In *Proceedings of the Workshop on Measuring Performance and Intelligence of Intelligent Systems (PERMIS)*, IEEE Conference on Control Applications, Mexico City, Mexico, September 2001.

Designing Metrics for Comparing the Performance of Robotic Systems in Robot Competitions

Holly A. Yanco, *Member, IEEE*

Abstract – Many robotics competitions have been held over the past decade. These competitions often have the stated or unstated goal of comparing different robotic systems and their research approaches. When designing the rules for a competition, there are several ways to compare the performance of robots: objectively, subjectively, or a mix of the two. This paper discusses several robot competitions that have been held and how the metrics for judging performance were designed.

I. Introduction

Robot competitions bring together a group of people interested in a particular problem to demonstrate and discuss ways to accomplish a given task. Competitions often influence the direction of research in robotics, which can be used to great advantage. Indoor navigation is considered by many to be a solved task now, and this accomplishment was driven by several years of office navigation competitions in the AAAI Robot Competition and Exhibition. The latest additions to the AAAI contest are Robot Challenge and Robot Rescue, both of which include many hard research problems. Despite these good examples, when designing a robot competition that will compare research institutions, it is important to consider that a particular competition could drive research for several years.

Rules for robot competitions can take one of three forms: a ranked competition with subjective scoring, a ranked competition with "objective"[1] scoring, and a non-ranked competition with technical awards. A subjectively ranked competition should have clearly stated areas that will be judged and suggest guidelines for the judging. An objectively scored competition should have easily quantifiable metrics (e.g., number of objects found or amount of time taken to accomplish the goal). A non-ranked competition allows for more flexibility in the design of rules, since the lack of rankings will prevent any contentions that might arise in a ranked competition.

Competition metrics can be useful to compare research approaches. However, it is often very difficult to directly compare different solutions to the same problem. For example, at the Robot Rescue competition in 2001, one entry had treads and was teleoperated, while another had wheels and AI control software. In this case, task completion is used as a metric, rather than judging the methods used to accomplish the goal.

Competitions may be head-to-head or have each competitor run separately in the competition arena. The advantage of a head-to-head competition is that it is much more exciting for spectators, as they can root for one team over another. However, individual runs can be much easier for judges to watch and score, especially when the task is not one that easily lends itself to head-to-head competition.

II. History of the AAAI and RoboCup Competitions

In 1992, the first annual AAAI Robot Competition and Exhibition was held in San Jose, California. The introduction of this event marked the first AI robot competition and brought together many of the major robotics research laboratories and universities. This inaugural year introduced a competition involving navigation and identification of locations marked with encoded poles. Navigation continued to be a major component of the competition for several years, with office navigation as the primary focus. At the time of these early competitions, indoor navigation for mobile robots benefited greatly from the intense work in the area; the competition drove research forward.

Computer Science Department, University of Massachusetts Lowell, One Univeristy Avenue, Olsen Hall, Lowell, MA 01854.

[1] Many "objective" scoring methods involve some amount of decision that must be made by the judges, which introduces some subjectivity.

The AAAI Robot Competition has evolved over its ten years to include several other contests, each with different research aspects. *Find the Remote* was an event at AAAI-97 where a vision system was necessary in order to locate specified objects. *Life on Mars* was another competition that encouraged the use of computer vision; competitors needed to find colored "aliens" in a field of black boulders, then put the "aliens" into a "lander" with a colored door. The *Hors d'Oeurvres Anyone?* competition, introduced in 1997, encouraged the development of systems with good human-robot interaction, by creating robot servers that would both bring food to people while trying to entertain or interact with people. The *Robot Challenge* was first held at AAAI-99; the goal of this event is to have a robot register for the conference and give a talk about itself at an appointed time, after being dropped off at the entrance to the conference hall. In 2001, the *Robot Rescue* event was added, bringing an urban search and rescue scenario to the AAAI Competition.

Another robot competition, RoboCup, started in 1997. The goal of RoboCup is to have robots playing soccer with humans by the year 2050. The first five years have encouraged research in this direction by having several robot leagues, each of which encourage the development of different aspects of the research problem. In the small league, a camera placed above the arena allows for off-board vision processing. Larger robots have on-board cameras. The Sony dog league encourages research in legged locomotion for soccer, and the humanoid league is promoting the development of human-like robots, although there have not been any humanoid league soccer games at this early date. In 2001, RoboCup added a Robot Rescue league, held in conjunction with AAAI-2002. RoboCup also has simulation leagues for both soccer and rescue.

