Dieter Fox
University of Washington

Some Experiences with RGB-D Perception in Robotics

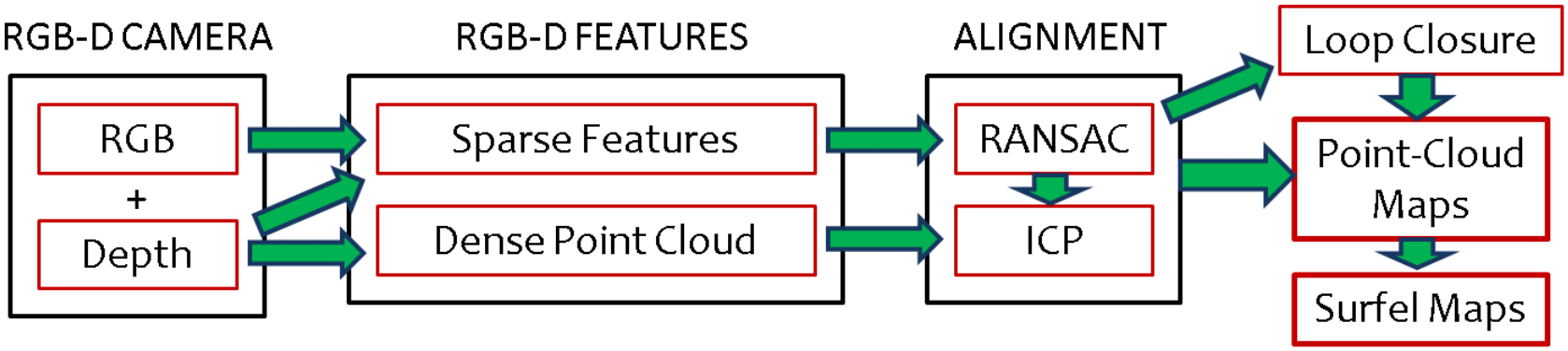
Joint work with
Peter Henry, Evan Herbst, Mike Krainin, Kevin Lai ,
Brian Curless, Liefeng Bo, Xiaofeng Ren, Richard Newcombe

Outline

- Environments
- Objects
- People
- Discussion

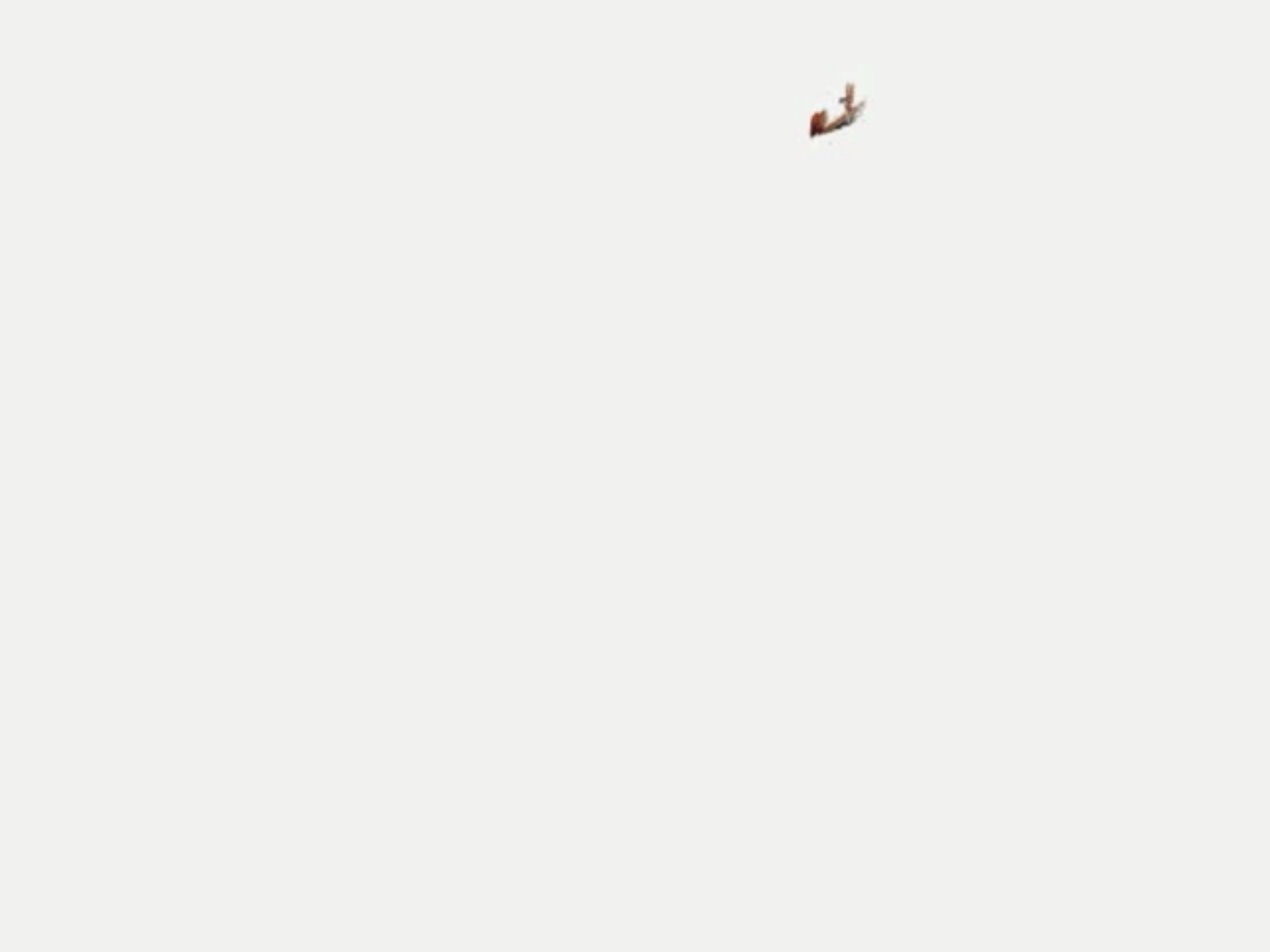
[Henry-Krainin-Herbst-Ren-F: ISER-10, IJRR-12]

RGB-D Mapping



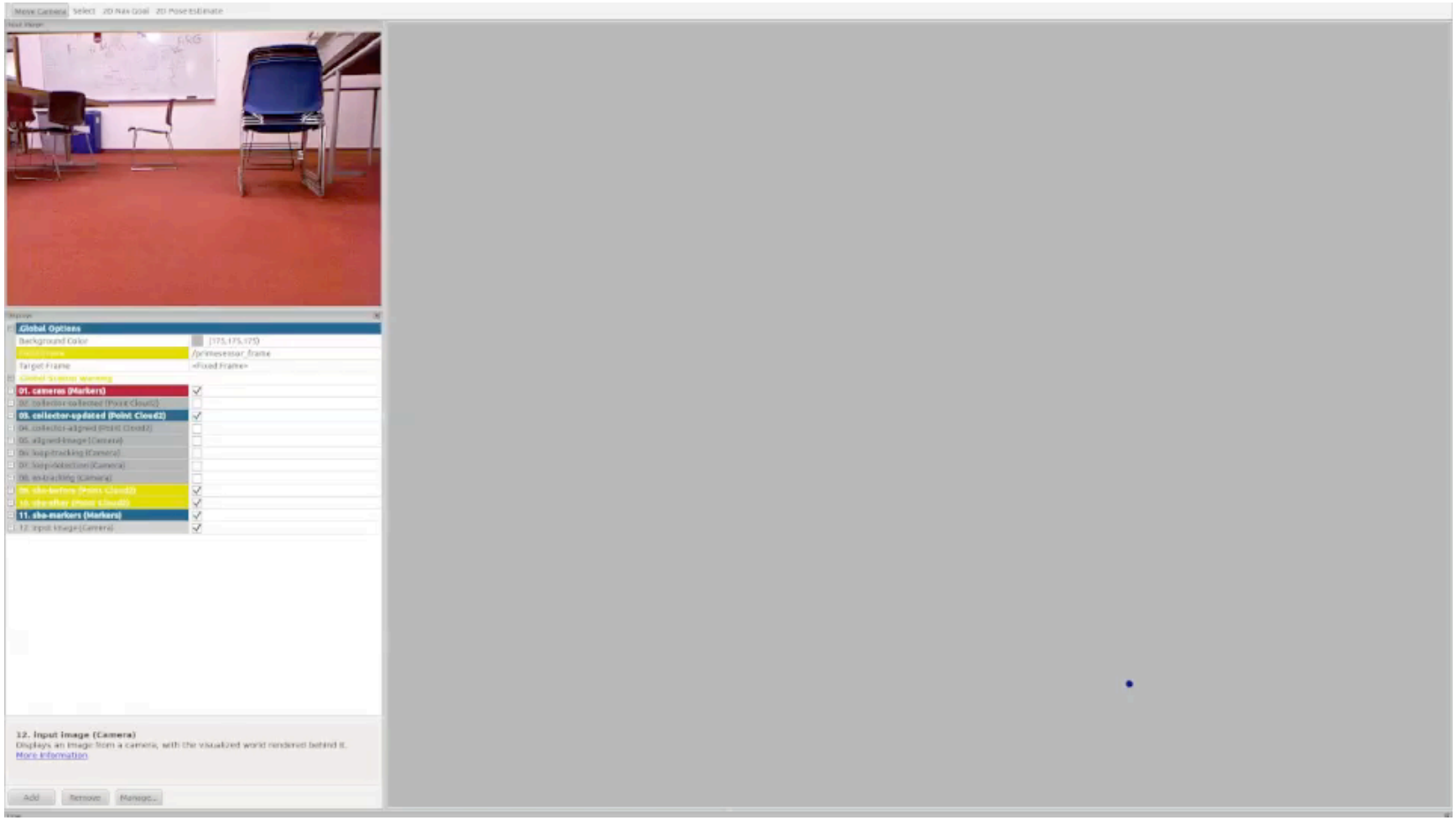
- Frame-to-frame alignment
- Loop closure detection (view based)
- Global alignment (TORO, SBA, G2O,…)

