The Knowledge Engineering Review, Vol. 26:1, 11–17. © Cambridge University Press, 2011
doi:10.1017/S0269888910000354

Robotics competitions as benchmarks for AI research

JOHN ANDERSON, JACKY BALTES and CHI TAI CHENG

Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, R3T 2N2 Canada;
e-mail: andersj@cs.umanitoba.ca, jacky@cs.umanitoba.ca, tkuggt@gmail.com

Abstract

In the last two decades various intelligent robotics competitions have become very popular. Arguably the most well-known of these are the robotic soccer competitions. In addition to their value in attracting media and capturing the minds of the general public, these competitions also provide benchmark problems for various robotics and artificial intelligence (AI) technologies. As with any benchmark, care must be taken that the benchmark does not introduce unwarranted biases. This paper critically evaluates the AI contributions made by various robotic competitions on AI research.

1 Introduction

Up to the latter half of the 1980s, artificial intelligence (AI) was criticized for producing small results, which served as a shaky foundation for grand theories. Many ideas operated provably but only on paper, and even those systems that were implemented operated only small conveniently contrived environments. Worse still, researchers used these small results to convince granting agencies, the public, and one another that much greater things were just around the corner, and this cycle continued.

In the late 1980s and early 1990s, researchers began to get a perspective on this cycle, and a number of important realizations resulted. The most significant of these was the argument that situated and embodied approaches would be a much stronger vehicle for advancing our field. We began to study the environments in which an intelligent entity was expected to perform, the interactions between the embedded entity, its expected activities, and its environment. Some of the PhD theses of the day argue for many of the principles of this approach (Agre, 1988; Chapman, 1990), which came to be known as the *intelligent agent* perspective from which much of today's research is performed.

A focus on situated interaction led to significant realizations about the importance of perception, the balance between reactive and deliberative reasoning, and the importance of empirical performance evaluations in realistic environments. At the same time, advances in microprocessors, memory, and batteries led to much more affordable mobile robotics platforms, and together these created a renaissance in embodied AI. The use of robotic vehicles allowed for immediate testing in much more realistic settings than were typically envisioned prior to this era, and led to significant changes in how we approach the design of intelligent agents. Robotic embodiment also allowed for more direct comparison between alternative theories, especially once a number of common commercial platforms (e.g. the Pioneer series of mobile robotics) became more commonplace.

It was in this environment that the first robotic competitions were envisioned. AAAI began holding their broadly varying series of competitions in 1991 (Balch & Yanco, 2002), while in 1992 Alan Mackworth proposed the grand challenge domain of robotic soccer (Mackworth, 1993). The reasons behind these were different, but together they sparked a range of efforts, from the

Federation of International Robot-Soccer Association (FIRA, since 1996) and RoboCup (since 1997) competitions (Fira, 2009; RoboCup, 2010) to the DARPA grand challenge, proposed in 2003 (DARPAGC, 2003).

Beyond encouraging embodiment and situated reasoning, there are a number of very good reasons for using robotic competitions to advance AI. The most obvious of these is increasing rigour by creating standard benchmark problems by which the performance of research can be judged. The fact that this takes place in the real world means that not every possibility can be anticipated in advance, and the brittleness of problem approaches must be considered. The act of competition in real time brings important real-world considerations too: a solution must be ready by a specific time, bringing elements of maintenance and adaptability into consideration when making design choices. While a benchmark could equally be applied in a laboratory situation, public measurement in a competition setting helps to eliminate some of the self-censorship present in traditional means of disseminating scientific research: if something does not go well, that fact is obvious. It may be much less obvious *why* something went wrong, but the evidence at least invites questioning. Such negative results are much more difficult to come by through reading the scientific literature.

