Improving Robotics Competitions for Real-World Evaluation of AI

John Anderson, Jacky Baltes

Autonomous Agents Lab
Department of Computer Science
University of Manitoba, Winnipeg, Canada

Kuo-Yang Tu

National Kaohsiung First Univ. of Science and Tech.
Kaohsiung, Taiwan

(Videos originally linked in this talk may be found on our lab website)

Situated/Embodied Approaches

- Have been a significant boon to AI research: discourages slippery slope of simplifications *("I'm not working on vision, so I'll assume my planner can always see things...and grip them...and not drop them...and....")*
- Chief among such approaches are robotic domains, which force us to at least partly consider many sensory/effectory elements and the uncertainty surrounding the real world
- Richness of real-world embodiment complicates system evaluation

Evaluating Intelligent Systems

- A complex embodiment can be brittle: if vision fails, elaborate motion control won't help much
- Physical dependency on the world causes many potential problems, exacerbated when >1 robot is involved
  - repeatability in positioning, failure of equipment, difficulty in measurement/data gathering
  - Even keeping a large number of pieces of equipment running for evaluative purposes is challenging
- Part of an overall problem of lack of control: randomizing positioning, consistent distribution of random events
- Simulation still has a role: improving control, consistency, reliability

Risks of Simulated Evaluations

...but it can't be relied upon as the sole means of evaluation

- added risk that one may be simplifying the environment (self-bias) – overly optimistic results
- removing some of the very factors that drive research: *failures reveal problems and set new goals*

- In part as a reaction to the problems of controlled experimentation and repeatability, while keeping things in the real world, *structured challenges* and *competition scenarios* for evaluating AI systems
  - distinct concepts, but overlapping elements

Structured Challenges

- Provide everyone with a common grounding for potentially different research focuses
  - Ability to speak the same language from a task perspective, and objectively compare performance
- A broad set of controls over a domain, partly to define a problem and partly to ensure fair comparison, coupled with metrics for measuring success
  - Some of these are also controls to deal with the limitations of current technology (adaptable over time)
- For any significant domain, these end up being *lengthy and legalistic* (e.g. foot size vs. height in a humanoid robot)

Structured Challenges

Competition Environments

- Competition scenarios involve direct evaluation of approaches by comparison *in situ*
- My approach must work *now*, not three hours from now
  - Drastically changes the value of some elements: robustness, ease-of-setup, maintainability, *parsimony*, are incredibly important
  - They should be in the lab too, *but there's nothing forcing us to consider it*
- Eliminates self-censorship in dissemination: reporting successes but not failures – there is as much or more to learn from why some approaches did not work!
- Still debate on applicability of the model to education, but little question of the positive benefits laid out here

Goals of a Good Challenge

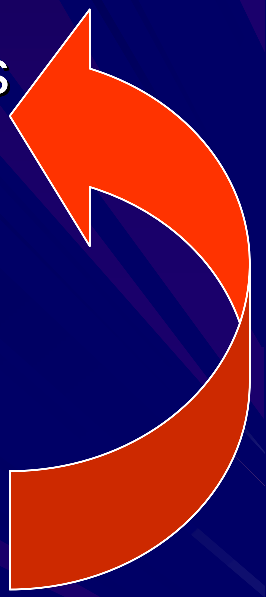
- Grounded benchmark for comparison of approaches, with hard real-world constraints
- Ability to change format as technology develops
- Grounding for research in many areas: ML, MAS, Vision, Planning, Mechatronics…
- Connected to applications
- Motivating to students – showcase research
- *Motivating to the public/funding – PR (fast-moving, attractive)*
- *In a competition setting, everyone has a motivation to do well (win?)*

Common Challenges/Competitions

- Robotic Soccer is the most commonly-known of these challenge areas. First proposed by Alan Mackworth, UBC (1992), First implemented by Kim Jong-Hwan, KAIST (FIRA, 1996) followed by Kitano Hiroaki, Sony (RoboCup, 1997)
- Also broader challenges, such as RoboCup Rescue

