Enhanced Visual Scene Understanding through Human-Robot Dialog

Matthew Johnson-Roberson, Jeannette Bohg, Danica Kragic
Computational Vision and Active Perception Lab
Centre for Autonomous Systems
KTH, Stockholm, Sweden
mattjr,bohg,danik@csc.kth.se

Gabriel Skantze, Joakim Gustafson, Rolf Carlson
Dept. of Speech, Music and Hearing
KTH, Stockholm, Sweden
gabriel,jocke,rolf@speech.kth.se

Introduction

Current robots are capable of autonomously completing many tasks that are challenging both in perception and manipulation.However, autonomous behavior is still only possible under many assumptions and within a controlled environment. One of the key challenges in robotics is to relax previously made assumptions and thereby enable a robot to act in new situations and handle increased uncertainty.

In this paper, we are specifically dealing with the problem of scene understanding in which the robot has to correctly enumerate how many separate objects there are in the scene and to describe them in terms of their attributes. Only after a full understanding of the scene has been reached, models of new objects can be extracted, labeled and stored in working memory for later re-recognition. Furthermore, it has also been shown that segmentation eases tasks like visual recognition and classification processes by reducing the search space, (Björkman and Eklundh 2005). Thus, a valid scene model forms the basis for a general symbol grounding problem (Harnad 1990).

An example for an embodied robotic system that tries to segment a scene into several objects with an active vision system is proposed in our previous work (Johnson-Roberson et al. 2010). Dependent on how close objects are to each other or how similar they are in appearance, some segments might incorrectly group two or more objects together. For a robot to be able to accomplish a certain task based on such a scene model, it has to (i) detect uncertainties in that model and (ii) confirm or improve the uncertain object hypotheses to achieve a sufficiently good understanding of the scene.

The approach taken up in this paper is to put a 'human in the loop'. We propose a novel human-robot-interaction framework which combines state-of-the-art computer vision and a natural dialog system. Thus, a human can rapidly refine the model of a complex 3D scene. This process is visualised in Figure 1.

Contributions and System Overview

The system is comprised of three large components as shown in Figure 2: First, the vision system uses stereo cameras to produce a point cloud of the scene. This point cloud is then clustered by performing an initial segmentation grouping points with similar traits. Second, the scene analysis module determines areas of the scene that are the poorest object hypotheses and seek human arbitration. And finally, the dialog system allows a human operator to provide responses to the robot's questions in a natural manner. Based on this, the scene model can be refined by re-seeding the initial scene segmentation.

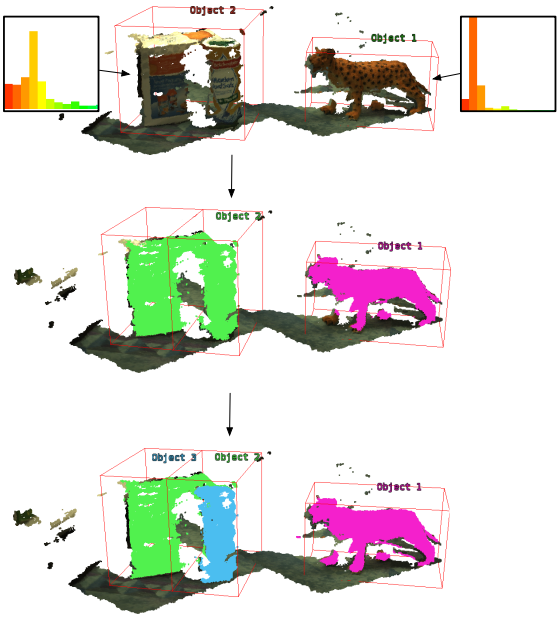

Figure 1: Depiction of the different stages of the scene understanding. Top: Point cloud from stereo matching with bounding boxes (BBs) around segmented objects and their hue histograms. Left segment containing two objects has a higher *hue entropy* (0.66 as opposed to 0.24) and is therefore considered more uncertain. Middle: Initial labeling of the segments. Left segment is re-seeded by splitting BB in two parts based on human dialog input. Bottom: Re-segmented objects. Three objects are correctly detected.

