

# Scene Understanding through Audio-Visual Fusion

E. Martinson, B. Fransen, E. Lawson

**Abstract**— Scene understanding involves the integration of a wide variety of information to produce a thorough description of the robot's environment. By integrating spatial, visual and audio cues, we could provide a greater amount of understanding than can be obtained using one of the modalities alone. In this paper, we describe our current work on using audition to enhance existing object detection and tracking systems. By combining these systems together, we can recognize known objects and provide additional information about unknown objects.

**Keywords**- Scene Understanding, Audition, Object Recognition

## I. INTRODUCTION:

Scene understanding is a large problem with many different aspects. Object recognition, person recognition, and active tracking are three sub-problems of scene understanding that we have been investigating as part of our human-robot interaction framework. In addition to studying visual and spatial properties of the scene, we are now beginning to investigate visually guided acoustic inspections of objects in the scene. Knowing that an object or area is of potential interest, an acoustic investigation of reflectance and/or sound transmission properties can add to scene understanding. The fusion of such acoustic knowledge, which suggests material and/or structural composition, with visual information (e.g. size and shape) could be used to infer object function.

Robot audition, which studies sound characteristics of a robot's environment, requires a sensor array of microphones similar to the array of pixels in a camera. The microphone array is used to study both the output and absorption properties of scene elements which is integrated with our visual information for analysis. We will begin with describing our approach to scene understanding through RGB-D cameras and then review how, through spatial information, we are investigating and integrating audition into our world model.

## II. RGB-D SYSTEM

We provide a visual inspection of the scene using object recognition, person recognition, and active tracking to enable scene understanding and human-robot interaction. Our current line of investigation is based on both the low level recovery of information and methodologies necessary to integrate that information for the purpose of developing scene understanding. We have implemented systems that analyze the world through RGB-D input. Our current hardware configuration is comprised of a Swiss Ranger (SR) 3000 integrated with a color camera (Figure 1). The two cameras are mechanically calibrated such that their centers of projection are parallel.

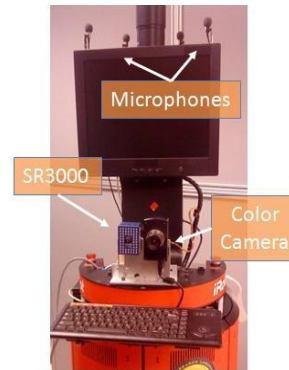


Figure 1: Robot with sensors to provide spatial, visual and audio information.

The translation and projection matrix is solved by using a calibrated test pattern. While the SR 3000 is used on all our systems, we use a Lumenera Lw295C and Point Grey fireflyMV for experiments. The integrated system provides both depth and color.

We use the rigid object detection and tracking functionality described in Fransen et al. [3] to generate object and/or scene models from integrated range and color imagery (Figure 2). The system extracts local spatial and intensity gradients, integrated over a region, to measure pose and position change for an object. For the purpose of tracking rigid objects, spatial information is only necessary during training. At run time, an object's position can be ascertained solely using appearance changes. As a tracker, the algorithm continually updates the known position and pose information of an object over time. As a recognition system, the same algorithm can be applied in a scanning fashion to detect an object in the scene.

For articulated person tracking, a learned skin color model is utilized in conjunction with depth to recover arm and face positions [2]. To model skin tone, a detected face region is sampled to extract a person's unique hue. A hue model is then used in conjunction with a body pose model to extract hand, elbow, and face positions for a person moving in front of the camera. Figure 3 demonstrates face and pose tracking.



Figure 2: Integrated depth and color system.

This work was supported by the Office of Naval Research under job order numbers N0001408WX30007 and N0001410WX20773. The views and conclusions contained in this document should not be interpreted as necessarily representing official policies, either expressed or implied, of the U.S. Navy.



Figure 2: Left: Rigid object tracking applied to faces; Right: Body pose tracking.

The combined set of visual capabilities available to our robotic systems, including object modeling, scene reconstruction, and person tracking provide a solid basis for interaction with people, not just because they themselves convey substantial information about the scene, but also because their availability facilitates understanding of auditory phenomena that are otherwise impossible to decipher.