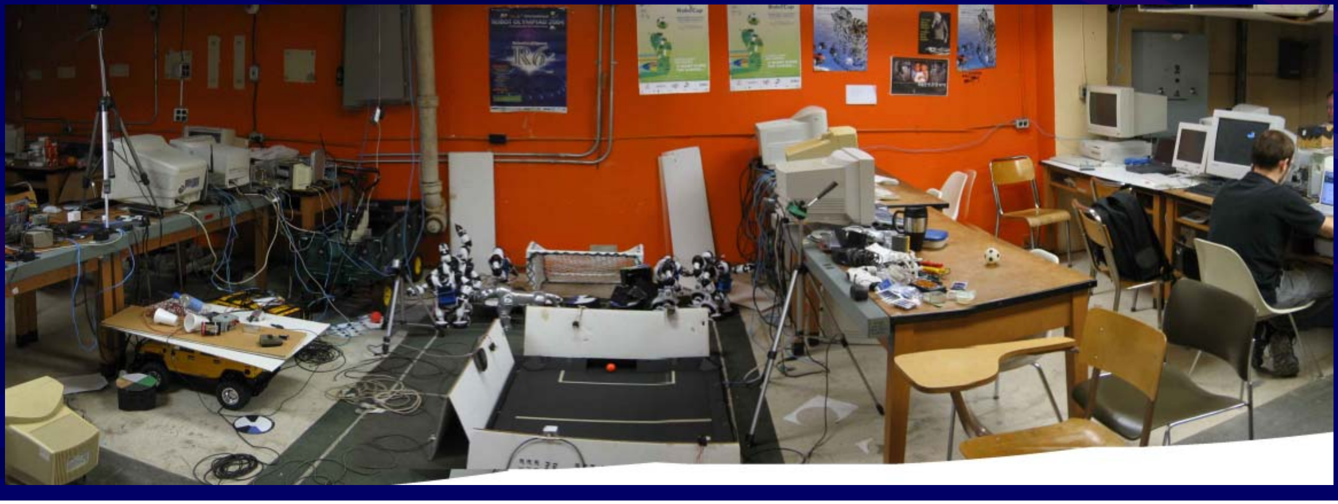
Problems from a Research Standpoint

- Desire to win changes the goal significantly: research matters most on paper, not necessarily in practice
- Scouring rules for minutia to find subtle/obvious ways of getting ahead of everyone else
- Advocating rules useful to performing well in the confines of the current controls, as opposed to advancing research (e.g. foot cameras)
- Rules grow exponentially as a result, leading to a subtle increase in acceptance of this (negative cycle)
- Difficult to remain devoted to promoting the research
- This also holds from an organizational standpoint: flashier = better = more spectators and sponsor $



Robotic Soccer Should Work (almost) Anywhere

From another standpoint, we're saying it doesn't even have to work in the confines of an average lab anymore!



RoboCup, 1997-Present

- Ever-increasing reliance on specialized hardware
  - e.g. powerful kickers, dribble bars, omnidirectional drives, chipkickers
- Unlikely such devices will ever be removed – motivation is self-perpetuating, and arguably it makes for flashy play (organizational bias)
- Too much focus on single-instance solutions that don't advance research beyond this single problem instance (myopia/forgetting why we're here)
- *Excellent orange-golf-ball retrieval-and-launch systems*
- Rescue: Strong bias to teleoperation because of additional complexity, limiting value to a lot of AI...also similar problems with fixation on cool hardware