Intertwined with the idea of improving scientific rigour is an intent to capture the interest of both scientists and the general public. If a problem is shown to be complex and interesting, and if it can be made relevant in the real world, it is likely to attract the attention of more scientific resources, both in terms of the devotion of scientists' time as well as a translation into research funding. There is strong merit both socially and scientifically to inspiring the general public as well. Just as the race to land a human on the moon served as grand challenge to unite the minds of mankind in the 1960s, it is conceivable that the goal of building functioning, interacting, generally intelligent robots can serve a similar purpose today, and with similar effects: inspiring young people towards science and engineering, and bringing peripheral research breakthroughs in related areas. One of the interesting elements of developing and managing competitions is the fact that these goals are not necessarily in harmony with scientific progress: something that serves as a strong benchmark may not look visually inspiring to young people, for example, and elements of competitions that are publicly popular may contribute little directly to science.

The purpose of this paper is to review some of the contributions of robotics competitions from a research standpoint, as well as to critique some of their weaknesses in order to improve future competitions. We begin by examining the AAAI robot competitions, move to variations on Robotic Soccer (and other RoboCup competitions, such as the Urban Search and Rescue (USAR) competition), and finally examine the current FIRA HuroCup.

2 AAAI/IJCAI robotics competitions

AAAI (along with its sister conference IJCAI), have been running robotic competitions and exhibitions since 1991. A good historical overview is (Balch & Yanco, 2002), while these competitions also publish a summary each year in the *AI Magazine* and also an associated workshop publication describing entries. Part of AAAI's mandate is to generate interest in AI, as well as to serve as a central organization for researchers in the field. As the first hosted international robotics competitions, they served both of these mandates well. AAAI's modern competitions differ from many others mainly in their broad focus and lack of emphasis on the competitive element. While there are awards and a winners list, as past attendees we can attest that there is a sense of camaraderie and a much more laid back atmosphere than most other competitions. Part of this is the broad nature of the tasks typically involved. Early competitions had a strong focus on specific tasks (e.g. cleaning up a tennis court). However, as early as 1997, broad tasks such as serving hors d'Oeuvres (including audience interaction), and registering and participating in a conference were used. These provided challenges that were significantly beyond contemporary (and even current) technology, and served to inspire and capture the imagination of attendees (though it is comparatively rare that the general public makes its way to AAAI and IJCAI). Because of the complexity of these challenges, teams were encouraged to demonstrate approaches to pieces of a

problem. This made entrants extremely difficult to compare, making these undesirable from the standpoint of setting benchmark problems. This factor also served to make the exhibition element at least as strong as the competitive element: if I am working on a different but related facet of a problem than you, I am more likely to see the differences and commonalities of our work than focus on minutia in order to place more highly than you. In Balch & Yanco (2002), discussion about the earlier more specific competitions mentions the fact that entrants found themselves having to explain why their performance was in fact due to good science and not taking advantage of rule loopholes or last-minute hacks. This is one of the first mentions of a problem also common to robotics competitions: a focus on winning as the end goal, rather than the underlying scientific advancement that is intended. More recently, the same problem was identified in the reinforcement learning competitions (Whiteson *et al.*, 2010).

AAAI also introduced the National Institute of Standards and Technology (NIST) USAR testbed in the 2000 competition (Jacoff *et al.*, 2003). In this competition, robots are expected to traverse unmapped terrain with significant challenges both physically (e.g. stepping fields, multiple levels) and perceptually (glass, mirrors, repeated visual patterns, moveable debris that will confound wheel-encoders), while looking for simulated human victims. This is more strongly associated with providing a benchmark problem than the other competitions that have emerged from AAAI (though robotic soccer has also been used as an event at the AAAI robotics competitions in the past). Both USAR and soccer are much more strongly associated by the research community with RoboCup, however, and we will discuss these in the next section.