Example Mapping Run



[Huang-Roy-Bachrach-Henry-Krainin-Maturana-F: ISRR-11, IJRR-12]

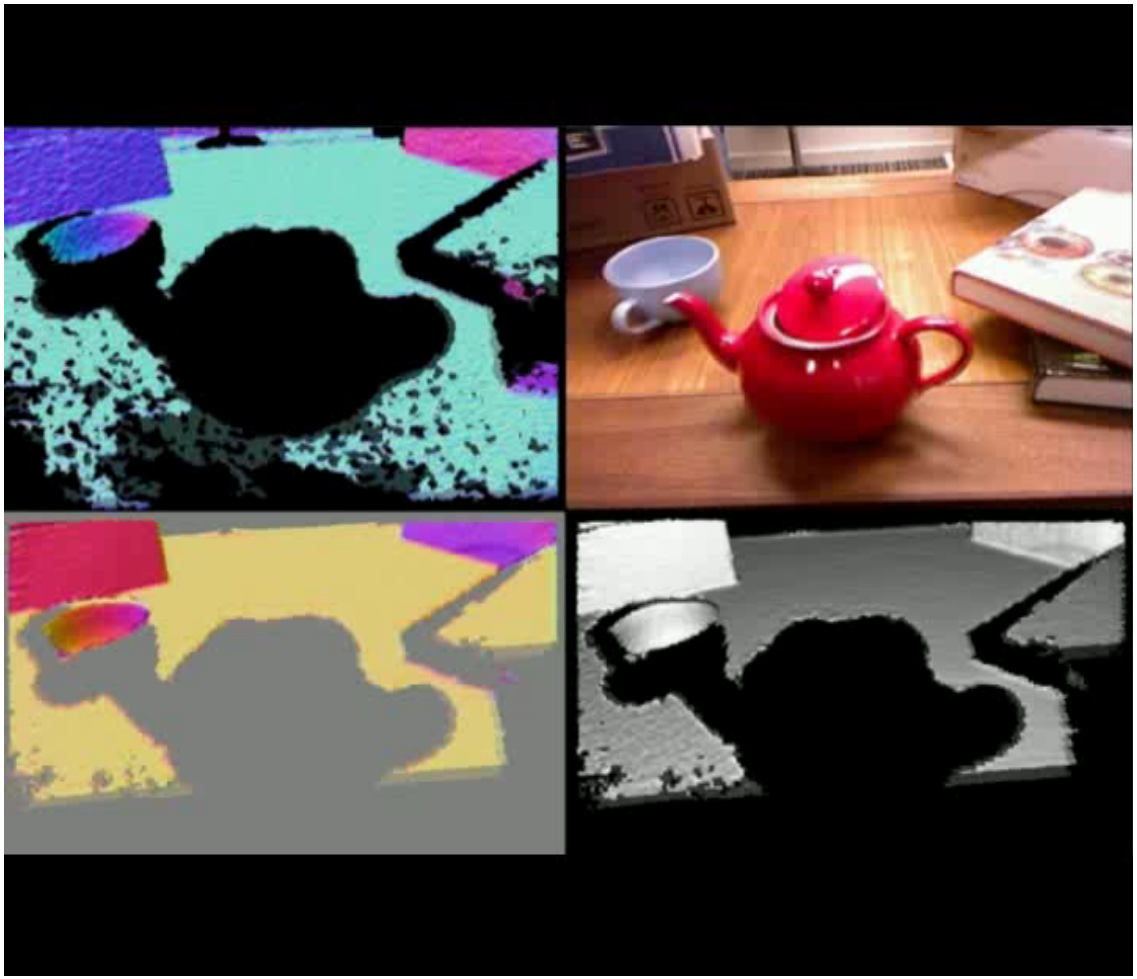
RGB-D SLAM on Quadcopter



[Newcombe etal: ISMAR-11]

KinectFusion

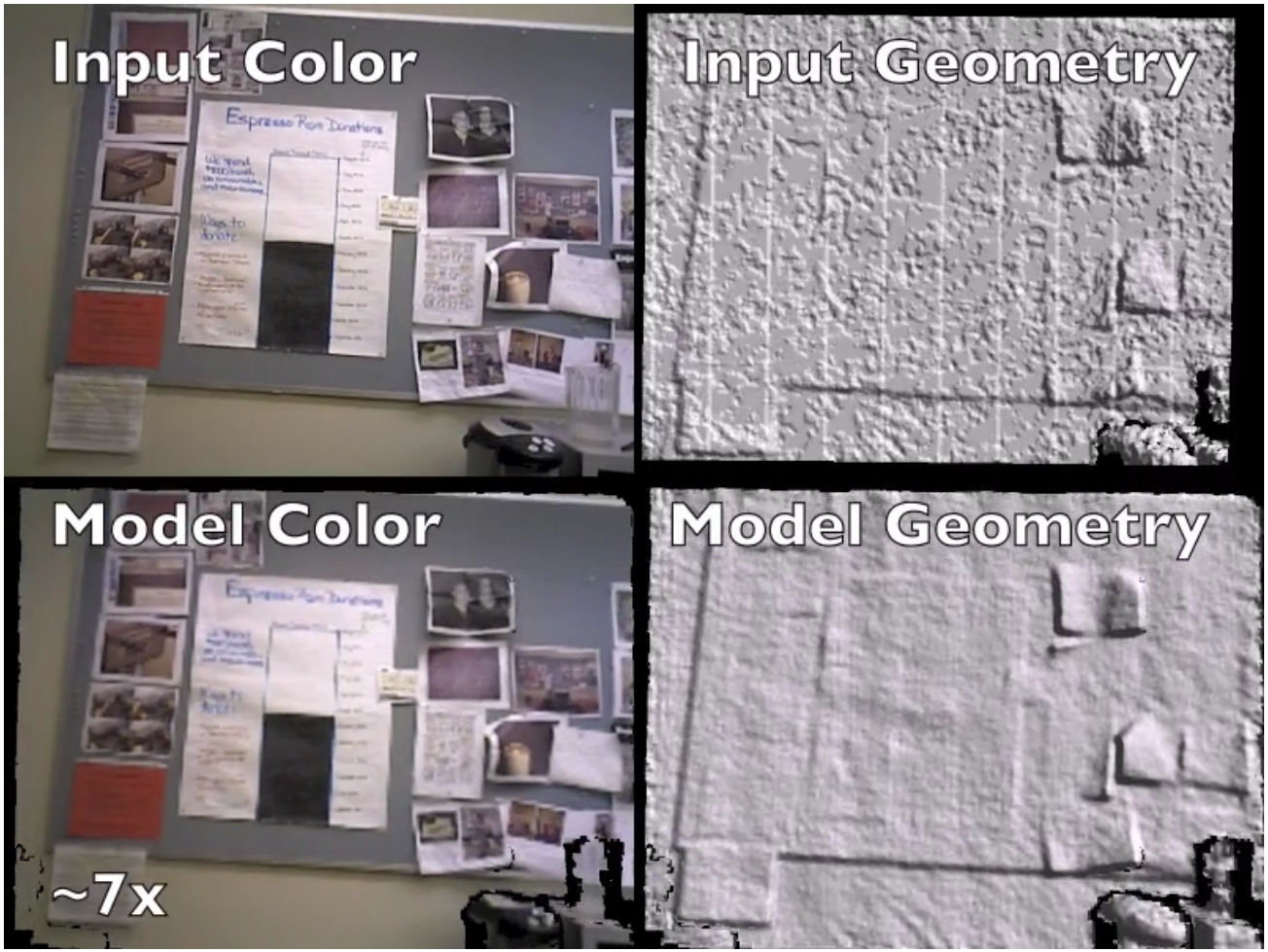
- Implicit surface
- Frame to model
- GPU-optimized
- Limited size



