Point Clouds and Range Images for Intent Recognition and Human-Robot Interaction

Richard Kelley, Amol Ambardekar, Liesl Wigand, Monica Nicolescu, Mircea Nicolescu
Department of Computer Science and Engineering
University of Nevada, Reno
Reno, Nevada 89557
{rkelley, ambardek, wigand, monica, mircea}@cse.unr.edu

***Abstract*—The wide availability of inexpensive RGB-D cameras such as Microsoft's Kinect presents an opportunity to significantly improve the sensing capabilities required for human-robot interaction. In this paper we present an intent recognition system that demonstrates the potential of RGB-D cameras for systems that must interpret human activities. We also discuss systems we are currently developing and touch on techniques we are using to make such systems scalable to significantly larger data sets.**

I. Introduction

To engage in effective interaction with humans, robots must have robust sensing capabilities: they must be able to detect, track, and ultimately understand human objects and activities. Successful interaction requires not only that these tasks be performed reliably, but that they be performed quickly as well. Although roboticists have made significant progress tackling these challenges using traditional video cameras, the widespread availability of inexpensive cameras that capture depth information presents an opportunity to significantly advance the state of the art for systems that must interpret human activities. In this paper, we outline our current work developing systems that use RGB-D cameras to improve human-robot interaction. We begin by reviewing related work. We then provide brief descriptions of the systems we have developed and discuss preliminary results. We conclude by describing ongoing work to improve and extend these systems using parallel (message passing and GPU) approaches.

II. Related Work

In traditional camera systems, depth information has to be inferred through stereo algorithms [6]. Creating dense depth maps from traditional stereo rigs is challenging and computationally expensive. More recently, projected texture stereo has been used to improve the performance of traditional stereo [2]. Along these lines, Microsoft's Kinect provides a low cost system that provides dense 640x480 depth maps at 30hz.

To process the output of these systems, there are essentially two options: range images and point clouds. Many of the techniques of classical computer vision are applicable to range images [6]. In contrast, point clouds require different techniques [3]. Although significant progress has been made in point cloud processing, researchers are only beginning to closely examine how such advances may be applied to human-robot interaction.

The problem of recognizing intentions in humans requires that a system move beyond observable stimuli to infer unobserved goals and desires of an agent [1]. To date, the vast majority of intent recognition systems rely at least to some extent on vision [5]. Although this approach makes sense in light of the importance of vision in human-human interaction, inferring information about the 3D world from 2D images adds a layer of complexity to the problem that RGB-D cameras can easily avoid.

III. Intent Recognition with Point Clouds

In the intent recognition problem, a robot observes one or more agents interacting with each other and the environment, and must infer the goals and predict the actions of each agent. Intent recognition can be approached through multiple sensing modalities, though in most systems vision plays a significant role. Previous systems have used video processing techniques such as appearance modeling and foreground background segmentation to track objects and people [5]. Our approach builds on these ideas, but uses point cloud processing techniques (provided by PCL) to handle segmentation and tracking.

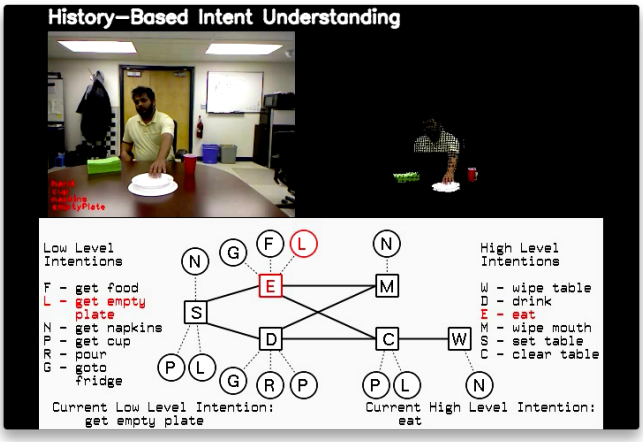

Fig. 1. Screenshot showing the use of RGB and point cloud information to estimate the intentions of a human interacting with household objects over an extended period of time.

We assume that the observer robot is stationary, and observes a human interacting with various objects in a household or office setting over time (see Fig.1). As it operates, the system takes as input a point cloud from a Kinect and performs the following steps:

1) **Foreground-Background Segmentation.** The system begins by segmenting out uninteresting regions from the cloud. In our office scene, this includes the floor, walls, and the table on which the objects of interest lie. This segmentation is performed using standard tools in the Point Cloud Library. The output of this stage is a set of clouds corresponding to the objects of interest in the scene.
2) **Appearance-based Object Recognition.** Offline, we train a Gaussian mixture model for each object we wish to recognize. At run time we use these models to perform classification for each segmented cloud. The output of this stage is a set of object labels and positions (centroids) computed from the classifier and the cloud information in step 1. This includes the locations of the human's head and hands.
3) **Interaction Modeling.** Once the position of each object is known, the system tracks the human's motion across consecutive frames to determine if he or she is reaching for anything. If a reaching action is detected, that information is sent to the intent recognition system for further analysis.
4) **History-based Intent Recognition.** Lastly, when the low-level classifier detects that a human is reaching for an object, the intent recognition module attempts to infer the intention underlying that activity. To perform this task, the system maintains a history of inferred intentions for each agent in the scene. Using a simple temporal logic knowledge base to represent activities, the system interprets the history to produce an estimate of the higher level intention of the observed agent.

IV. Ongoing and Future Work

In this section, we highlight three extensions that we are currently developing. The first extends our work on intent recognition by using more detailed human pose recognition as an input to the intent recognition system. The second system registers views from multiple cameras to monitor the objects and actions in a larger space. Lastly, we are working on parallel algorithms to support our other projects, both for message-passing and GPU-based systems.

A. Full-Body Intent Recognition

The intent recognition system described above ignores most of the geometry of the human's interaction with the scene. Although the task of labeling the parts of a human is challenging, recent work using the Kinect achieves good results in real time [4]. We are currently implementing a system based on this work, along with a system that uses detailed descriptions of human pose as an input to an intent recognition system. We expect that by using more information about a human's geometry, robots will be able to produce significantly better estimates of human behavior and intentions.

B. RGB-D For Ambient Intelligence

In addition to single-camera intent recognition, we are also developing a system that integrates multiple range images to perform tracking and behavior recognition across larger spaces. Currently, the system uses frame differencing in each of multiple cameras to track a human moving across their joint field of view. We plan on extending the system to create a smart environment spanning several offices and labs.

C. Parallel Algorithms

To efficiently process the volume of data required by several kinects, we are currently developing systems that process range images and point clouds in parallel. The system combines message passing parallelism to handle coarse-grained parallel tasks (processing video sequences from multiple kinects) and GPU algorithms to handle data-parallel tasks.

V. Conclusion

In this paper we have shown a number of ways in which human-robot interaction can benefit from the use of RGB-D cameras. By exploiting the wealth of information that such cameras provide, robots can learn to effectively and efficiently interact with humans.

References

[1] Y. Demiris. Prediction of intent in robotics and multi-agent systems. In *Cognitive Processing*, pages 151–158, 2007.

[2] Kurt Konolige. Projected texture stereo. In *ICRA*, 2010.

[3] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *International Conference on Robotics and Automation*, Shanghai, China, June 2011.

[4] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] Alireza Tavakkoli, Richard Kelley, Christopher King, Mircea Nicolescu, Monica Nicolescu, and George Bebis. A visual tracking framework for intent recognition in videos. In *Proceedings of the 4th International Symposium on Advances in Visual Computing*, ISVC '08, pages 450–459, Berlin, Heidelberg, 2008. Springer-Verlag.

[6] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.