3 RoboCup/Federation of International Robot-Soccer Association robot competitions

Robotic Soccer is a good example of a task that can both serve as a benchmark problem and inspire both scientists and and the general public. It is a complex, real-time domain that involves perception and multi-agent strategy, and leaves room for almost every area of AI one can think of, from learning to planning to reasoning about uncertainty. It is also a fast-moving game that is played for fun by most human cultures, so it has a broad reach. Two organizations, FIRA (Fira, 2009; who had the earliest robot soccer competition) and RoboCup (RoboCup, 2010; which has grown to be by far the largest robot competition in existence) have championed this as challenge problem for the robotics community. RoboCup has arguably done the better job, through organizing its various competitions towards the umbrella challenge of having a robotic soccer team beat the best human team by 2050. There is no doubt that both of these competitions have had a huge impact on the scientific community, and have brought about many important results.

Each of these organizations has a range of soccer competitions organized by size and physiology (various sizes of wheeled and humanoid robots). With each of these organizations having been in existence for more than 10 years, both have served to illustrate another good example of competition environments: they are adaptable over time and allow difficulty to be increased as technology improves. Various RoboCup leagues have rule changes over time ranging from removing walls that stop robots from leaving the field unintentionally, removing fiducial markers that allow easy localization, and changing the size of playing fields to account for greater manoeuverability.

The evolution of both RoboCup and FIRA also illustrate some of the problems that can occur in competition environments. In the quest for a good benchmark, the designers of the competition attempt to anticipate variation as much as possible in terms of competition rules. For example, consider humanoid robotic soccer: positioning a foot behind the ball is a difficult problem, rendered much easier if a camera is placed on the foot; similarly, larger feet make it easier for a robot to stay stable, and larger robots can cover ground more quickly. All these and many other variants must be considered in terms of competition rules, along with many changes to rules to make the game adaptable to play by robots. The number of elements that must be considered so that play is fair to all teams leads to a strong focus on the rules, and on exploiting elements that are not

contained within them. The result is instead of developing the adaptable, flexible approaches that one would want to associate with AI, entrants are tacitly encouraged to create special-purpose solutions useful for operating within an extremely specific set of rules: something not far off in spirit from the limited laboratory situations that competitions were in part designed to combat. In both RoboCup and FIRA competitions, teams also have a great deal of input to the rules, leading to self-reinforcing behaviour: if one has invested a great deal in a specialized solution, it is natural to want to make that solution applicable in future.

Consider the small-size league at RoboCup, for purposes of example (though all leagues in robotic soccer competitions are at least partly guilty of many of the same elements). The performance of any team in this league relies much more on fast omni-directional drives and specialized devices such as chip-kickers and dribble bars. The former of these allows the ball to be shot at great speed from anywhere on the field, and gives the sound and appearance of a pool ball being shot rather than anything resembling human soccer. The latter is a device that allows the ball to be brushed from overhead, allowing it to move with the robot in any direction (much like picking the ball up and rotating it in place). These two devices lead to teams deemphasizing intelligent performance in place of elements that do not have an equivalent in human soccer, but are easier to develop than good AI. However, they will likely never be removed, in part because of the vested interest in the teams that have developed them. Another strong factor, however, and one that is important to discuss in the context of the contributions of robotics competitions, is the fast moving appearance they provide. The latter is an important factor to consider in terms of evaluating the contributions of a competition: these games are exciting to watch, and provide the general public (unlike AAAI, RoboCup attempts to be a large public draw) with a good show. It can be argued that this is part of the inspiration that a challenge problem is supposed to provide, and that it helps pay for what is in fact an expensive event to stage. However, there is a fine line between providing a public spectacle and advancing science, and the relative strength of each of these must be considered: there is a great danger of style taking precedence over substance. This is not the only example of a questionable balance in RoboCup: Sony AIBOs, for example, were kept around long after they ceased being manufactured, in part because they looked cute to the public.