Curless etal: SIGGRAPH-96

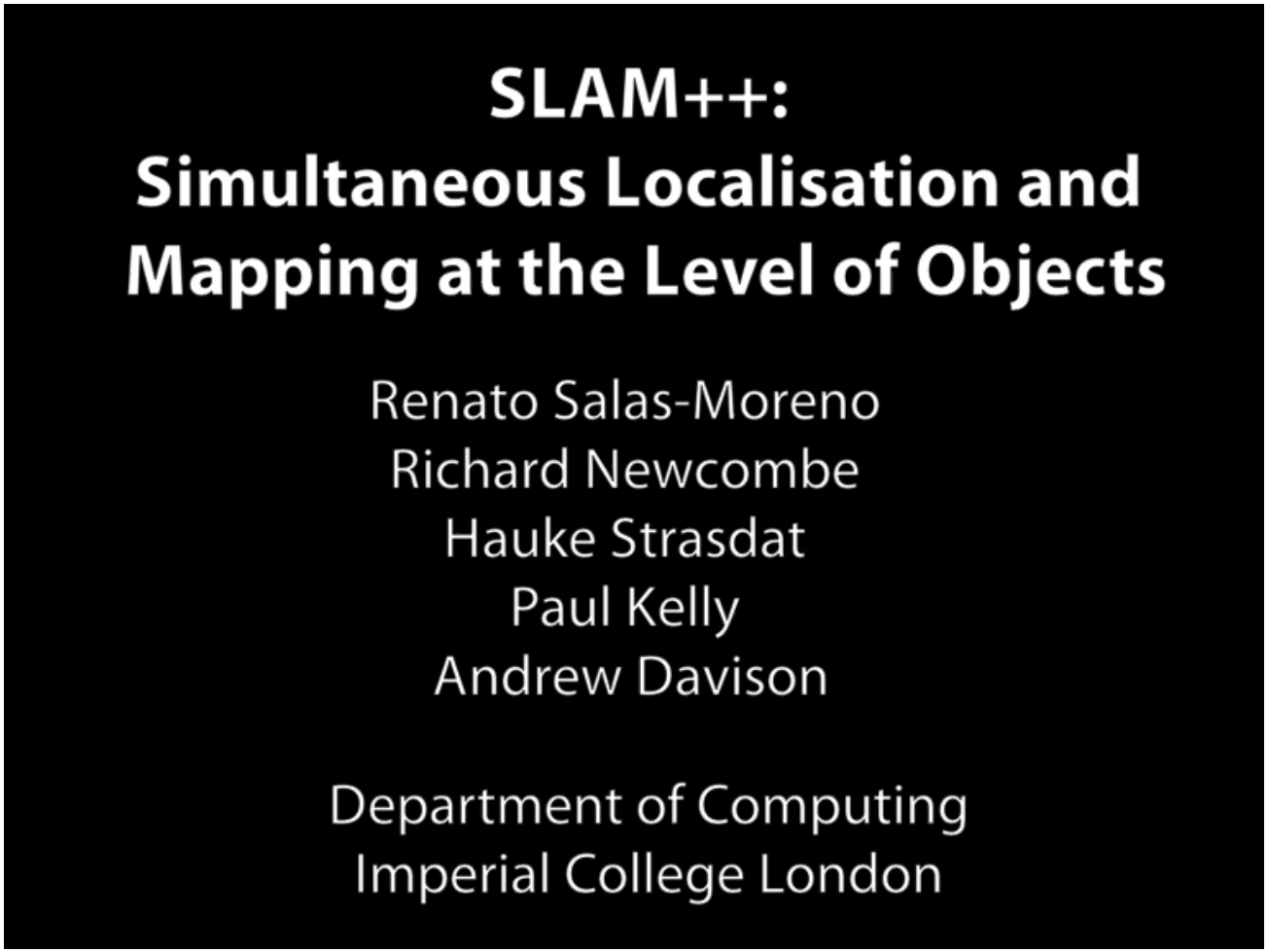
Patch Volumes: KinectFusion with Loop Closure

[Henry-F-Bhowmik-Mongia: 3DV-13]



SLAM++
SLAM at the Level of Objects

[SalasMoreno-Newcombe-etal: CVPR-13]



Mapping

- Current focus:
  - Scaling up both in space and time
  - Quality of reconstruction (calibration, alignment, …)
  - Dealing with and detecting changes in the environment
- Exploration: where to go, how much detail
- Representation: SDF, mesh, octree, point clouds, geometric primitives, object models?
- Semantics vs. geometry

Outline

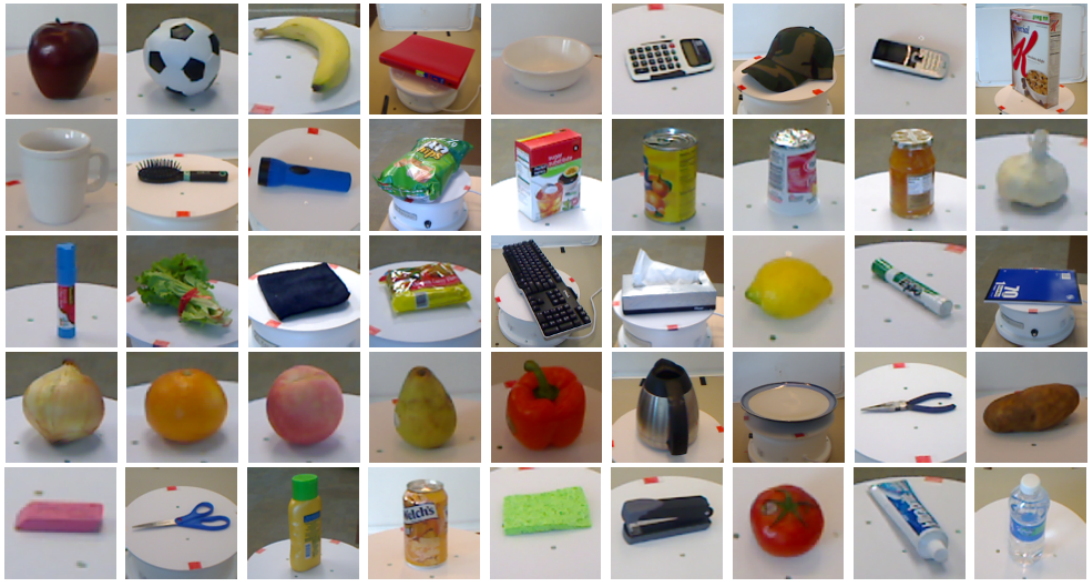
- Environments
- Objects
  - Recognition and Detection
  - Modeling and Manipulation
- People
- Discussion

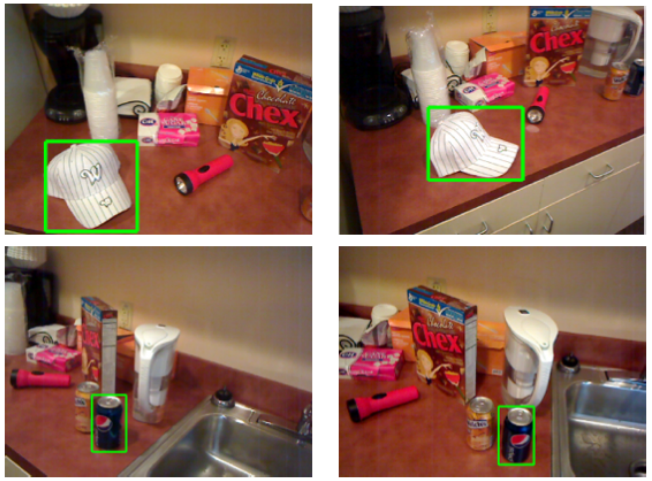
Outline

- Environments
- Objects
  - Recognition and Detection
  - Modeling and Manipulation
- People
- Discussion

[Lai-Bo-Ren-F: ICRA-11]

RGB-D Object Dataset

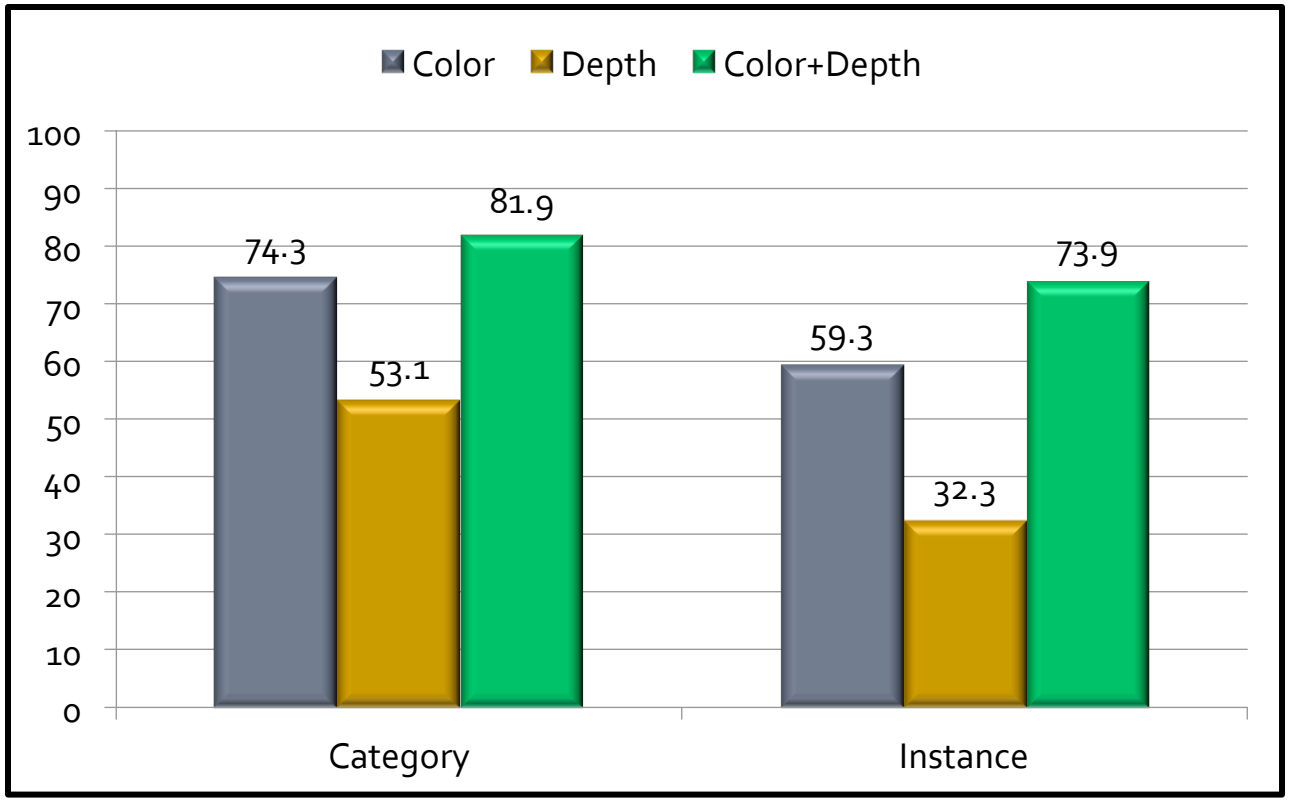




- Videos of 300 objects in 51 categories
- Take advantage of depth for segmentation
- Similar dataset by Willow Garage