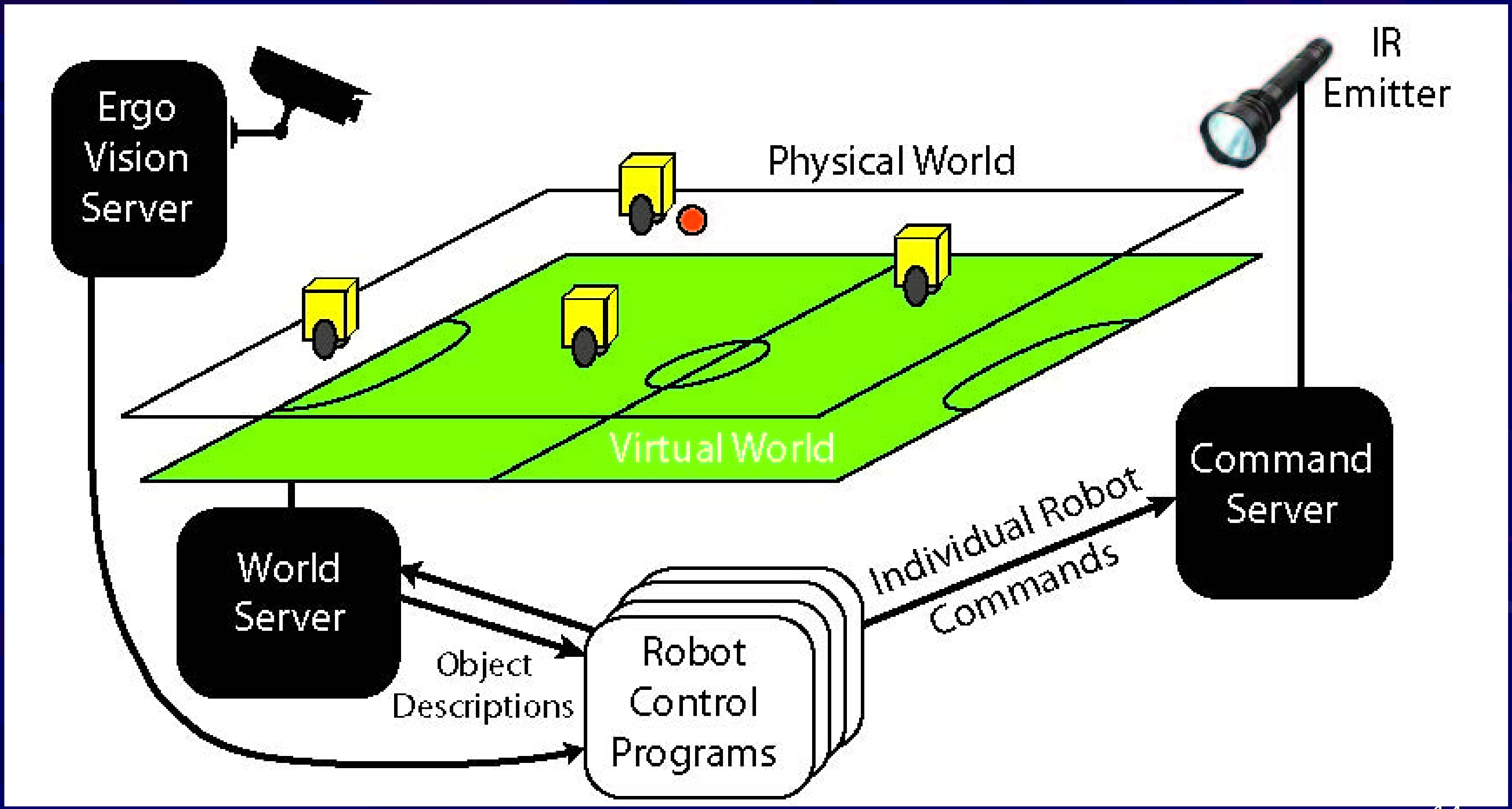
Two Improvements

- How can we improve challenge-based and competition-based environments?

