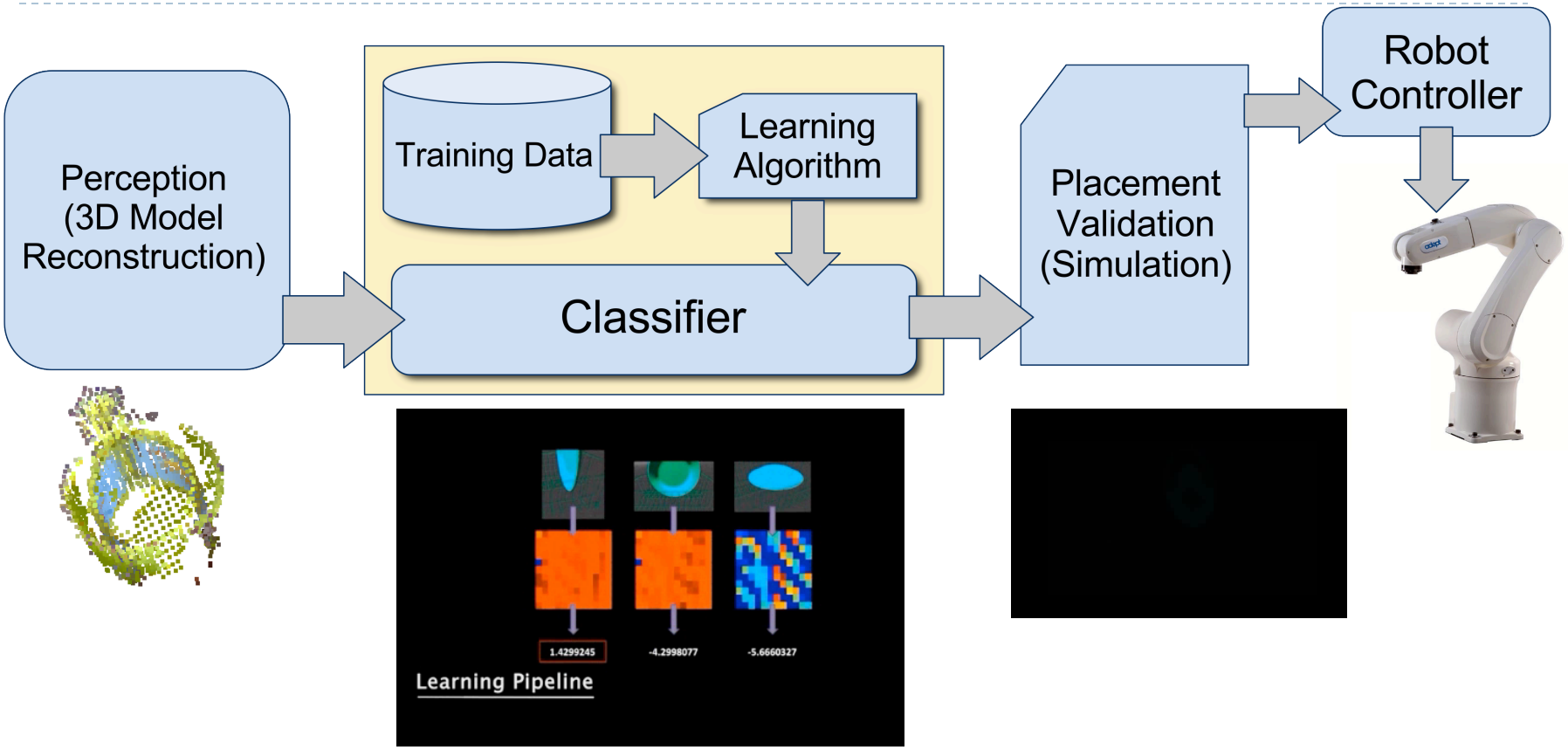




# Problem Specification

- Input: Point cloud of an object (e.g., plate) and placing environment (e.g., a dish rack).

- Output: a **stable** and **preferred** placement specified by 3D location and 3D orientation of the object.
  - Stability
    - stay still after placing and stand small perturbation.
  - Preference
    - E.g., plates and pens are flat on a table, but vertical in a dish-rack and a pen-holder.