Vision System The initial segmentation is performed using saliency points as labels in a Markov Random Field

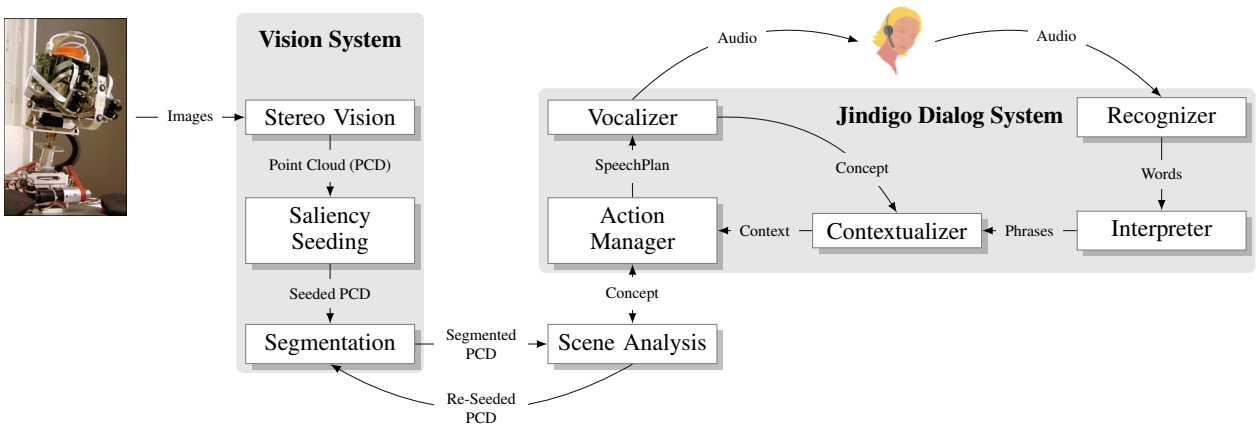

Figure 2: Overview of the System. Left: Armar III Active Stereo Head. Middle: Vision System that obtains stereo images from the cameras and outputs a segmented point cloud. Right: Architecture of the Jindigo Dialog System. Vision and Dialog System are communicating through a Scene Analysis Module.

(MRF) graphical model framework. This paradigm allows for the identification of multiple object hypotheses simultaneously and is described in full detail in (Johnson-Roberson et al. 2010). In brief, for each initial cluster generated from saliency points, a Gaussian Mixture Models (GMMs) is utilized to model the color properties. The unary cost of each node in the MRF is the likelihood of membership to an object hypothesis' color distribution. The pairwise cost is derived from a kd-tree neighborhood search directly on the point cloud and enforces smoothness between adjacent labels. Segmentation is finally performed by multi-label energy minimization.

Dialog System The dialog component is based on Jindigo, a Java-based open source dialogue framework (Skantze 2010) and is responsible for the spoken interaction with the user, as shown in Figure 2. When the Scene Analysis component (SA) needs input from the user, it sends a request to the Action Manager (AM) in the Dialog System. The AM then checks that this request can be performed and situates it in the dialog context. As the dialog system interprets the user input into scene refinements, the AM sends these updates to the SA. As soon as the scene analysis has been refined, the SA may send a new request. After each refinement, the SA also sends the current set of objects and their properties to the AM, so that the AM can resolve expressions like the largest segment.

Refining the Scene Model through Human-Robot Interaction The Scene Analysis module fulfills two tasks. First, it detects object hypotheses that are most likely to be incorrect. This is based on computing the entropy of the color hue histogram and 3D point histogram for each segment. The entropy for these two distributions should to be lower for segments containing only one object than for segments merging several objects. This is based on the intuition that objects are relatively homogeneous in their attributes. A human operator is queried for the most uncertain object hypotheses first to obtain information about the number of objects in that segment and their relative position to one another.

The second task of this scene analysis module is to make use of the input from the human operator. As exemplified in Figure 1, the bounding box around the object hypothesis is divided dependent on the correct number of objects and their positioning. Based on this, the initially segmented points are relabeled and new GMMs for the region are iteratively calculated. Then a new graph is constructed based upon the membership probability of each point to the new models. Energy minimization is performed for the new regions and the process is repeated in the region until convergence.

Results and Conclusions

We tested our approach on 20 specifically challenging scenes with in total 67 objects with similar appearance and positioned close to each other. 33 of these were correctly segmented while 17 segments each merged two objects together. Using the entropy measure to spot incorrect segmentation resulted in 39% fewer unnecessary queries compared to a random selection. For evaluating segmentation performance, all scenes were manually labeled to be used as ground truth. On average we achieved an increase of 10% in performance when comparing the scene segmentation before and after user interaction. These results show that the proposed framework allows for a rapid enhanced visual scene understanding through human-robot dialog.

References

Björkman, M., and Eklundh, J.-O. 2005. Foveated Figure-Ground Segmentation and Its Role in Recognition. *Proc. of British Machine Vision Conference*.

Harnad, S. 1990. The symbol grounding problem. In *PhysicaD: Nonlinear phenomena*, volume 42, 335–346.

Johnson-Roberson, M.; Bohg, J.; Björkman, M.; and Kragic, D. 2010. Attention Based Active 3D Point Cloud Segmentation. In *IROS 2010*. accepted.

Skantze, G. 2010. Jindigo: A Javabased Framework for Incremental Dialogue Systems. In *Proceedings of Interspeech*. submitted,`www.jidingo.net`.