- Improving control and repeatability outside of competitions: *Mixed Reality*
- Improving Competitions by emphasizing breadth and adaptability: *FIRA Hurocup*

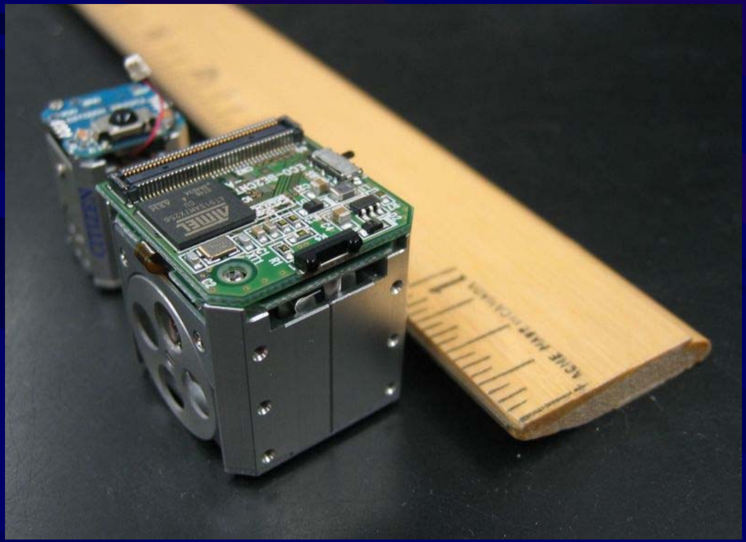
Mixed Reality

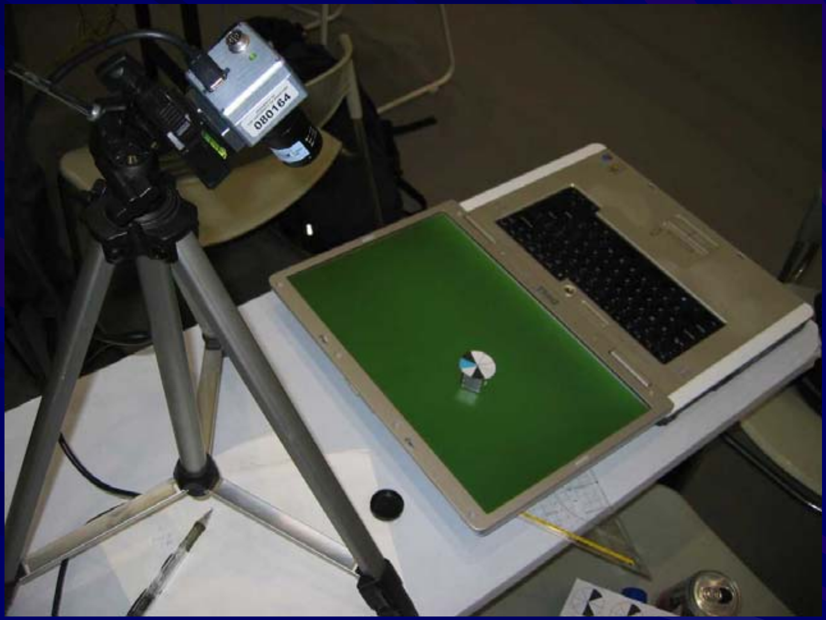
- A combination of both a physical and virtual environment, bringing the advantages of both



Mixed Reality

- Useable with any robots: size and field size are obviously correlated





- MR League in RC since 2007, based on our E-League

Evaluation Advantages

- Virtual element allows much of the control that simulation provides, but grounded over physical element
- e.g. control of randomness, positioning in when evaluating a dynamic path planning approach

- Repeatability is enhanced; breakdowns still occur but accurate re-placement is supported
- Automated detection of occlusions

Evaluation Advantages

- Virtual elements allow consistent control of any almost any element desired: e.g. hockey, a virtual puck allows more detailed modeling of when transfer from one agent to another should occur (and consistency in judging goals)

- Sticks can be modeled, allowing various types of shots (virtual actions – hardware independent!)
- The world can be made more stringent (turns on ice – prohibit hard angles)

In a Competition Environment

- In a competition environment, allows consistent hardware, consistent virtual actions between teams: concentrate on developing good AI!
- *Rules still need to be written to promote the spirit of adaptability/breadth as opposed to encouraging one-shot solutions*
- e.g. in education, we work with a series of challenges, where it is known the current one will be the basis for the next, but not what elements are going to be added
  - path following, path planning, dynamic path planning, abstract behaviors, complete applications