# Learning Approach

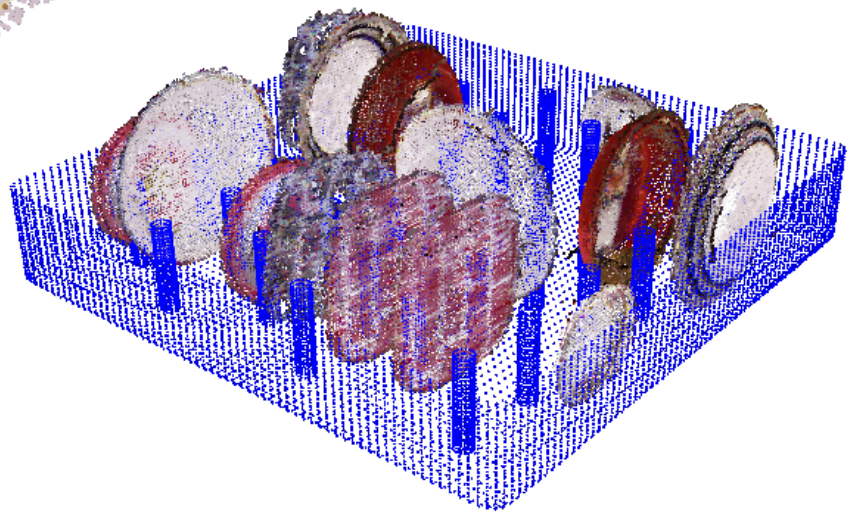


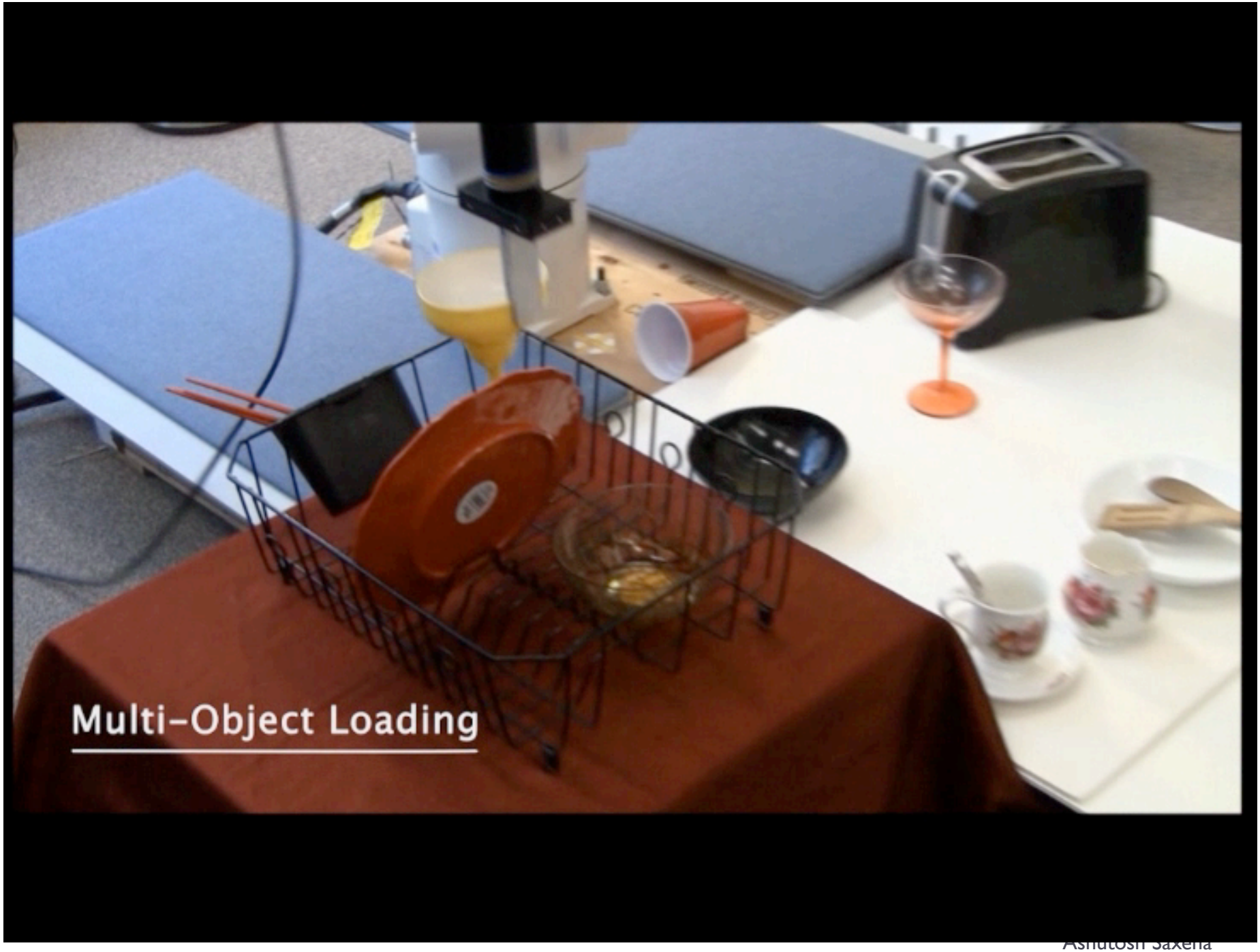

Supervised Learning.

- Features X for placement $\xi$
- Learning algorithm: X → y
