A Robot Team that Can Search, Rescue, and Serve Cookies: Experiments in Multi-modal Person Identification and Multi-robot Sound Localization

D. Blank, G. Beavers,
W. Arensman, C. Caloianu, T. Fujiwara, S. McCaul, C. Shaw
Department of Computer Science and Computer Engineering
University of Arkansas, Fayetteville
{dblank,gordonb}@uark.edu

Abstract

A team of three robots placed second in both the *Urban Search and Rescue* and *Hors d'Oeuvres Anyone?* events at the 2000 American Association of Artificial Intelligence autonomous robot competitions. This paper describes the multi-robot, low-cost sound localization technique, and the multi-sensor, person recognition system used in the AAAI contests by the robot team.

Introduction

Each year in association with its yearly conference, the American Association of Artificial Intelligence (AAAI) hosts a series of competitions designed to challenge, and highlight, autonomous mobile robotics research from around the world. This past summer, we took our team from the University of Arkansas to compete at AAAI-2000 in Austin, Texas.

At the conference, three robot contests were held: a repeat of the previous years' *Hors d'Oeuvres Anyone?* competition, where robots were to serve food to conference attendees; a new event called *Urban Search and Rescue* (USAR); and a long-term, very difficult problem called the *Robot Challenge*.

All of the contests allow teams from colleges, universities, and other labs to show off their best attempts at solving common robotics tasks in a competitive environment. Teams compete for place awards as well as for technical innovation awards, which reward particularly interesting solutions to problems.

The Robot Challenge is so-called for good reason: the goal of the challenge is to create a robot capable of attending a conference on artificial intelligence, including finding its way to the registration booth, registering, and even hobnobbing with the other attendees. In addition, the robot has to present a paper, complete with a question-and-answer period! This is meant to be a decade-long challenge. Although we did not compete in this event, a few teams have begun to attempt the challenge.

Although all of the problems are open ended, we set two specific goals for our team: locate people in the search and rescue task using sound, and attempt to identify people in the serving contest. All of the solutions we present here were written in our open source Extendible Robot Control Language (XRCL), and are available off of our website at `http://ai.uark.edu/` [1].

Urban Search and Rescue: Sound Localization

The objective of the USAR contest is to give participants the opportunity to work in a domain of practical importance. Robots had to enter a fallen structure, find simulated human victims, and direct human rescuers to them. Victims (represented by manikins) could be identified by their body heat, motion, sound, or skin color.

This year marked the first year for the USAR event. The National Institute of Standards and Technology (NIST) designed and built a USAR structure in which the victims were to be located and rescued. The impressive structure contains areas of easy, medium, and hard degrees of difficulty for autonomous mobile robots to move about. The "easy" area of the USAR course was still a challenge for many of the robots, because it contains glass walls (hard to detect for laser), curtains (hard to detect for sonar), and other objects that fell below the line of the robot's sensors. However, the medium and hard areas of difficulty were designed to give robotics researchers something to attempt for the next few years (see Figure 1.) For example, the hard area contained a ramp and large holes that robots were to avoid lest they come crashing down to the area below. Body heat was simulated with heating pads, and NIST had rigged mechanical devices to provide motion and sound.

Our self-imposed goal in the USAR contest was to identify the location of noises made in the arena. One technique that has been used to estimate the location of the source of a sound is to mount three microphones on a robot at the vertices of an equilateral triangle of about 0.2 meters on a side. A sound wave will thus arrive at the different microphones at times differing by as much as one half millisecond. Given an accurate measure of the difference of arrival times, the sound source can be computed to lie on a branch of a hyperbola having one of the microphones at its focus. Each pair of microphones thus determines a hyperbola and the intersection of multiple hyperbolas is the source. This methodology requires specialized multichannel A/D hardware. For example, a system must be capable of receiving input from multiple sound sources. In addition, the device must be capable

Figure 1: NIST's USAR arena. This picture shows one of the "hard" areas. Note the debris and ramps that make it difficult for autonomous robots.

of sub-millisecond resolution. Due to these limitations we decided against using this methodology.