Standard Features

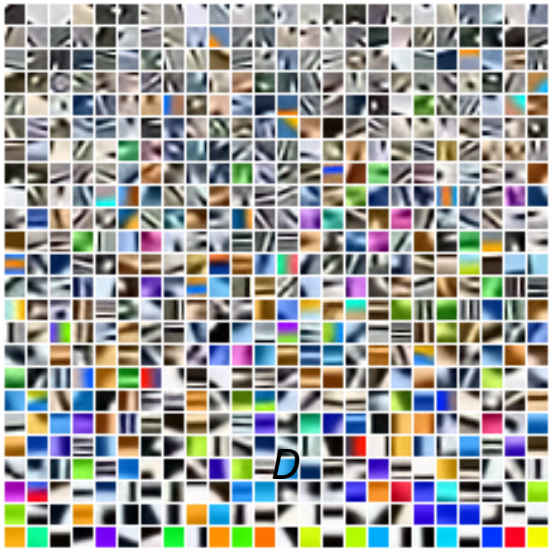
(Spin Images, SIFT, Bounding box)



[Bo-Ren-F: NIPS-11, ISER-12, CVPR-13, IJRR-14]

Hierarchical Matching Pursuit

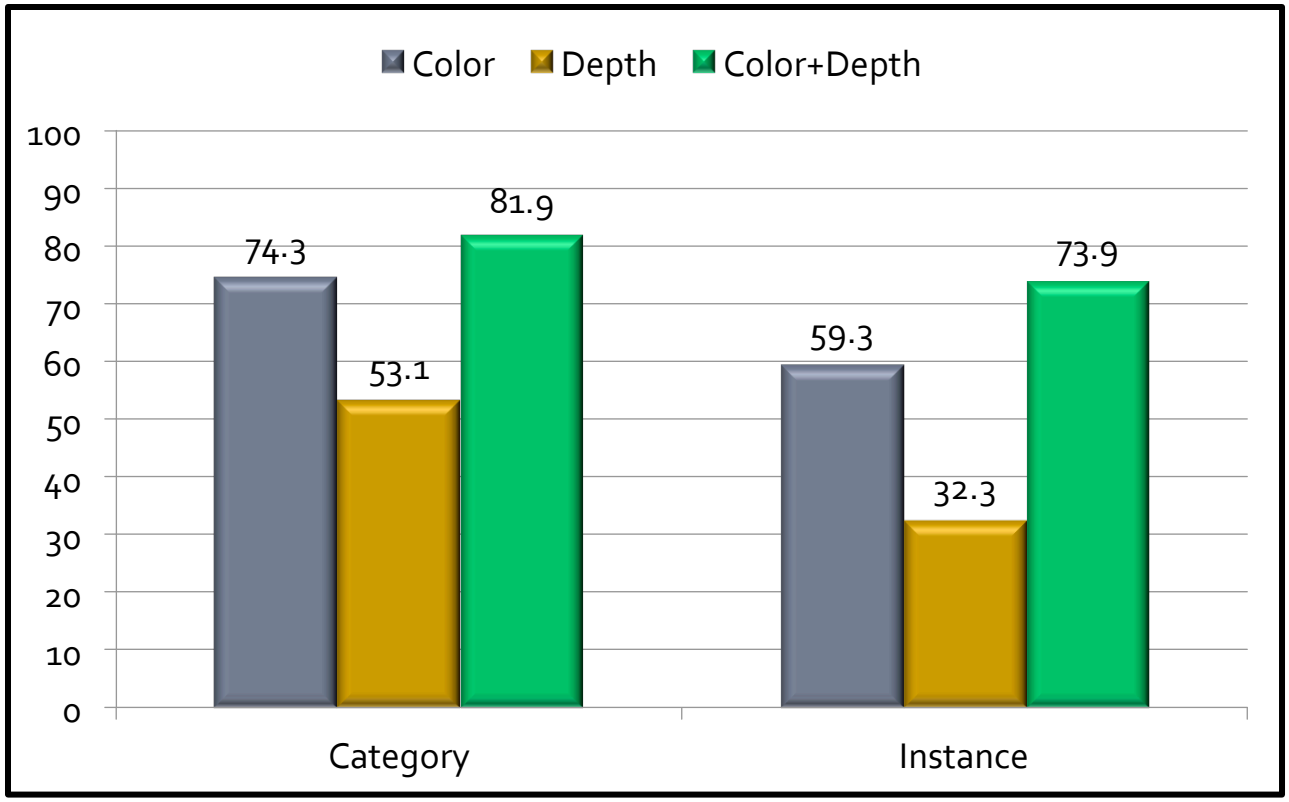
- Unsupervised feature learning
  - Sample images and patches
  - Use K-SVD to learn color and depth dictionaries over
    - small image patches (level 1) and
    - level 1 sparse codes pooled over image regions (level2)



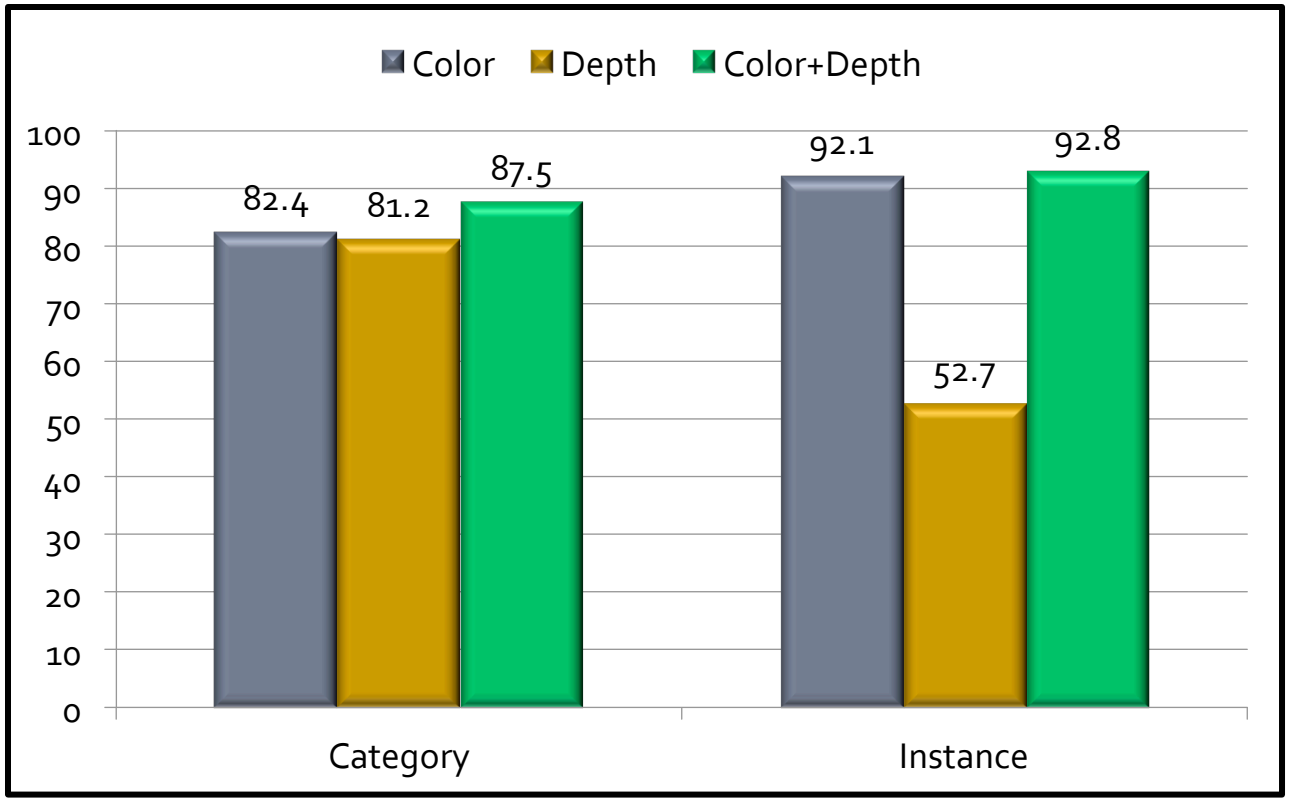
- Classification
  - Compute 2 level sparse code over image / segment (> 100,000 dims)
  - Train / evaluate linear SVM

