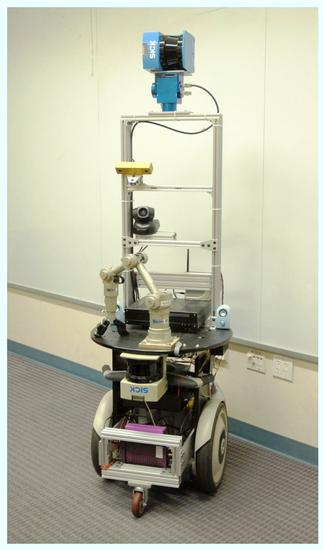
Holistic Scene Understanding from Visual and Range Data

Stephen Gould, Paul Baumstarck, Morgan Quigley, Andrew Y. Ng, Daphne Koller

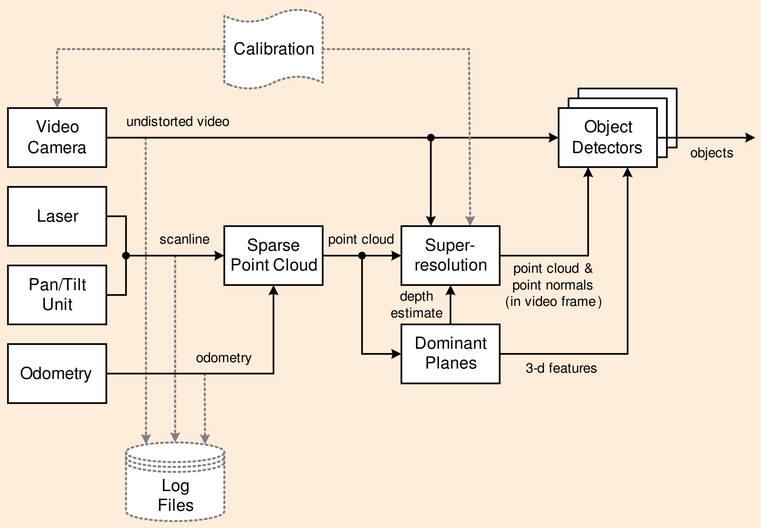
1 Overview

- Robust object recognition is a long standing goal of computer vision.
- This is especially difficult in a robotic setting where the objects of interest are small (e.g. 32 x 32 pixels) and ill framed.
- 3-d contextual information can significantly improve object recognition accuracy.



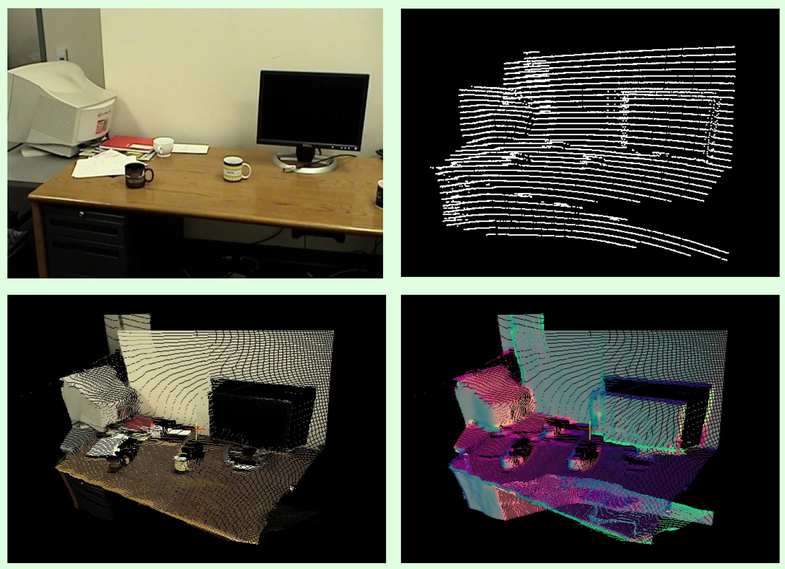
- Our work is part of the ongoing **STAIR** (STanford Artificial Intelligence Robot) project which has the long-term goal of integrating techniques from all areas of AI to build a useful home/office assistant robot.