### III. AUDITORY INVESTIGATION

Auditory perception is commonly used for sound localization and speech understanding [2], but audition potentially consists of substantially more information, some of which is not available visually. In particular, objects in the environment affect the flow of sound around them, both relative to their size and shape, which are determinable by depth images, and by their material composition and internal structure. In general, reverberations from soft objects are of lower amplitude than those off of hard objects, and sound waves are more likely to travel through light objects than heavy ones. If reasonable fidelity environmental models could be provided visually, then acoustic knowledge could assist both in identifying objects and identifying their purpose in the environment. Acquiring this level of visual integration with such guided auditory investigations is a central theme of this combined effort.

Our current line of investigation in this work is identifying surface material properties, specifically focusing on a material's reflectivity, or the ratio of reflected amplitude to incident amplitude [5]. Given an unknown surface that a robot can get close to, the simplest method for identifying reflectivity is an active process, where the robot moves close to the unknown material surface, generates an impulse sound via an on board speaker, and listens for the response. The coefficient, however, depends on both the frequency and incident angle. Most published data deals with random incidence angle reflection, which tends to be lower than normal for low frequencies, and about the same for higher frequencies. Therefore, by coupling the measured impulse response with surface angle estimations from the RGB-D system, we can compare new surfaces with a reference material (e.g. concrete).

A more involved approach for large or unreachable objects, is to compare measured auditory feedback with an acoustic simulation of the environment. In the vicinity of highly absorbent materials, sound volume should be significantly depressed. Figure 3 demonstrates this reduction near curtains and a couch in an otherwise reverberant environment using a ray-tracing based simulation [1]. This same model verification

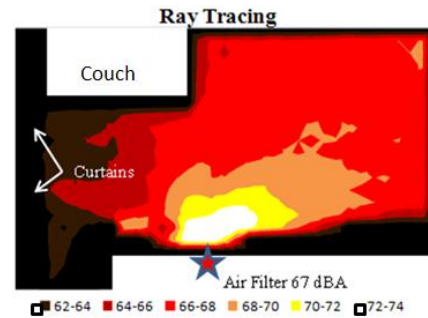


Figure 4: Reverberation of sound in a simulated environment

based-approach could be used to estimate transmission through unknown objects as well, which would have a similarly large impact sound propagation.

What makes this simulation feasible, however, is having enough knowledge. An effective acoustic simulation requires knowledge of sound sources, which can be acquired autonomously by our mobile robot [4], and environmental knowledge such as a 3D obstacle map and surface materials, which are provided by visual inspection. The obstacle map can be acquired with depth cameras, augmented with available a priori knowledge, and the material properties can be determined through a combination of audio and video.

In preliminary work, we incorporated the models created by the RGB-D system into the acoustic simulator for purposes of estimating the volume of an unknown sound source. We assumed known wall positions but unknown furniture configurations, and that the environment is highly and uniformly reflective. The latter assumption is true for most office environments. In 10 trials with different sound sources ranging from 55-72dBA, using the robot generated depth map resulted in a mean estimated volume error of 1.9dBA.

### IV. SUMMARY

In this paper, we described the capabilities of our existing robotic visual systems, focusing on scene understanding through object and/or people modeling, tracking and recognition. We then presented new work fusing this visual effort with auditory information for purposes of material identification. Together, visual, spatial and auditory information provide a more complete description of environment, facilitating participation in human-robot interaction.

### V. REFERENCES

- [1] Elorza, D.O., *Room Acoustics Modeling Using the Raytracing Method: Implementation and Evaluation*, Licentiate Thesis, Dept of Physics, University of Turku, 2005.
- [2] Fransen, B.R., Morariu, V.I., Martinson, E., Blisard, S., Marge M., Thomas, S., Schultz, A.C., Perzanowski, D., "Using Vision, Acoustics, and Natural Language for Disambiguation", *HRI 2007*: 73-80, 2007.
- [3] Fransen, B.R., Herbst, E.V., Harrison, A.M., Adams, W., Trafton, J.G., "Real-time Face and Object Tracking", *IROS 2009*: 2483-2488, 2009.
- [4] E. Martinson and A. Schultz, Discovery of sound sources by an autonomous mobile robot, *Autonomous Robots* 27 (2009)
- [5] Raichel, D.R., *The Science and Applications of Acoustics*, New York, NY: Springer-Verlag, 2000.