It is important to note that despite soccer being a challenging game, it is still a very restricted domain compared to much general human activity. For example, even on an unmodified soccer field, localization is not a strong challenge compared to the general localization problem, because of constrained elements (field lines, team uniforms). Perception, while challenging, is still also far easier in this constrained domain. Because of these restrictions, less structured challenge problems have emerged (as noted, also partly because of the AAAI competitions), such as those embodied in NIST's USAR test arenas (Jacoff *et al.*, 2003) mentioned earlier. Localization and perception are two extremely challenging problems here, because of the lack of ability to assume regularity in the domain. Similar problems in encouraging one-time solutions over general problem solving can also be seen in these competitions. Recently, for example, the RoboCup rescue moved to using angled floor panels (convex and concave) to replace flat surfaces in the simplest of its arenas. Rather than developing very general approaches to traversing terrain, teams were encouraged to deal with this precise type of angled flooring. Although the NIST arena does contain more challenging areas that feature stepping fields, entrants restricting themselves to one area of the arena need only work on one specific solution. In the past, the RoboCup Rescue has also encouraged specialized hardware over general solutions in terms of its scoring. For example, a multiplied score in 2005 could be tallied by detecting a single victim multiple times using different forms of sensing. This allowed a human operator to, for example, perceive a victim themselves through their robot's camera, and then use the alternative forms of sensing (e.g. a heat, motion, CO_2 detectors) provided on the robot to re-flag the location of the already-known victim. This promotes a high score through additional hardware that may not be necessary, rather than focussing on qualities of adaptability and generality that we usually associate with intelligence. Because of the extreme difficulty of this problem, most approaches rely on at least some form of teleoperation, which does encourage improvement in human–robot interaction, but does not serve

as a strong incentive to focus on autonomy. Since most teams are not fully autonomous, this becomes self-reinforcing: rule changes tend not to encourage autonomy, because it is not an advantage to existing teams.

4 HuroCup: improving evaluation through breadth

An obvious way to overcome the problems of special-purpose solutions is to emphasize and require breadth, robustness, and versatility in all aspects of a competition. This section describes our work on FIRA HuroCup humanoid robot competition as an example of how this breadth can act as an efficient benchmark for AI systems and provide meaningful research results for the participants.

The FIRA HuroCup is the oldest intelligent humanoid robot competition, with the inaugural competition taking place in June 2002 with five teams. Since the HuroCup event is organized as part of FIRA, the initial plan was to develop a soccer competition for humanoid robots. However, it became quickly apparent that soccer did not provide a good benchmark problem for humanoid robots. Since soccer was played on a flat hardwood surface, many teams quickly developed efficient walking gaits and kicks for this surface. The main challenge then was to develop localization (where is the soccer player on the playing field?) and mapping (where are the other players and the ball?) methods for the players. However, localization and mapping are not specific problems for humanoid robots and research in these areas can be done without much change from wheeled or other walking robots.

Therefore, the HuroCup committee decided to focus on open research problems that are more closely associated with humanoid robots in particular. The main open research problems in humanoid robotics fall into several areas:

Active balancing Humanoid robots must be able to walk over various even and uneven surfaces. They also must be able to adapt their walk to changes in the weight and balance of the robot (lift-and-carry, weight lifting).
Complex motion planning Humanoid robots can perform many different actions. The sheer number of these movements mean that they can not all be pre-programmed. Instead a humanoid robot must be able to plan new motions online (e.g. opening a door with an elbow or foot if the robot is carrying barrier to operate a light switch or to pick up a box from under a table).
Human–robot interaction a humanoid robot must be able to interact naturally with a human which entails that it is able to understand speech, facial expressions, signs, and gestures as well as generate speech, facial expressions, and gestures.

As one of the advantages of the humanoid form is its flexibility and applicability to a wide variety of problems, some of these areas are naturally associated with robustness and breadth (e.g. walking vs. walking on uneven terrain vs. walking while carrying a load).