- Choose $\xi$ with highest y.

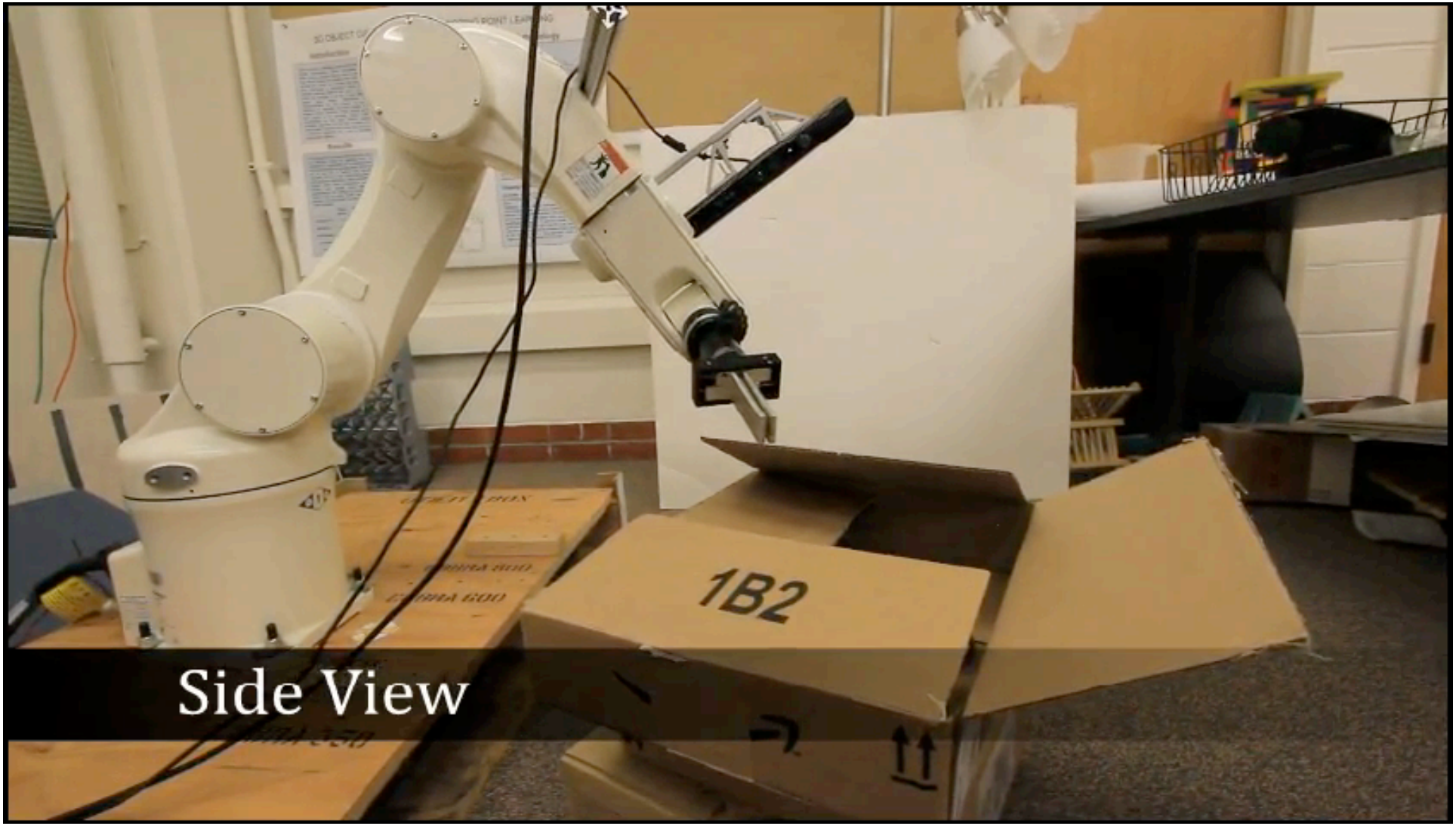
**Learning to Place New Objects**, Yun Jiang, Changxi Zheng, Marcus Lim, Ashutosh Saxena. In *RSS workshop on mobile manipulation*, June 2011.