Better Competitions: HuroCup

- Similarly, we want rules at the research competition level that equally encourage such breadth – there is no natural scaffolding as there is in an educational setting, though
- As an example of these, we consider our work with the FIRA HuroCup rules
- RoboCup is to the Oscars as FIRA is to Sundance
- Hurocup is FIRA's humanoid division: moved away from soccer because of the types of problems overviewed earlier, and because soccer (under the conditions we see at RoboCup) is not the best challenge problem for humanoids

Promoting Breadth

- Don't allow a single winner to be determined by a single small skill set
- Use multiple events, where good performance in a number of events with similar skill sets will not be enough
- Especially important in humanoids: the humanoid form is broader/more flexible than any other embodiment: challenges must be similarly broad
- Should reflect main research problems in humanoids: Active Balancing, Complex Motion Planning, Human-Robot Interaction
  - Researchers have a vested interest in leveraging this breadth – important to step beyond soccer

Challenges

- Must use a single autonomous robot with no alternations between events
  - All sensing and processing must be on-board
- Any special hardware developed for one type of event is probably a liability in others (at least in the context of being redundant while still affecting weight, COM, balance…)
- 8 different events, organized specifically so that doing well in one likely presents a challenge in others

Challenges

- Sprint: 3m in a straight line forward and then backward
- Marathon: 42.195m without being allowed to change batteries. 2m push-back if your robot falls and cannot get up on its own. Will be done outdoors in 2009

Challenges

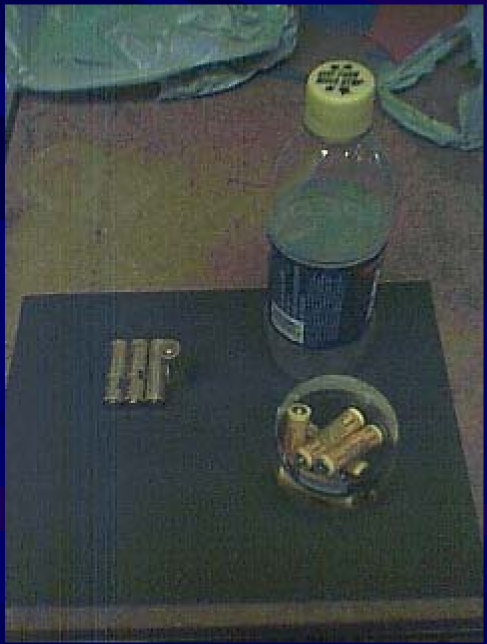
- Basketball: pick up a table tennis ball randomly placed in front of the robot, throw into a basket (previously: began with ball in-hand)

Challenges

- Weightlifting: lift as many CDs as possible on a bar: walk 30cm with the weight below the head, 30cm with the weight above – drastically but predictably changing COM