Instead, we proposed a technique for using commodity off-the-shelf (COTS) low-cost hardware. Specifically, we wanted to attempt to perform sound localization using standard PC sound cards and microphones. To accomplish this, we decided to put one microphone on each of three independent robots. Only one sound card was needed per robot and the robots could be moved far enough from one another so that millisecond accuracy would give a reasonably accurate estimation of the distance, e.g. a difference of 10 milliseconds would indicate a difference in distance from the source of about 3.5 meters.

This method, however, introduced other difficulties. If all of the sound sources were located on a single computer, synchronizing timing would be relatively easy. However, separating the microphones across three robots made global time synchronization an obstacle. The three robots that we had available can be connected via a wireless network, and there exist tools to sync such networked computers. Given that we had a global time common to all robots, the question still remained of how to mark a "sound event" with that time. Our idea was to tag significant sounds at their onsets with a global time. However, this turned out to be non-trivial. Each sound device has an associated series of hardware and software buffers that accumulate input before sending it on to the robot's operating system. It was therefore impossible to accurately associate the global synchronized time with the sound event without resorting to writing our own sound device driver.

Our solution to this problem was to treat the sound input as time. Rather than attempt to associate the buffered sound input with a separate time system (i.e., the system clock), we realized that sound was coming into the computer at a regular rate and could provide its own timing information. Knowing the sound card's sampling rate, our goal, then, was to simply count the digitized sound units as they came in. The number of samples divided by the sampling rate could then be used to calculate the time of onset of a sound event. Except for a small amount of "drift" that we corrected in the timing of the digitized sound, the technique of using the sound signal as time did indeed work. This methodology left one final problem: how to synchronize the three sound cards *together*. It was decided that a sound emitted at a position equidistant from the three robots could be used as an initialization so that times of arrival would be measured from the time of this common signal.

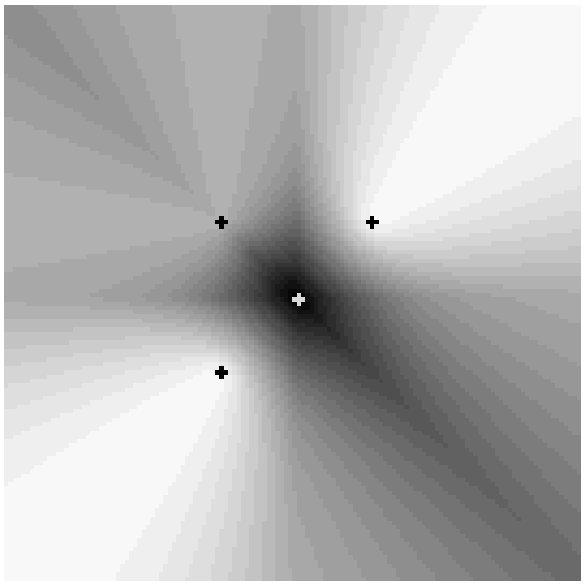

Figure 2: This image represents the differences between actual and computed sound timing information received from three robots (black crosses). Smaller differences are shown as darker areas. The white cross shows the actual location of the sound source. The area shown is 13.5 meters square sampled every 0.1 meters.

Having solved the problem of getting synchronized timing differences from the robots for a particular sound event, the location of the sound must then be computed. For this to work, the robots must, of course, know their approximate global location. Theoretically, the location of the sound can be computed exactly from the intersection of two hyperbolas, as described above. In practice, however, inaccuracies make such exact equation-solving methods impractical.

Our method used in the competition to locate the source of the sound was a simulation technique. Once we had the actual sound time differences from each of the robots, we then computed hypothetical differences from a grid of positions surrounding the robots. We then compared the actual timing data with the computed timing data at each of these sampled points, and computed the difference. Figure 2 shows these differences as gray scales. The black crosses indicate the positions of the robots. The gray scale values represent the differences between the computed timing data and the actual data with white representing large differences and black smaller differences. The white cross represents the location of the actual sound source. One could think of this as a likelihood map of the possible location of the sound.