Standard Features

(Spin Images, SIFT, Textons, Bounding box)



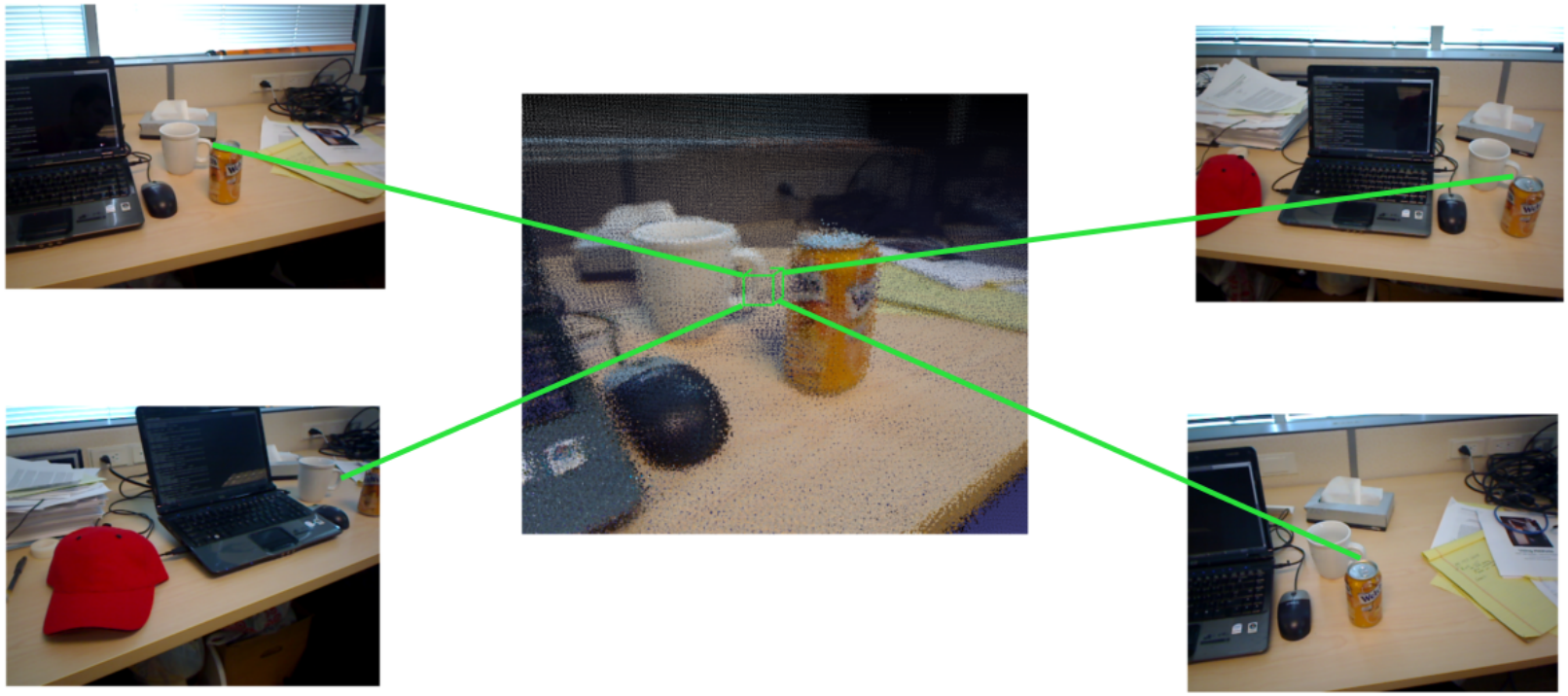
HMP Features



- VFH [RAM-12]: Category shape depth 56.0
- State-of-the-art on STL-10, Caltech-101/256, UCSD-Birds, MITScenes-67
- Excellent results on detection and scene labeling [Ren-etal: NIPS-12, CVPR-13]

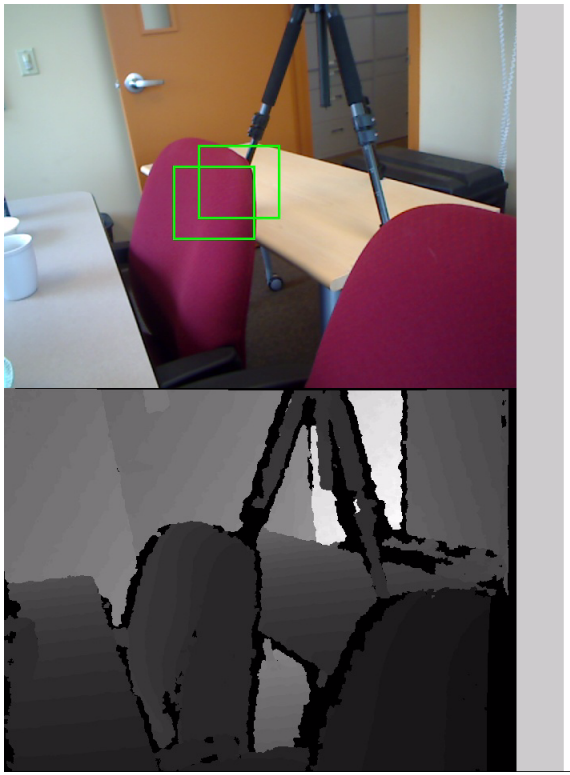
3D Scene Labeling from RGB-D Videos

[Lai-Bo-Ren-F: ICRA-12,ICRA-14]



- Input: RGB-D video sequence
- Build map and accumulate detection scores in 3D voxels
- Slide HMP shape features over 3D map
- Label voxels using shape-dependent MRF