2 Multi-sensor robotic vision system



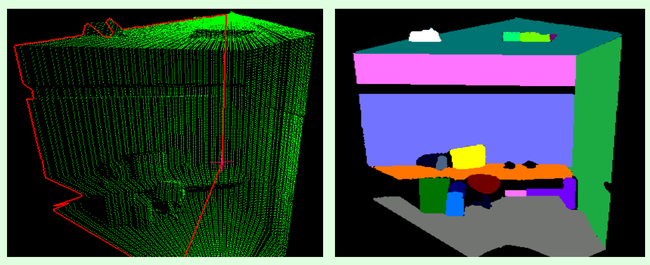
- Laser scan lines are combined with odometry and pan information to form a sparse point cloud.
- Dominant planes are extracted from sparse point cloud using region-growing algorithm.
- Super-resolution MRF infers 3-d location (dense point cloud) and direction of normal vector for every pixel in the image.
- Object detectors combine 2-d and 3-d cues for better object recognition.
- Architecture supports efficient multicore processing.
- "Anytime" implementation of core algorithms allows us to meet hard real-time constraints for robotic systems.

3 Super-resolution sensor fusion

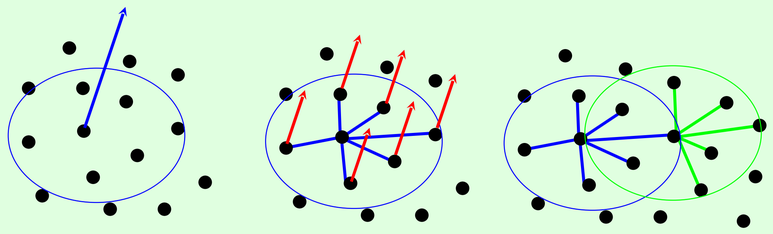


- **Super-resolution MRF** [Diebel, 2006] used to infer depth value for every image pixel:
  - singleton potential encodes our preference for matching laser measurements;
  - pairwise potential encodes our preference for planar surfaces.
- Algorithm can be stopped at anytime and later resumed from previous iteration.
- Use camera intrinsic parameters to reconstruct dense point cloud, and SVD over local neighborhood to estimate surface normal vectors.

4 Dominant plane extraction



- Plane clustering based on greedy, region-growing algorithm with smoothness constraint [Rabbani, 2006].
- Extracted planes used to initialize super-resolution MRF and for 3-d feature calculation.



(a) Compute normal vectors using neighborhood

(b) Grow region over neighbors with similar normal vectors

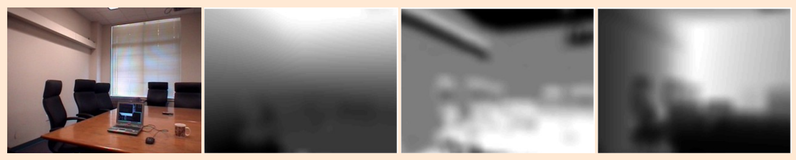
(c) Use neighbors with low residual to expand region

5 2-d object detectors

- We use a sliding-window object detector to compute object probabilities given image data only:
  - Features are based on localized **patch responses** from pre-trained dictionary and applied to image at multiple scales [Torralba, 2007].
  - **Gentle-boost** [Friedman, 1998] classifier applied to each window.

6 3d features

- Compute 3-d features over candidate windows (in image plane) by projecting window into 3-d scene.
- Features include statistics on **height** above ground, **support** (vertical or horizontal), and **size** of object.

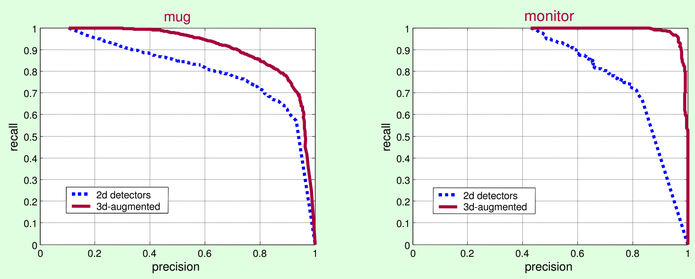


- Other features: **surface variation** (computed as variance over normal vector direction) and **distance to dominant planes**.

7 3d-augmented object detectors

- Use features to reduce computation required by patch-based 2-d detectors, i.e. **attention** mechanism.
  - For example, flat/planar regions are unlikely to be objects and can be discounted before performing expensive patch response calculations.
- Augment feature vector with **log-odds** ratio from 2-d sliding-window classifier.
  - Advantage of this approach is that 2-d classifier can be trained separately.
- Learned logistic model gives probability of object given image and laser data for given location and scale.

8 Experimental results



References

[1] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS* 2006.

[2] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," Stanford University Tech. Rep., 1998.

[3] T. Rabbani, F. A. van den Heuvel, and G. Vosselman, "Segmentation of Point Clouds Using Smoothness Constraint," in *ISPRS*, 2006.

[4] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," in *PAMI*, 2007.