Our sound localization method worked quite well in the

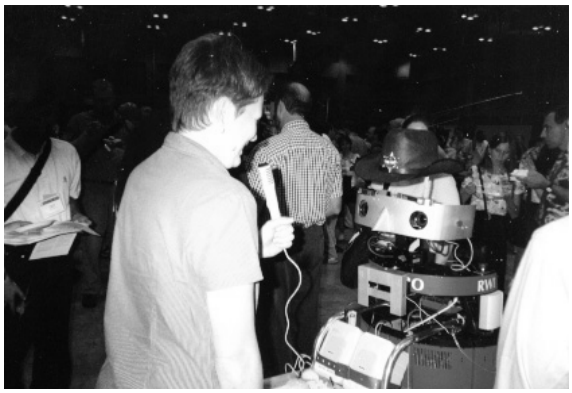

Figure 3: Elektro dishes out cookies and conversation while attempting to re-recognize conference attendees using our multi-modal color histogram technique.

lab. When a sound event occurs near one of the robots or when the sound event is equidistant from all three robots, the system can reliably locate the source of the sound. However, the amount of error grows as the source of the sound differs from these two standard cases. Further experiments will need to be performed to determine why the method did not work in all situations. Some enhancements that may address this issue are discussed below.

The relative amplitude of the sound wave at the three microphones could be used to improve our calculations in two ways. First, since the amplitude will drop off as the square of the distance from the source, the amplitude can be used as another means to calculate the position of the source. Second, the amplitude could be used to weight each robot's timing data differently. For example, knowing that a sound is closer to one robot than the others (by examining the amplitude of the sound event, for example) we could give it more confidence, and thus a larger weight. Variable weighting based on amplitude for each robot's sound data also suggests a learning-based approach that we have begun to examine. Humans use the time differential of frequencies below 1KHz and differential intensity of frequencies above 4Khz as the primary horizontal cues for sound localization [3]. This suggests that both time and intensity should be used to locate sound sources.

In the actual competition, we were unable to locate any of the mechanical noises as possible victims. However, we did employ other methods for finding victims, such as those based on vision.

Hors d'Oeuvres Anyone?: Person Identification

The *Hors d'Oeuvres Anyone?* competition was held during the final evening of the AAAI conference with robots providing the snacks at the conference banquet.

Our goal was to re-identify people that the robot had seen earlier in the evening at the reception. After identifying a person, we had developed a method for recording their spoken name so that we could replay it later when we recognized their return (e.g., "Hello NAME, I see you have returned for more cookies.")

Our person identification solution is an extension of Swarthmore College's successful robot, Alfred, from the 1999 competition [2]. Our methodology, however, differs in important ways. Our system:

- employs a multi-modal approach (using laser and motion data with the images)
- eliminates distracting backgrounds
- can identify people regardless of their position in a scene
- does not rely on colors to locate people, but does rely on colors to recognize them
- is optimized by a genetic algorithm

As it is approximately 18 inches off the ground, our passive laser range finder begins the identification process by searching for "legs". When an obstacle is found that roughly matches the shape of a human, the vertical boundaries are then relayed to the vision system. Using these boundaries provided by the laser, we can "crop out" background on each side of the target area.

However, this technique can still leave background image data above the heads and shoulders of our target. To remove this extra data from the image, we examine motion inside the laser crop lines. This is accomplished with a simple pixel-based differencing method over a few video frames. Our assumption is that the background will be stationary while we can detect some motion in the person.

Using the motion information as a border, we can therefore reduce the image considered so that we have focused largely on just the portion of the image that we are interested in (i.e., the person).

However, if we change the size of the image considered using the cropping methods described, we can no longer use standard color histogram methods as they are based on pixel counts over a static image size. To compensate for variable image size, we normalized the color pixel counts.

The final step in the color histogram creation is the actual building of the histogram. To account for differences in light intensities, we plotted the color pixel counts in a grid determined by red/green and blue/green ratios. Unfortunately, the resulting histograms for people occupy a very small region of the color space (the middle row of Figure 4).