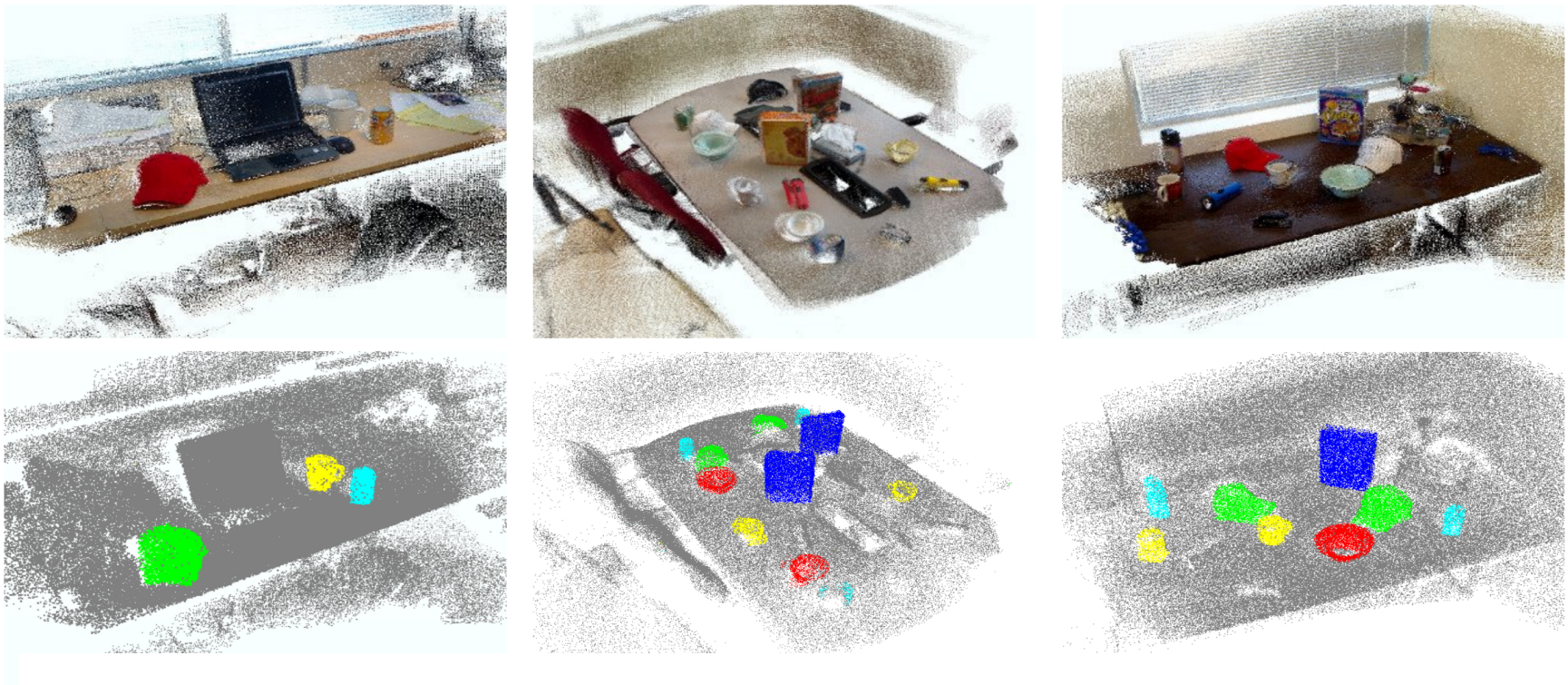
3D Object Labeling

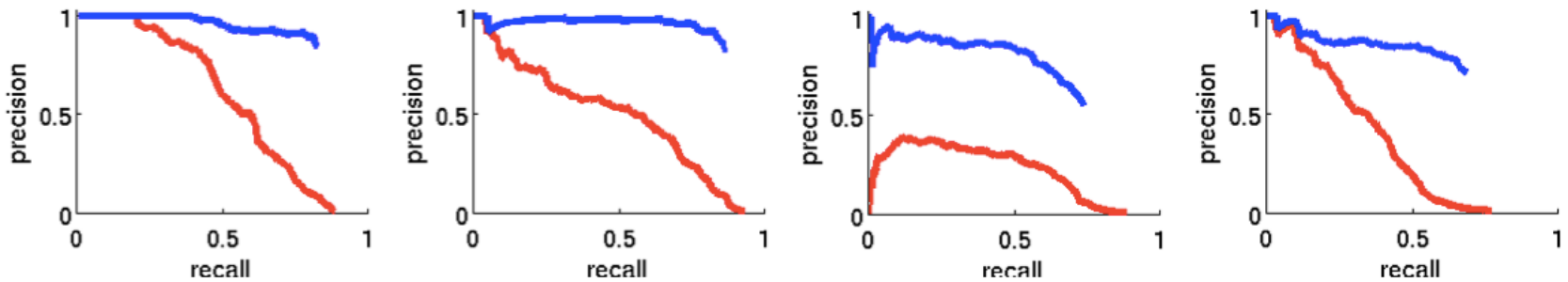


Cap
Cereal box
Soda can
Mug
Bowl

Background

Example Results

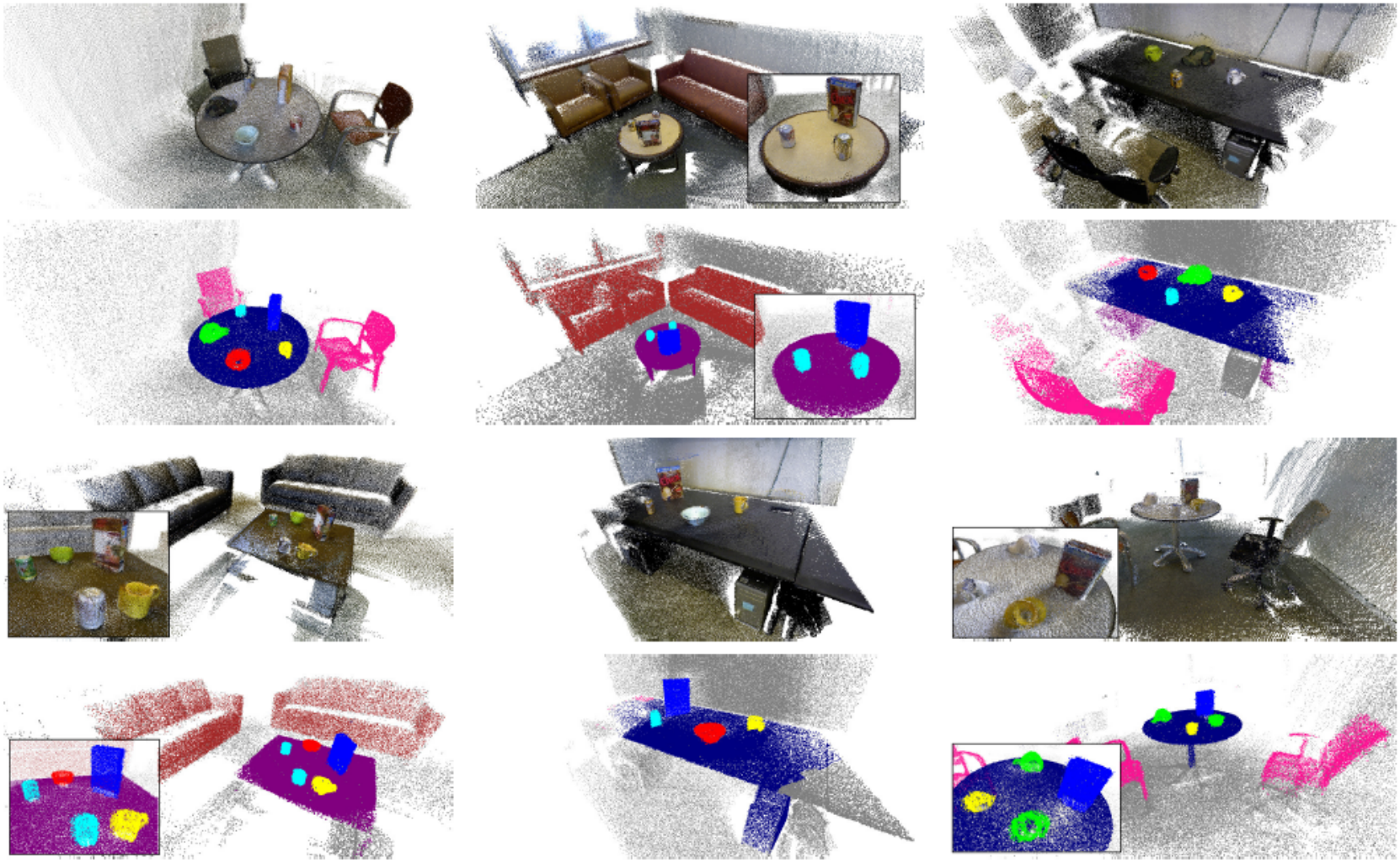




Living Room Scene

Chair
Cap
Coffee Table
Cereal Box
Table
Coffee Mug
Sofa
Soda Can
Bowl
Background

Scene Examples > 90% prec/recall



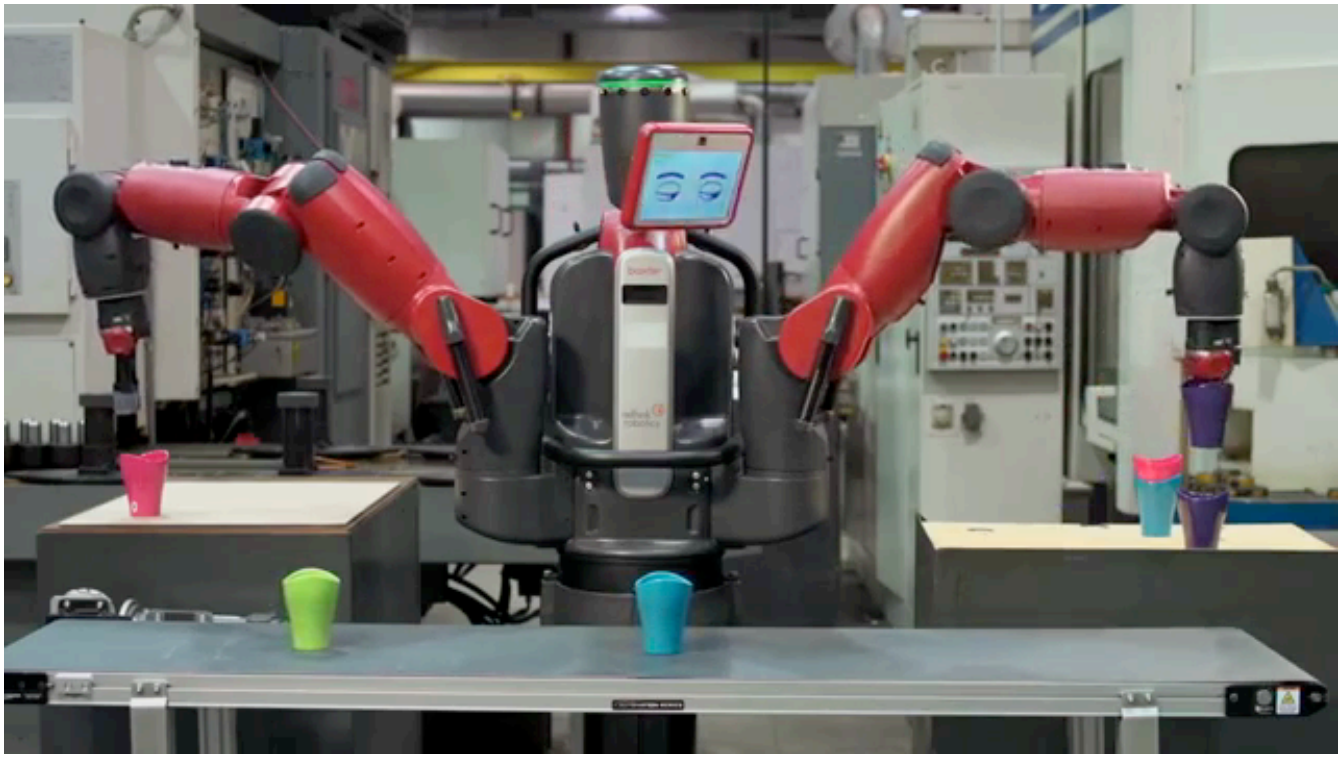
Object Recognition and Detection

- Good features for color and depth available
- Integration over time and space is beneficial
- Depth not crucial for recognition under benign conditions but provides shape context for detection and scene labeling
- Still no off-the-shelf systems out there
  - more fully labeled 3D datasets
  - not researchy enough?