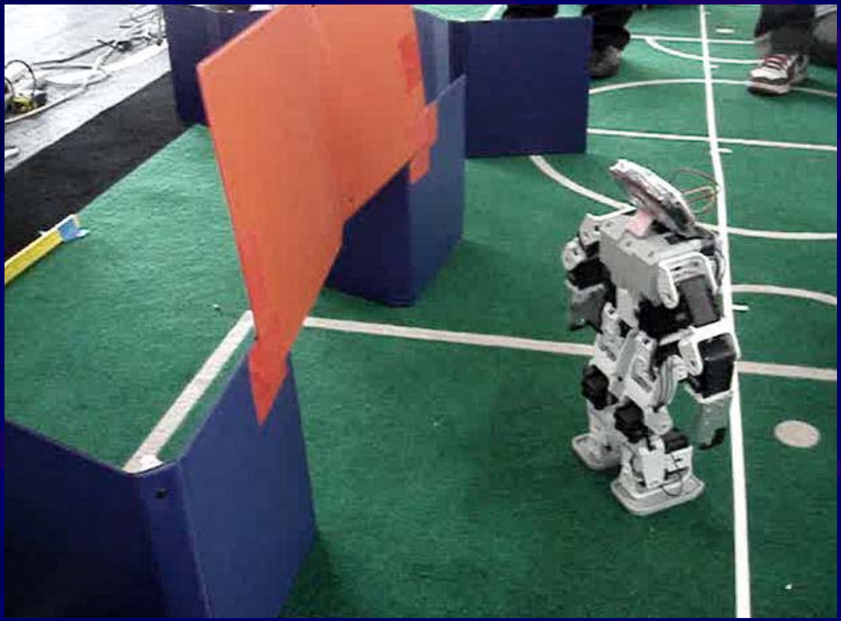
Challenges

- Lift and Carry: carry an increasing number of weights in a "backpack" over an uneven stepping field. The field is color coded to recognize heights (extremely advanced)



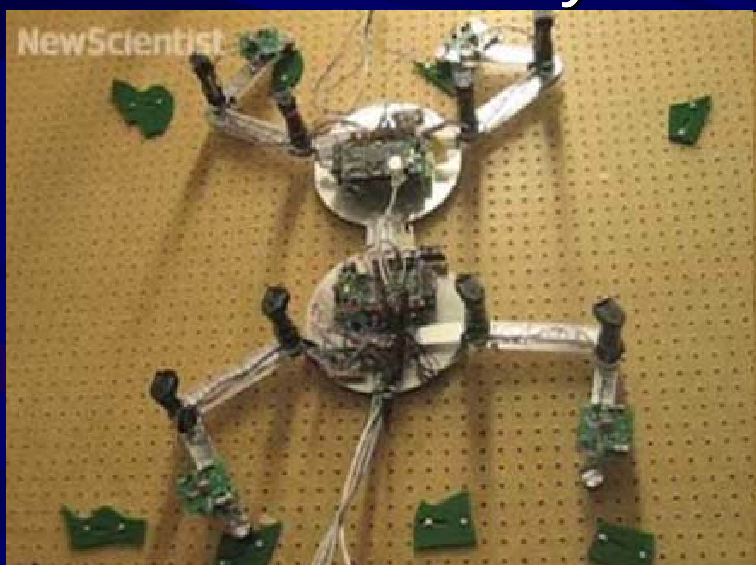
Challenges

- Obstacle Run: Move through a 3m long region without touching any of 3 types of obstacles: walls, holes, gates. Gates can be crawled through (complex motion planning)



Challenges

- Penalty Kick: one nod to soccer, involving strategy and dealing with another robot's actions. The ball is placed randomly.
- Climbing wall: new in 2009, climb a wall where foot and hand holds are placed randomly and must be visually recognized



In Combination

- These represent most elements of what we would expect from good, adaptive humanoid motion
- The relationship between them is important, e.g. fast start vs. long haul in sprint/marathon. Specialized equipment in one does not likely help the other
- Most evident in basketball, where teams rarely use any type of special motor for throwing, since increased weight would tend to decrease performance in WL/running
- Natural bias against many of the problems previously outlined with competitions/challenges

Lessons Learned

- Properly steering a competition can go a long way toward removing a focus on narrow, specialized solutions
- We have seen good things in HuroCup: e.g. in 2004 IR sensors were used by ~half the teams in the obstacle run to detect obstacles
- Undesirable from a human-like standpoint, but rather than disallowing them, we moved to gate/hold obstacles that were harder to take advantage of with IR
  - IR Disallowed after everyone quit using it prior to 2009

Summary

- We need to work toward better means of evaluation to advance work in AI
- Challenges and competitions have led us some of the way, but there are important weaknesses that we need to focus on to make these more useful
- Some of the control inherent in a mixed reality environment provides a nice bridge between simulation and the real world
- Competitions can be designed to avoid some of the pitfalls we see currently (follow the 2009 rule changes @ fira2009.org!)