# Multiple Object: Results




Multi-Object Loading

# Inferring 3D Articulated Structures.




**Inferring 3D Articulated Models for Box Packaging Robot**, Paul Heran Yang, Tiffany Low, Matthew Cong, Ashutosh Saxena. In *RSS workshop on mobile manipulation*, 2011.

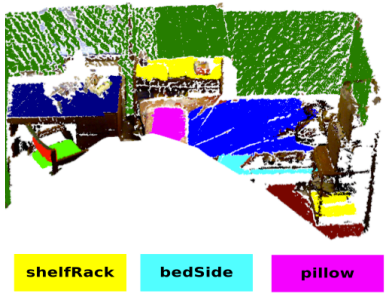
# Questions?

Download code and data at:

`http://pr.cs.cornell.edu/placingobjects`

`(Cornell Personal Robotics.)`

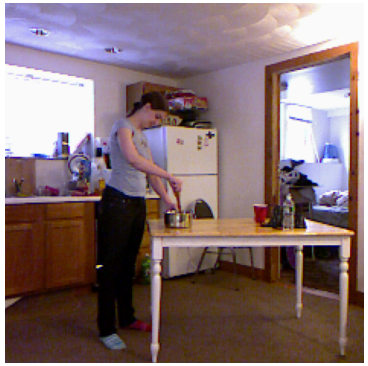
# Cornell Personal Robotics




Labeling 3D Scenes.

Learning Manipulation:

Placing, grasping, 3d articulated objects.



Human Activity Detection, Social Interaction.

More details, videos, data and code at:

http://pr.cs.cornell.edu

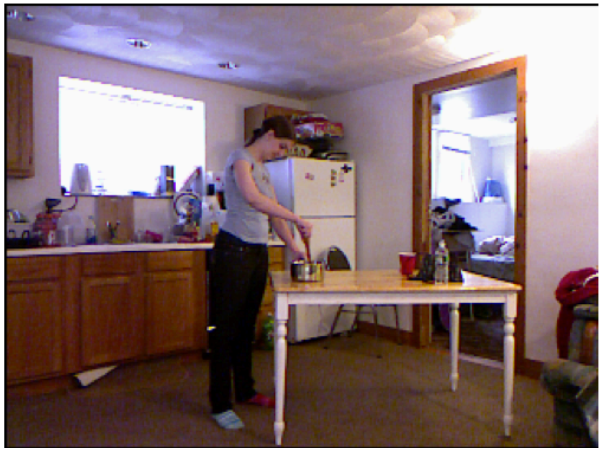
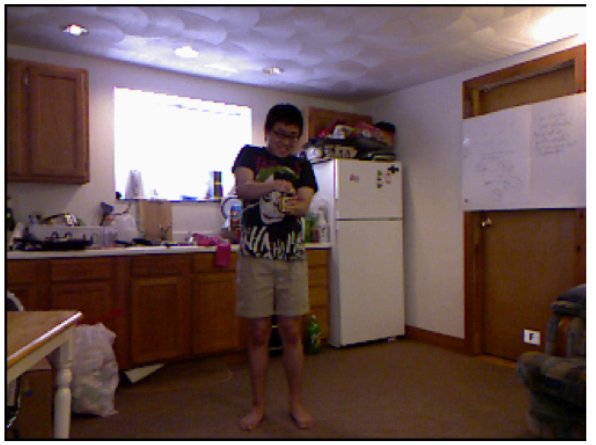
# Human Activity Detection