III. Designing Competitions and Metrics for Judging Performance

When designing any competition, the organizers must carefully consider the rules and scoring. The rules and scoring are often points of contention, so care must be taken to avoid skewing the algorithm towards any single research approach or robot base. Additionally, it is desirable to create a set of rules that are broad enough to encourage many different approaches, as this is likely to advance the state of the art more quickly.

Competitions fall into three categories:

1. Ranked competitions using subjective scoring based upon pre-specified criteria. The AAAI *Hors d'Oeuvres Anyone?* event is an example of this scoring method.
2. Ranked competitions using objective scoring using carefully spelled out criteria. The AAAI/RoboCup *Robot Rescue* event is an example of this scoring method.
3. Non-ranked competitions with technical awards. The AAAI *Robot Challenge* is an example of this type of competition.

A. The AAAI Hors d'Oeuvres Anyone? Event

The AAAI *Hors d'Oeuvres Anyone?* event was first held at AAAI-97 and has been an event in all of the subsequent AAAI Robot Competitions. The task of the *Hors d'Oeuvres Anyone?* competition is to serve hors d'oeuvres to people in a crowded reception. Robot servers should cover the entire space, in a attempt to serve as many people as possible. Entries may consist of a single robot or a team of robots.

The competition encourages human-robot interaction beyond driving food on a tray to people. In the first competition in 1997, one robot showed movie clips while serving food. Another team included a performance with their trio of servers, acting out a "Robotic Love Triangle." Almost all of the teams outfit their robots for the event, from masks to signs to butler uniforms. Some robots tell jokes when serving, while others try to greet people by name, using computer vision to locate a conference badge, extract the name region, perform character recognition, and then speak the result. Some of the years have provided bonus points for robots that could recognize VIPs by the color of the ribbons hanging from their conference badges.

Robots are also rewarded for recognizing that they need to reload their tray, either by counting the number of people served, by measuring the weight of the tray, or by using a computer vision system to judge when the tray is empty. Once the robot has determined that it needs more food (or a human attendant has made that decision for a robot unable to make its own determination), it should be able to guide itself back to a food reloading station. At this station, a human attendant reloads the food. While it would be desirable to have a robot reload its own food, there will need to be additional research into manipulators for mobile platforms.

When designing rules for competitions, it is important to consider the different robotic bases that researchers have in their labs. In this particular competition, the floors are flat and regular, allowing the majority of labs with wheeled bases to compete. The problem with many of the robot bases currently in use is that they are too short to interact effectively with people. To solve this problem, teams build structures on top of their robots to increase the robot's height to a person's waist height. Speech is also an important ability for robots in this competition; fortunately, relatively inexpensive systems are available to generate speech from text.

The robots are ranked using subjective scoring. In the 2001 competition, event judges awarded a subjective score of 1 to 10 in the following categories: ability to serve food, interaction with humans, interaction with other contestants, manipulation and sensing modes. To produce the final rankings for the event, the rankings determined by the event judges are combined with a popular vote. During the event, each attendee is given a token which is to be placed in the box of his/her favorite server. After the conclusion of the serving period, the votes are tallied and combined with the judges' scores to produce the rankings for the competition.

The metrics for determining the winner of this competition thus may have two disparate results: the crowd pleaser may not be the best technical entry. When designing a competition with metrics for technical judging and for popular voting, one should consider whether the two parts should have equal weight or if the technical aspects should outweigh the votes of non-roboticists. In the case of robotic servers, effective interaction with its audience is very important; a very technically-advanced entry that acts like a rude waiter may not be the best entry.

This competition is intended to serve as an entry level competition at AAAI. Undergraduate teams can be as successful as teams consisting of more advance robotics researchers. Additionally, the robot platforms can vary without too much of an effect on a team's competitiveness.

B. The AAAI/RoboCup Robot Rescue Event

In the *Robot Rescue* competition, the goal is to find victims in a collapsed building, which is represented by the Rescue Arena designed and built by the National Institute of Standards and Technology (NIST). The robots must report the location of victims to operators outside the arena. Entries may consist of a single robot or a multi-robot team.

The NIST designed rescue course has three areas: yellow, orange and red. In the yellow area, there are even floors, allowing wheeled bases to be used in the competition. The orange area has ramps and stairs with some rubble on the floor. The red area is the most difficult, with narrow collapsed areas and large amounts of rubble.

The differences in hardware and research approaches are more pronounced in this competition than in the *Hors d'Oeuvres Anyone?* competition, since two of the arena's areas are impassable to wheeled robots. In the 2001 competition, one team's entry was a custom built tracked robot that was teleoperated (future plans include the inclusion of AI software). Another entry used commercially available wheeled bases with custom AI software to navigate and locate victims. The wheels on the second team's entry precluded them from entering the orange or red areas. Since more points are earned for victims found in the more difficult areas, it is more difficult for a wheeled team to rank above an all-terrain team.