To expand this region in a manner that would maximize the differences among people, we evolved the parameters for a warping of the space using a genetic algorithm. The fitness of each distinct set of warping parameters was determined by computing the differences among a set of test images, with larger differences getting the highest scores. After running a standard genetic algorithm (complete with selection and mutation) for several hundred generations, this resulted in a set of parameters that would produce histograms occupying a much larger area and thus more meaningful data (the bottom row of Figure 4).

A re-recognition of a person is made by comparing a histogram to a database of already-encountered people. If the difference is within a threshold, we can determine them a match, otherwise the new person is added to the database

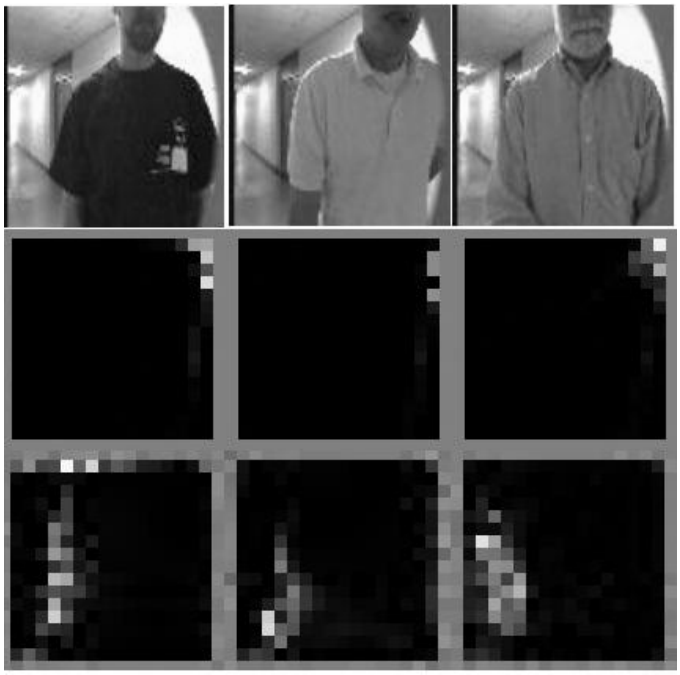

Figure 4: The top row shows the actual raw images grabbed from our video camera. The middle row shows the histograms computed from the raw images as described in the text. The bottom row shows those same histograms after applying the genetically evolved warping parameters.

along with a recording of their name. At this point, the robot system was also designed to ask questions of the person and record these data in the database with the histogram.

Our technique works well in identifying people that it has encountered before. Of course, this methodology depends on the color quality and quantities in an image. If a person were to change their appearance (say, by removing their jacket) the system would fail to recognize them. Further experiments need to be run to examine the capacity of the database and determine the overall accuracy of the method.

Conclusion

In preparation for participation in two events at the 2000 AAAI Robot Competition, our group developed a novel approach toward sound localization using distributed, low-cost sound cards, and further implemented a multi-modal approach for person recognition using laser and motion detection in building color histograms.

The procedure for determining the location of a sound source depends on each of the three distributed robots establishing the time of arrival of a sound by examining the sampled sound. Based on the difference in time of arrival each cell in a grid is assigned a value of the likelihood that it contains the source. The procedure works well in the lab, but less well in noisy environments in certain configurations. The procedure for identifying persons by their "color signature" worked well once a genetic algorithm determined how to weight the color spectrum so as to emphasize the differences between persons.

References

[1] D. S. Blank, J. H. Hudson, B. C. Mashburn, and E. A. Roberts. The XRCL Project: The university of arkansas' entry into the AAAI 1999 Mobile Robot Competition. In *Proceedings of AAAI Workshop on Robotics*, 1999.

[2] L. Meeden, B. Maxwell, N.S. Addo, L. Brown, P. Dickson, J. Ng, S. Olshfski, E. Silk, and J. Wales. Alfred: The robot waiter who remembers you. In *Proceedings of AAAI Workshop on Robotics*, 1999.

[3] G. Reid and E. Milios. Active stereo sound localization, technical report cs-1999-09. Technical report, York University, Ontario, Canada, 1999.