Jae Y. Sung, Colin Ponce, Bart Selman, Ashutosh Saxena.

Cornell University.

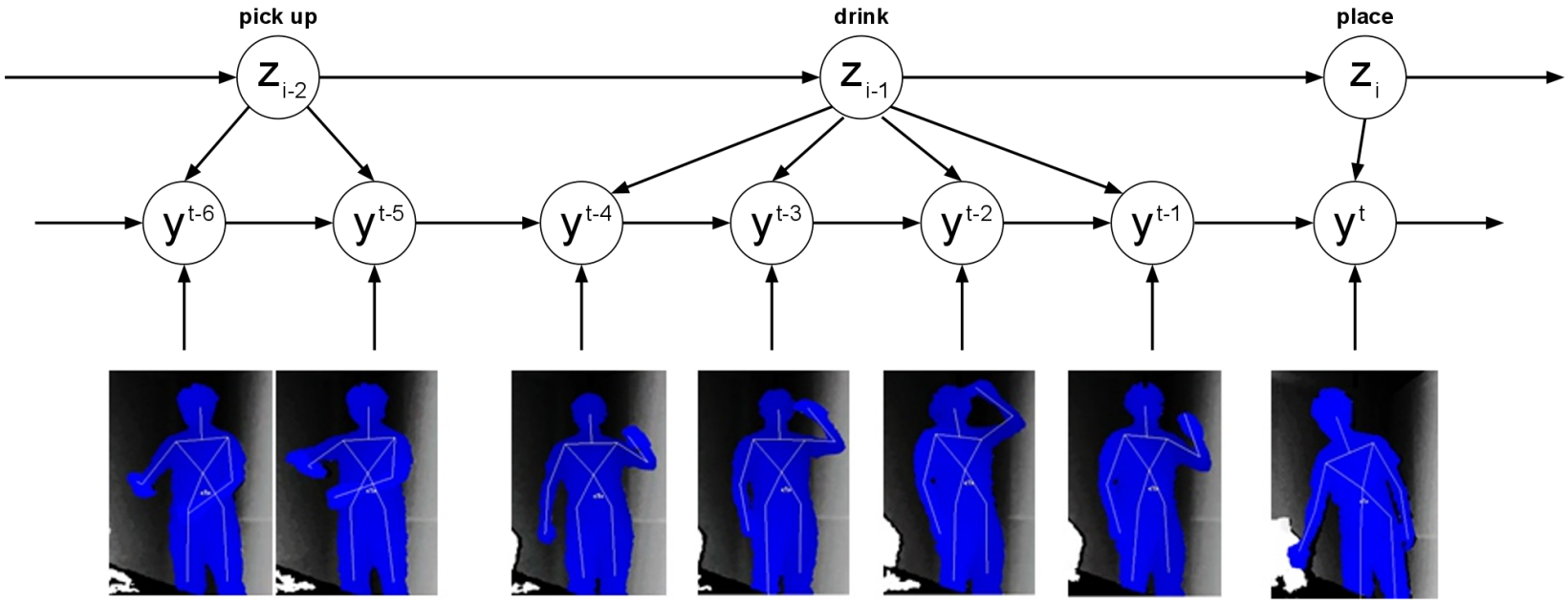
# Detect Common Activities

- More natural human robot interaction.
- Challenges
  - Cluttered Environments, new subjects.
  - Detect activities: *rinsing mouth, brushing teeth, wear contact lens, talk on phone, drink water, open pill container, cooking-stirring, cooking-stirring, talking on couch, relaxing on couch, write on whiteboard, work on computer.*
  - Against other ***random*** activities.