In deciding on challenge events, we and the other members of the HuroCup committee looked for those that would specifically advance research in these areas, as well as considering what would most encourage robust solutions and work well in a public challenge environment. To avoid exploiting rules in one large challenge environment attempting to encompass all humanoid skills, we instead focussed on dividing the FIRA HuroCup into a series of events that each test a subset of interacting humanoid skills. Scores in the individual challenges are summed, so that in order for an entry to do well in the HuroCup, a *single robot* must perform and score well across the range of events. Any special hardware development that focusses on doing well in one type of activity becomes redundant in others that do not require that specialization. Such additions can also be severely detrimental in two ways. First, given the limited time available in and around a competition, additional hardware, and control that serves no purpose in some events draws support and resources away from the limited pool available to a team as a whole. More directly, the addition of such equipment may be strongly detrimental to the performance of other events (e.g. specialized arm motors making the robot more top-heavy, making adaptive balancing more difficult).

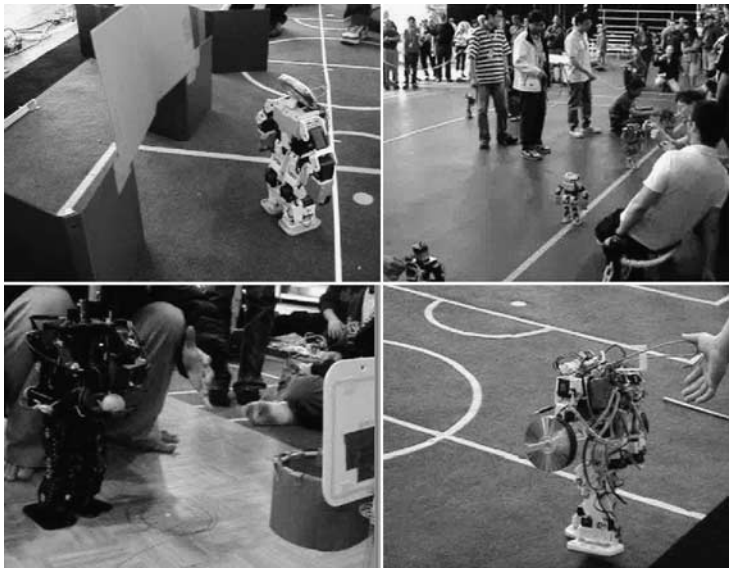

Figure 1 Four events in the 2007 and 2008 HuroCup. Top: obstacle run (L), marathon (R); bottom: basketball (L), lift and carry (R)

All HuroCup events require a fully autonomous robot that has all sensing and processing on board. No interaction from a human is allowed. HuroCup 2009 consists of the following eight events, some of which are depicted in Figure 1:

Sprint The humanoid robot must walk a distance of 3 m in a straight line forwards and then backwards. This means that a robot must possess at least two fast walking gaits. This event is really intended as a starter event which allows beginning teams to score some points. The remaining events are more difficult.

Obstacle run The humanoid robot must cross a 3-m long region to reach the end zone without touching any of the obstacles. There are three types of obstacles: walls, holes, and gates. A robot must not step into a hole, but can crawl through a gate to reach the end zone.

Penalty kick is similar to penalty kicks with the difference that the ball is placed randomly in front of the robot.

Lift and carry A robot must carry an increasing number of weights over an uneven stepping field. The stepping field is colour coded so that the robot can recognize steps. This is an advanced challenge and many teams have problems with it.

Weight lifting The weight lifting competition was introduced to provide a slightly simpler active balancing challenge than lift and carry. A robot must lift as many CDs as possible. However, since we did not want to test the shoulder motor strength, the robot must walk 30 cm with the weight low and then 30 cm with the weight above its head. This means the centre of mass of the robot changes drastically, but predictably and the robot needs to compensate.

Basketball A humanoid robot must pick up a table tennis ball randomly placed in front of the robot and throw it into a basket.

Marathon A humanoid robot must cover a distance of 42.195 m as fast as possible without being allowed to change its batteries. The event was the first HuroCup event that takes place out-doors, which means that teams must cope with more uneven surfaces and lighting conditions.

Climbing wall A humanoid robot must climb up a wall where foot and hand holds were placed randomly. This event was introduced in 2009.

The combination of events represent most of the range of activity expected of a humanoid, and the requirement of doing well in a range of events ensures breadth in evaluation. To do well in a dash, for example, a robot must have a fast start, but need not have fine control once it is moving.