Outline

- Environments
- Objects
  - Recognition and Detection
  - Modeling and Manipulation
- People
- Discussion

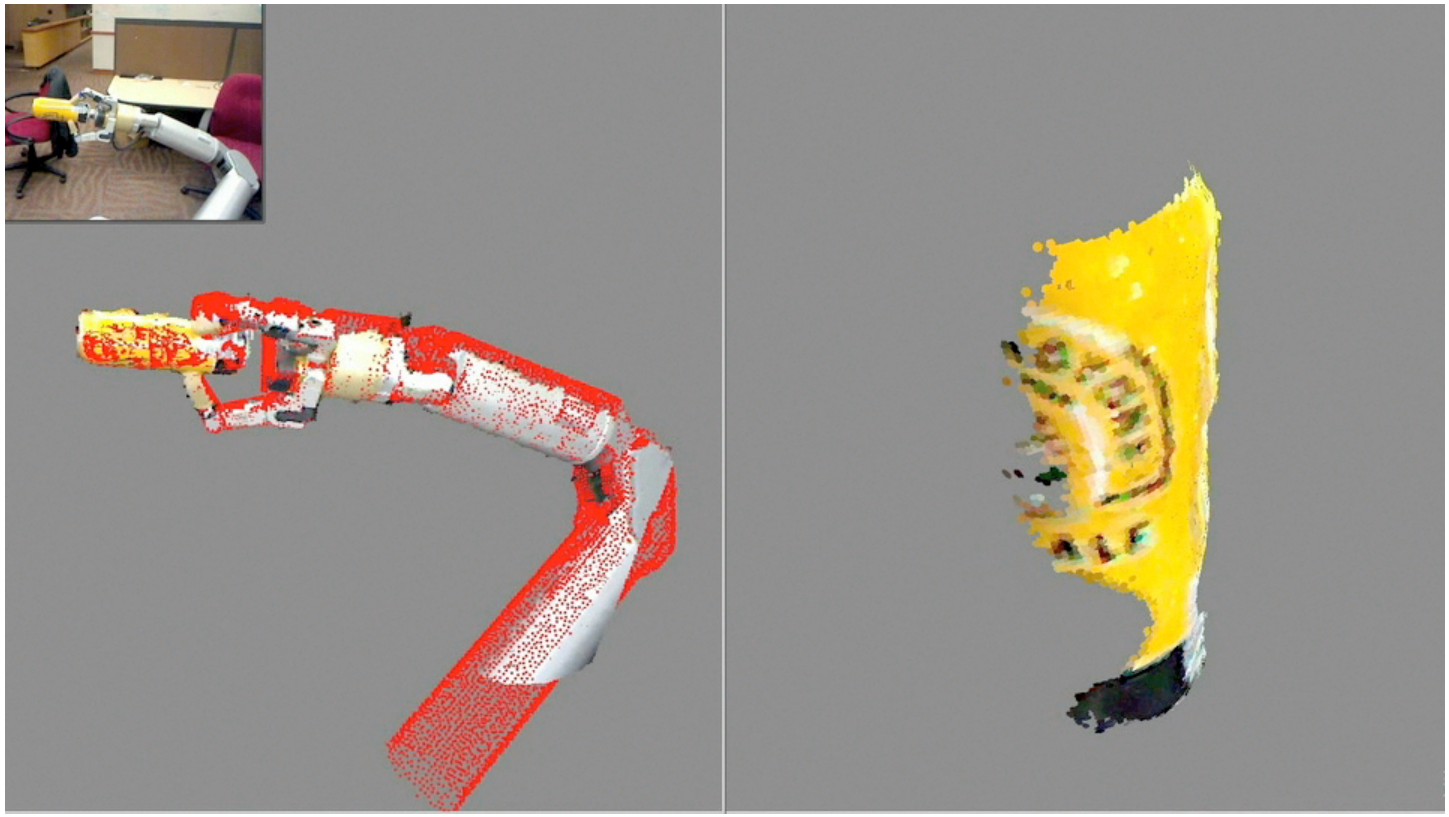
Manipulation



- Uncertain object pose and shape
- Uncertain manipulator pose due to cable stretch
- Where to grasp

[Krainin-Henry-Ren-F: IJRR-11]

Joint Tracking and Modeling

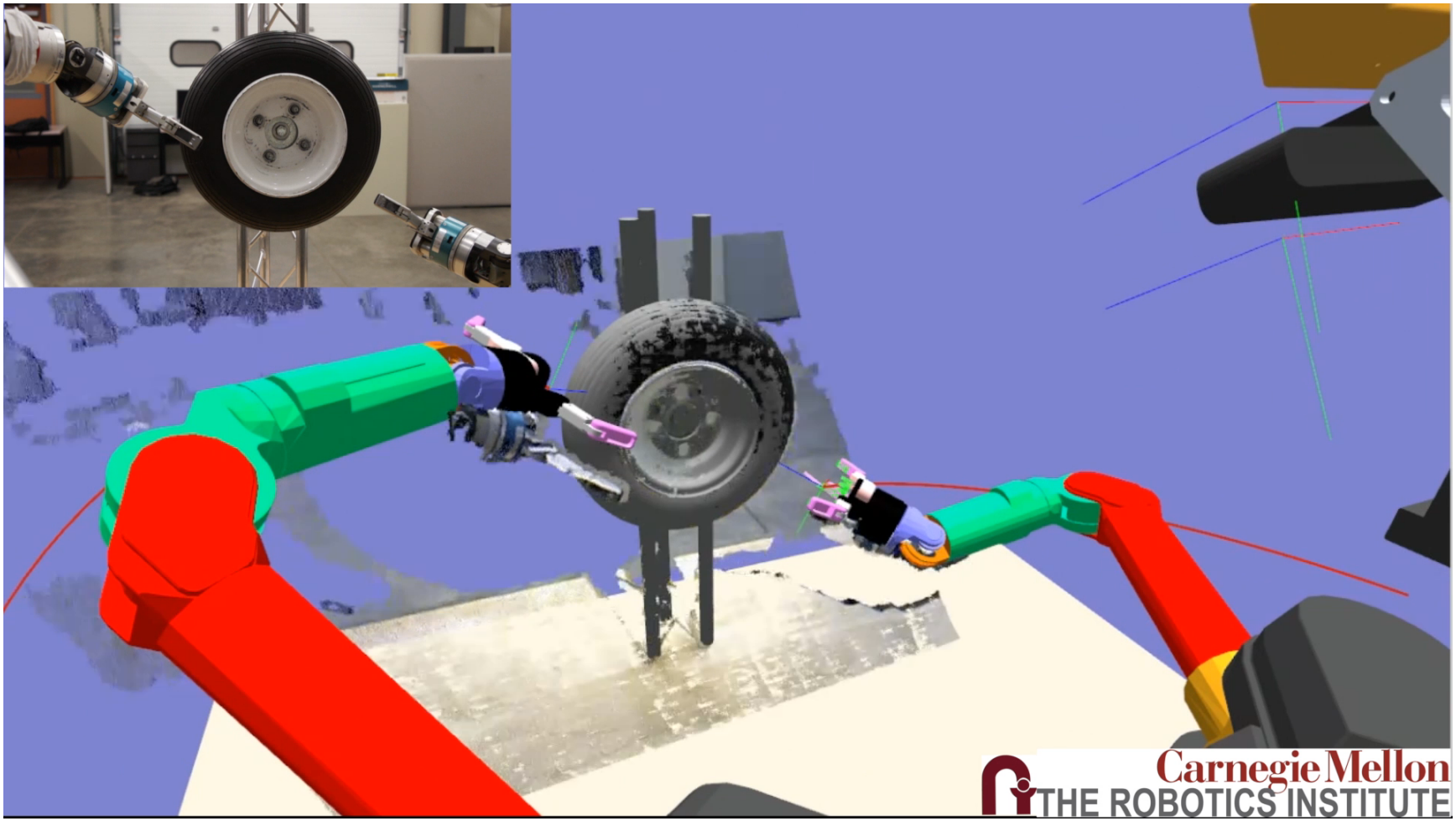


- EKF with articulated ICP over manipulator joint angles, camera pose and pose of (partial) object

Courtesy of Drew Bagnell

Closed-Loop Servoing

[Klingensmith-etal: ICRA-13]



[Krainin-F-Curless: IJSRR-11, ICRA-11]

Active Object Modeling

Object Modeling and Manipulation

- Depth information extremely useful for grasping, modeling, and manipulating objects
- Enables robust exploration of objects and DOFs
- So far, reasonably simple control, not a lot of sophisticated physics-based reasoning
- Touch sensors /skins big step forward

Outline

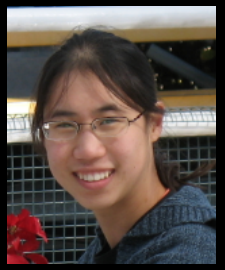
- Environments
- Objects
- People
- Discussion

[Ziola-Grampurohit-Landes-Fogarty-Harrison: VL-HCC-11]

Interactive Task Assistance and Playing: Lego OASIS

[Lei-Ren-F: UbiComp-12]