The Robot Rescue event debuted at AAAI in 2000. In 2001, the competition was held jointly at the co-located IJCAI-2001 and RoboCup-2001 conferences. At AAAI-2000, teleoperation was not allowed, as the focus of the AAAI competitions is the development of the algorithms. However, the inclusion of the RoboCup community, which includes many roboticists on the mechanical engineering side, warranted a change to this rule. The focus shifted from judging how the robot performed its task to how well it performed its task. A joint rules committee consisting of AAAI and RoboCup representatives designed the rules for the 2001 competition.

The rules of the competition focused on the desired outcome in a real search and rescue situation. It is important to be able to find all of the victims quickly and to report their locations to people outside the building. The reported locations should be accurate, and it is best if the robots are able to generate a map that would allow human rescuers to find the victims quickly. In a real rescue situation, it is better to have fewer human operators required for a robot, since there are restrictions on who can enter the "warm zone" around a disaster area.

The joint rules committee identified several variables to be used in judging the competition. All were spelled out carefully, resulting in an objective scoring algorithm.

The variables for the scoring algorithm are as follows:

- N is a weighted sum of the number of victims found in each region divided by the number of actual victims in each region.

- C_i is a weighting factor to account for the difficulty level of each section of the arena: C_{yellow} = .5, C_{orange} = .75, and C_{red} = 1.0.
- N_r is number of robots that find unique victims.
- N_o is the number of operators.
- A is an accuracy measurement for the location of each victim: A = F/V. F is equal to 1 if the victim is in the reported volume, and 0 otherwise. V is the volume in which the reported victim is located, given by the operator in the warm zone to the judge. The average accuracy is used in the scoring algorithm.

Each team ran for twenty five minutes; the best two scores from four runs were used to determine the final score. The algorithm for determining the score of a round is as follows:

$$Score = N * \frac{N_r}{(1+N_o)^3} * \overline{A},$$

where

$$N = \sum_{i=\{yellow,orange,red\}} \frac{C_i * N_{victimsDetectedS_i}}{N_{actualVictimsS_i}}$$

In order to receive a ranking in the competition, the competitors needed to meet a minimum score requirement, which was equivalent to finding all of the victims in the yellow zone. No competitor earned the minimum score in 2001, although two teams were close. Instead of rankings, two technical awards were presented by the judges, one which rewarded the development of mobility for rescue and the other which rewarded the development of AI algorithms for rescue.

C. The AAAI Robot Challenge

The task of the AAAI *Robot Challenge* is to have a robot attend the National Conference on Artificial Intelligence. The event is started when a robot is dropped off at the entrance to the conference center. The robot needs to find the registration desk for the conference, which it may do by asking people for directions and assistance. After registering, the robot needs to find a specified conference room and give a talk about itself at a specified time.

The event is very challenging for the robotics field and includes many open research problems. The intent of the event is to encourage senior robotics researchers and graduate students to bring their work to AAAI. Since there are many areas of research involved in this problem, it would be difficult to rank the competition entrants. Instead of rankings, judges may give technical awards. Examples of possible awards are innovation in localization and navigation, innovation in robot vision or sensor technology, innovation in human-robot interaction, innovation in real-time planning, innovation in manipulation, and excellence in collaboration and integration. The advantage of a non-ranked competition is also that people may be more willing to demonstrate work in progress, resulting in additional communication between researchers.

IV. Conclusions

When designing performance metrics for competition, a rules committee must decide what is important. Task completion may be the most important goal, as it is in the *Robot Rescue* competition; it may not be important how a victim is found, as long as the person can be rescued. Other competitions may choose to allow partial completion of the specified task, judging instead a demonstration of good research and/or intelligence. Some of the aspects of the *Hors d'Oeuvres Anyone?* rules include this approach. The initial stages of the *Robot Challenge* also reward partial completion, although the ultimate goal is task completion.

A competition must also decide whether it aims to showcase new research or systems that are ready for deployment. In the case of the *Robot Rescue* event, wheeled robots may be used to demonstrate new algorithmic capabilities, but can not score as highly as a tracked robot in the more difficult areas. In contrast, the *Robot Challenge* allows new research to be showcased and eliminates most of the performance pressure with the removal of rankings.

All of these approaches have valid purposes. When designing a new competition and set of rules, determining the desired outcomes of the event should be the first task. This step will help to determine whether the scoring should be objective or subjective. The next step should be designing rules that can include multiple robot bases and research approaches. Whatever the design, the rules should be clearly spelled out and available as far in advance of the competition as possible.