# Algorithm

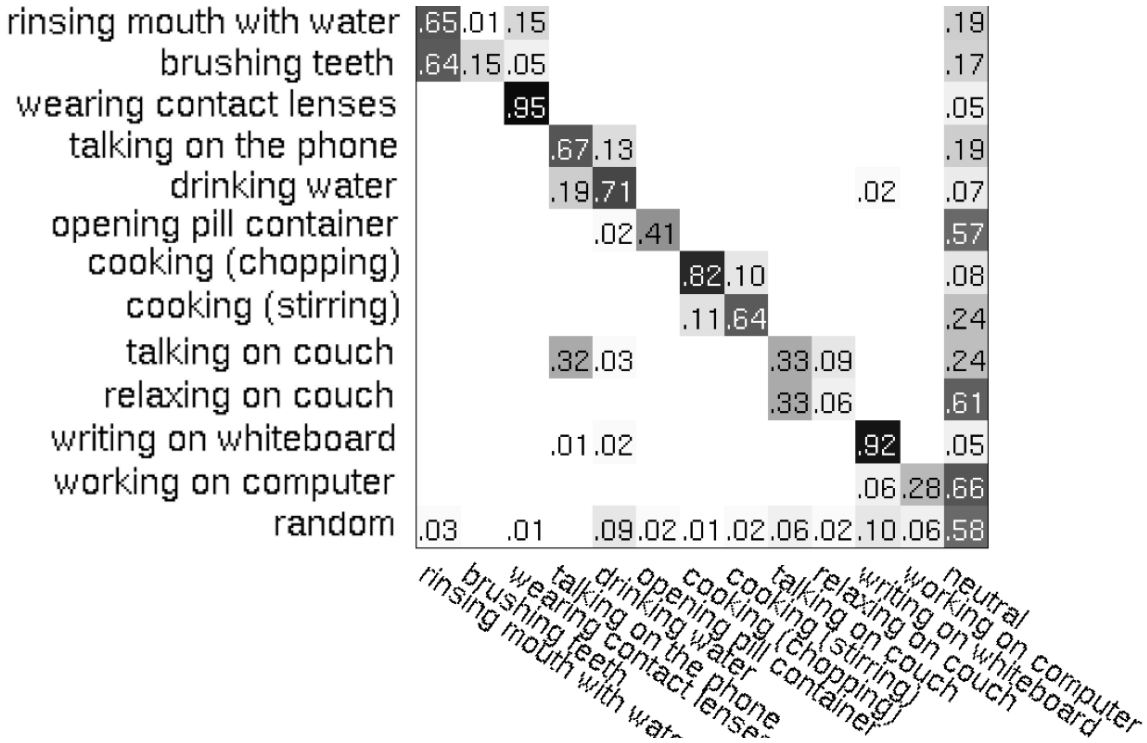
- Hierarchical MEMM model.




**Human Activity Detection from RGBD Images**, Jae Y. Sung, Colin Ponce, Bart Selman, Ashutosh Saxena. *In AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.

# Activity Detection: Results

- Seen Person: 84.3% ($F_{0.5}$ score).
- New Person: 64.2% ($F_{0.5}$ score).



Human Activity Detection from RGBD Images.

Jae Y. Sung, Colin Ponce, Bart Selman, Ashutosh Saxena.

In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.

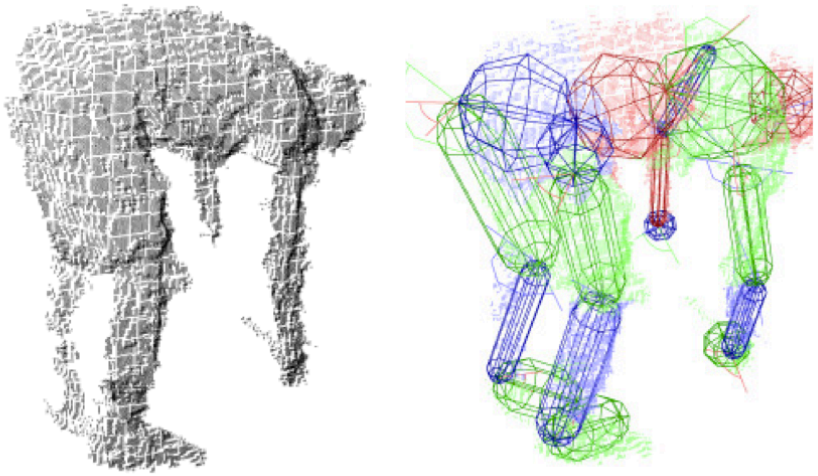
More details/code/data at:

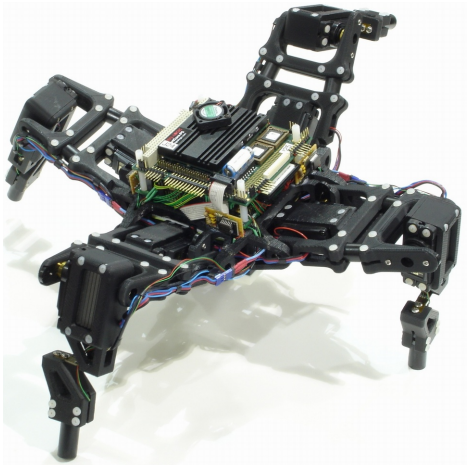
`http://pr.cs.cornell.edu/humanactivities`

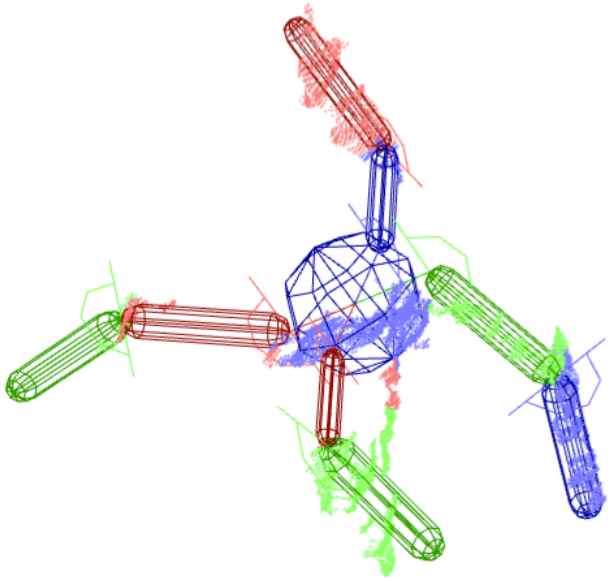
# Extracting Kinematic Models

See Daniel Ly's poster today.

**Pose estimation from a single depth image for arbitrary kinematic skeletons**, Daniel Ly, Ashutosh Saxena, Hod Lipson. In *R:SS workshop on RGB-D cameras*, 2011.