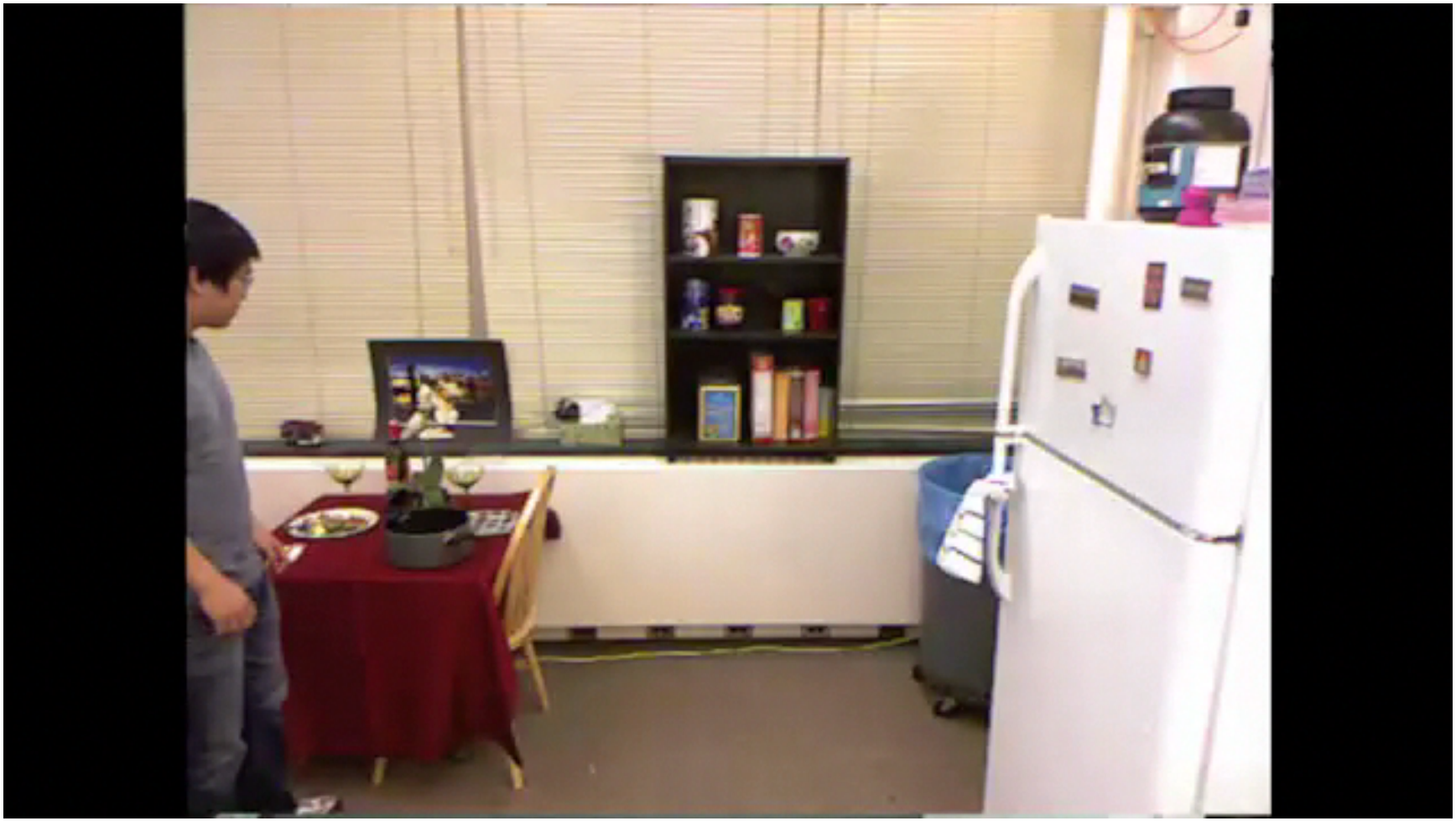
Tracking Cooking Activities



Courtesy of Ashutosh Saxena

Anticipating Human Activities

[Koppula-Saxena: RSS-13]



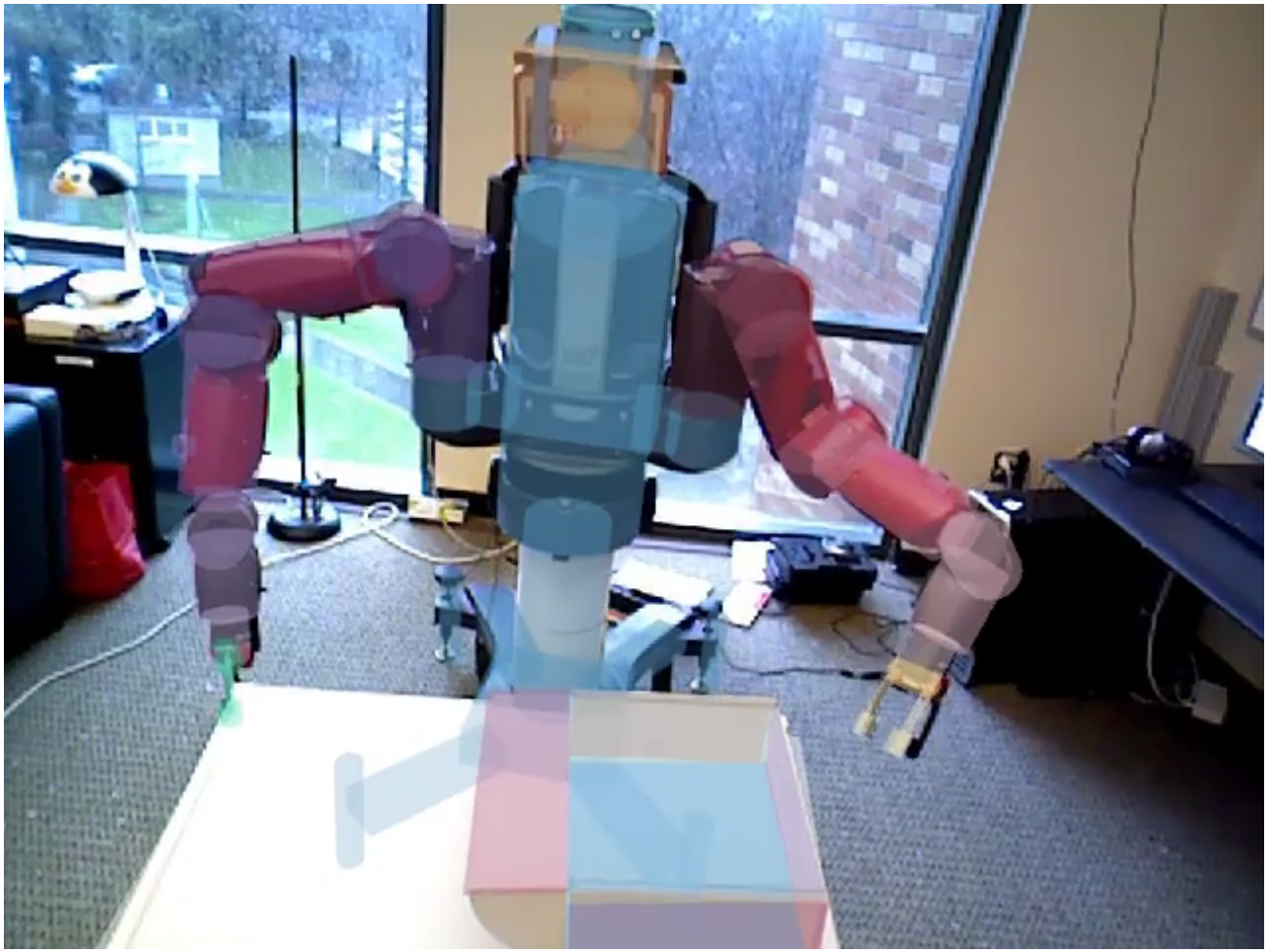
[Schmidt-Newcombe-F: RSS-14]

DART: Dense Articulated Real-Time Tracking

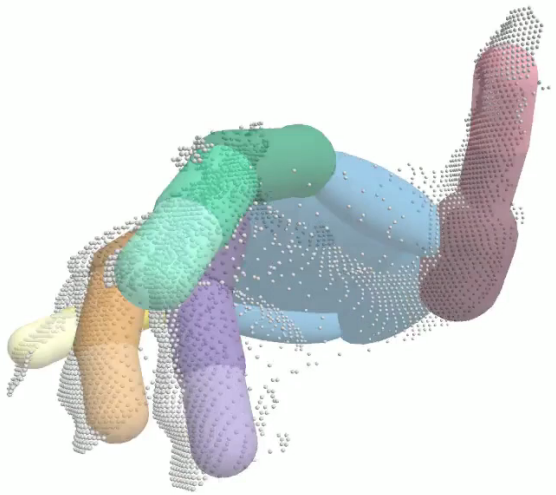
Using Articulated Signed-Distance Functions


- Goal: General tool for real-time tracking of arbitrary articulated objects
- Input: Shape models of parts along with joint structure
- Insight: Efficient optimization via articulated signed distance functions

Tracking Baxter (20 DoF) and Box (8 DoF)



Hand (27 DoF) and Human Body (42 DoF)



People

- Working with people is extremely important for robotics
- Body and object tracking provide ideal context for activity recognition and anticipation
- Still just at the beginning
  - Fine-grained activity recognition (what and how)
  - Complex, multi-step and concurrent activities
  - Joint task solving and learning

Outline

- Environments
- Objects
- People
- Discussion

Conclusions

- Depth cameras provide shortcuts for low-level vision but do not readily solve any higher-level problem

- Enable non-experts such as robotics, HRI, or HCI researchers to build on reasonably robust vision and 3D information

- Robotics is about interacting with the world
  - Depth provides immediate access to "where things are"
  - Navigating, grasping, planning, interacting