On the other hand, completing the marathon (Figure 1, second from top) requires following a lined track over a long period of time. Specialized hardware for either of these does not likely help the other. This is even more obviously seen in basketball, where teams have rarely attempted to use any special throwing motors, since the extra weight will tend to decrease performance in weight lifting and running events. Whereas most other robotics competitions still developed special technology to solve specific problems (e.g. teams in the small-sized league at RoboCup have developed special rubber mixtures to increase maximum acceleration of their robots and a range of kicking devices), HuroCup robots are still a core humanoid robot with two legs, two arms, and a camera.

The events are constantly updated to reflect the current state of the art. In 2009, for example, the climbing wall event was added, but major changes to other events were also introduced. In the lift and carry, the robot must pick up the weight lifting bar whereas previously, it could start with the bar in hand.

Breadth in the HuroCup is not just achieved through a range of events, but in the spirit of its rules. For example, the HuroCup rules deliberately leave breadth in the specification of all measurements and colours. The last author still remembers incredulously that he had to repaint centre lines at 2:00 am for the RoboCup 1999 small-sized league competition, because a team showed that the lines were 2 mm wider than specified in the rules and and claimed that their vision system was unable to compensate for the extra width. In general, properly steering the evolution of a competition can go a long way in removing the focus on narrow, specialized solutions. In the 2004 HuroCup, for example, roughly half of the team used infra-red distance sensors to detect obstacles in the obstacle run – an obvious piece of specialized hardware that should be illegal. Instead of simply disallowing those types of sensors, the obstacle run competition was extended to include gate and hole obstacles, which cannot be detected easily with these sensors. This led to fewer teams using infra-red distance sensors without having to disallow them in the rules and few teams complained when infra-red sensors were finally disallowed for 2009.

The organization of the HuroCup competition also incorporates lessons learned from other robotics competitions, to make improvements from both participant and spectator viewpoints. Teams rotate swiftly through the events (e.g. every one to three minutes a new robot will perform a task), and thus spectators do not need to watch a badly performing robot for 30 minutes. All events are repeated on different days, allowing a second chance when technical problems occur.

5 Conclusion

In this paper, we have looked back at almost 20 years of robotic competitions, and highlighted some of the pitfalls associated with employing competitions as scientific benchmarks. We argue that breadth and versatility must be embraced both at the event level and in terms of the spirit of the rules in order to avoid a strong focus on special-purpose solutions or loopholes in the rules.

References

Agre, P. E. 1988. *The Dynamic Structure of Everyday Life*. PhD thesis, MIT.

Balch, T. & Yanco, H. 2002. Ten years of the AAAI mobile robot competition and exhibition. *AI Magazine* **23**(1), 13–22.

Chapman, D. 1990. *Vision, Instruction, and Action*. PhD thesis, MIT.

DARPA. 2003. *DARPA Grand Challenge Conference 2003*. `http://www.darpa.mil/grandchallenge04/conference_1a.htm` (retrieved December 16, 2010)

FIRA. 2009. *Fira HuroCup Rules*. `http://www.fira2009.org/gamerules/hurocup-laws.pdf`

Jacoff, A., Messina, E., Weiss, B. A., Tadokoro, S. & Nakagawa, Y. 2003. Test arenas and performance metrics for urban search and rescue robots. In *IEEE International Conference on Intelligent Robotics and Systems (IROS-2003)*, Las Vegas, NV, 3396–3403.

Mackworth, A. K. 1993. On seeing robots. In *Computer Vision: Systems, Theory, and Applications*, Basu A. & Li X. (eds). World Scientific Press, 1–13.

RoboCup. 2010. *RoboCup Website*. `http://www.robocup.org` (retrieved December 16, 2010)

Whiteson, S., Tanner, B. & White, A. 2010. The reinforcement learning competitons. *AI Magazine* **31**(2), 81–94.