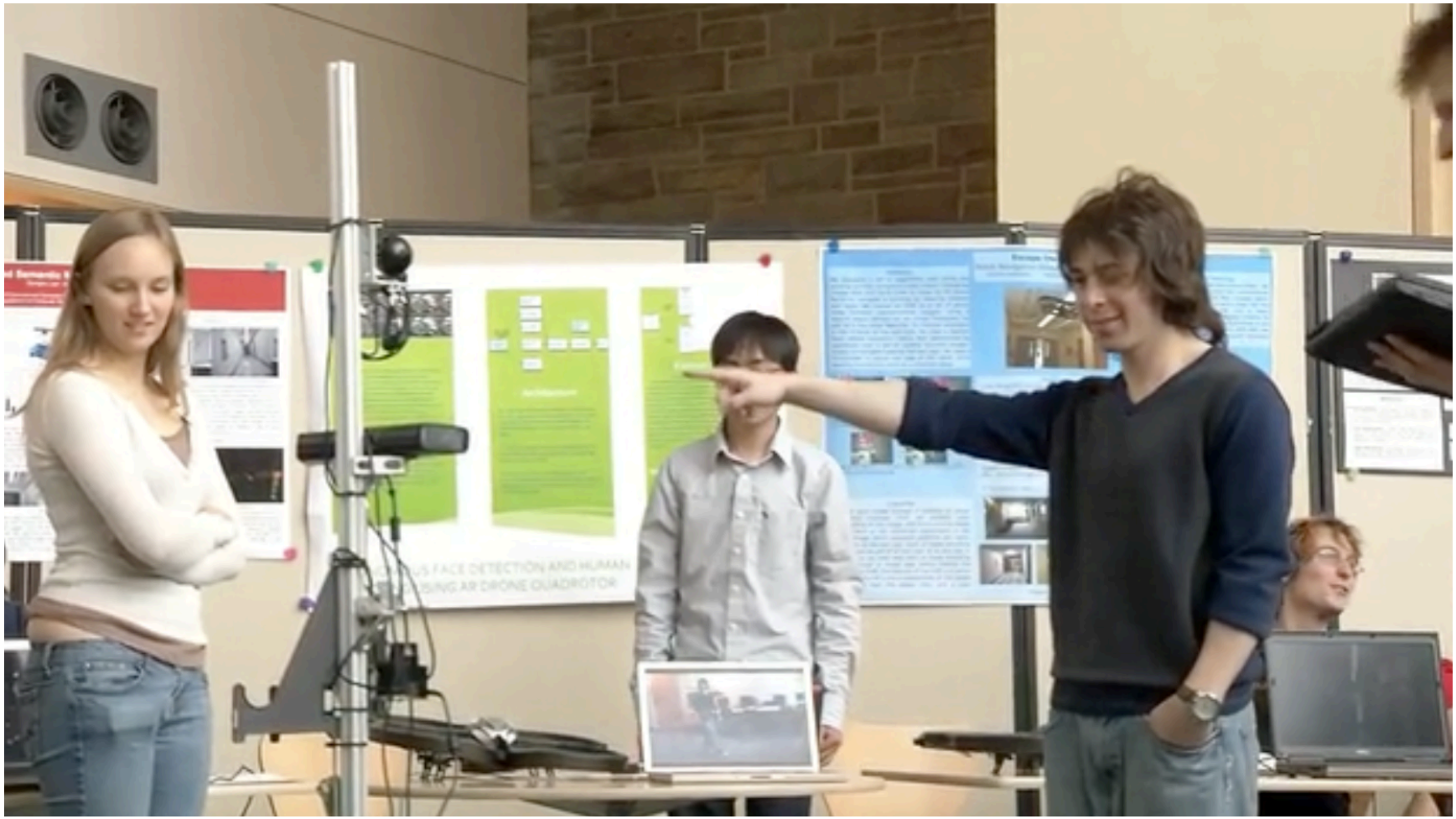






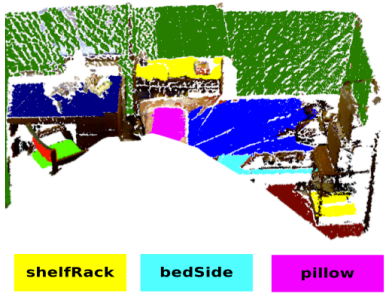
# Manipulating Humans!

See the Blue Robot in the corridors at RSS!



Ashutosh Saxena

# Cornell Personal Robotics

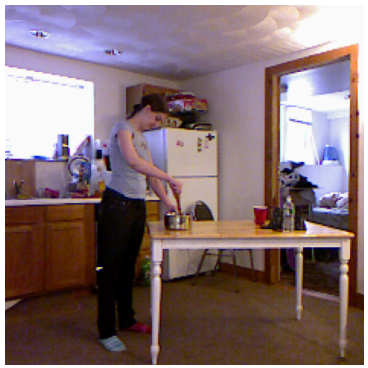



Labeling 3D Scenes.

Hema Koppula, Abhishek Anand, Thorsten Joachims.

Learning Manipulation: Placing, grasping, articulated objects.

Yun Jiang, Marcus Lim, Changxi Zheng, Stephen Moseson, Matthew Cong, Paul H. Yang, Tiffany Low.



Human Activity Detection, Skeletal Extraction, Social Interaction.

Jae Y. Sung, Colin Ponce, Bart Selman, Igor Labutov, Jason Yosinski, Daniel Ly.

`http://pr.cs.cornell.edu`

# Questions?

Cornell Personal Robotics:

http://pr.cs.cornell.edu

Thank you