Benchmarking Through Competitions

F. Amigoni, A. Bonarini, G. Fontana, M. Matteucci, V. Schiaffonati

***Abstract*—Competitions are a widely used and successful tool to promote scientific and technological progress in robotics. However, their usefulness for scientific research and for successful transfer of robotic technology from laboratories to industry may be increased by introducing a scientific approach to their design, and by structuring them as *benchmarking competitions*, i.e., in such a way that they can also be considered as benchmarks for objective performance assessment. This is the goal of RoCKIn, an FP7 project focusing on the development of robot benchmarking competitions.**

I. Competitions and Benchmarking

The Robotics community is running hundreds of competitions every year; among them we cite the DARPA challenges, RoboCup Soccer / Rescue / @Home / @Work competitions, the ICRA Robot Challenge, the Robot Cleaning Competition, and FIRST Competitions[1]. The outcome of these competitions is usually a ranking among participants, highlighting the best performers.

Individually and collectively, robot competitions have a very positive effect in encouraging participants to tackle challenging problems, thus promoting advances in robotics state of art. However, most robot competitions usually suffer from limitations when considered from a scientific and benchmarking perspective. For example, their results cannot be used as a benchmarking tool, which strongly reduces their potential impact as a mechanism to promote robotics to industry and limits their usefulness to research groups. Indeed, convincing companies to embrace technology originating in autonomous robotics to create new products and markets is, in fact, very difficult without established tools to assess the real-world performance of such technology and to compare different approaches. For these reasons, we believe that the times are ripe for a new way of designing robot competitions, to make them more scientifically grounded and, at the same time, more suitable for the role of objective benchmarks.

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement 601012.

[1]A list can be found at *http://robots.net/rcfaq.htm.*

The RoCKIn project[2] aims at providing tools for *benchmarking through competitions* to the robotics community. The aim of RoCKIn is to design and set up *scientific robot competitions* able to increase scientific and technological knowledge. This will require a strong attention to rigorous experimental methodologies, tempered by the need to retain the features that have made robot competitions into successful tools for the promotion of progress in robotics: first of all the "fun" and "challenge" elements that are key parts of any successful competition. RoCKIn will then exploit the scientific foundations of its robot competitions by endowing them – by design – with an additional role: that of *benchmarking tools*. In other terms, while the outcomes of the RoCKIn robot competitions will retain their traditional value of producing a ranking among competing solutions at competiton time, the experimental setting of the competition will also take on the more general significance of benchmarking procedures.

A. From Competitions to Scientific Competitions

To explore the use of competitions as benchmarking tools, it is necessary to adopt a scientific approach, investigating whether and how competitions can be treated as scientific experiments. Key differences, in fact, exist between these categories.

Competitions are designed to produce a ranking at a specific moment, while scientific experiments are aimed at proving some property in a way that, once established with a successful experiment, can be considered as assessed and can be used as the starting point for further research. Experiments have the key property of being repeatable, while competitions generally cannot be repeated under exactly the same conditions. An experiment should

[2]RoCKIn (*http://rockinrobotchallenge.eu/*) started in January 2013 under the 7th Framework Programme of the European Commission. Its partners are: Associacao Do Instituto Superior Tecnico Para A Investigacao E Desenvolvimento (Portugal, Coordinator); Università degli Studi di Roma La Sapienza (Italy); Hochschule Bonn-Rhein-Sieg (Germany); KUKA Laboratories GmbH KUKA (Germany); Politecnico di Milano (Italy); Security Challenge Limited (UK).

be reproducible, while the specifications for competitions are sometimes too vague to enable the exact reproduction of what happens there. Finally, competitions mainly push towards the development of viable solutions, while experiments are more focused on verifying hypothesis, sharing results, and increasing knowledge.

Even if current robot competitions cannot be considered as pure scientific experiments, the best features of competitions and scientific experiments could be successfully merged to kick off robotics innovation. This process requires careful handling of non-trivial methodological issues, but has the potential to create a new tool: the *scientific competition*, i.e., a competition where the tests faced by robots can be considered as scientific experiments. Nowadays, the growing attention of the robotic community towards an increased scientific rigour in experimental work and towards benchmarking in robotics [1] [2] [3] creates a favourable environment for this type of effort.

B. From Competitions to Benchmarking

Robot benchmarking can be defined as an objective performance evaluation of a robot system/subsystem under controlled, reproducible conditions. Benchmarking, though rife with practical difficulties, plays a fundamental role in robotics, since it enables objective comparisons among different systems in a common predefined setting, and promotes reproducibility and repeatability, thus ensuring a rigorous experimental approach. A benchmark includes a set of metrics together with a proper interpretation, allowing the evaluation of the performance of the system/subsystem under test according to well-specified objective criteria[3]. In particular, a benchmark can be used to certify properties and functionalities, and therefore takes a key role in demonstrating the worth of specific solutions to prospective adopters, be they companies contemplating the realization of new products, or their clients interested in the purchase of such products.

Competitions have a number of favourable features that would be useful to benchmarking. First of all, competitions are *appealing*: (some) people like to compete for personal or institutional affirmation. Competitions also take place with *regularity* and *precise timing*, and are good *showcases* of the current state of the art. Finally, competitions usually switch the focus from single subsystems towards *complete systems*, thus helping participants to take a broader view of robotics and highlighting the influence of subsystem integration on the overall system performance. In addition to these obvious benefits of competitions, there are additional advantages that are not so apparent. For instance, public competitions promote *critical analysis* of experimental work by taking it out of the lab, 'into the light". Another advantage of competitions is that they allow to split the cost and effort of setting up complex experimental installations among many participants. For the reasons stated above, designing and developing robot competitions that can also act as scientifically-sound tools for benchmarking is a worthy goal.

There are also disadvantages in using competitions as benchmarking tools. Most of these are related to the fact that – differently from what happens in the laboratory – a competition is a setting where time is strictly limited, and accessibility of the experimental setup to each individual participating team is subject to strong constraints. For instance, strict limitations usually apply to the number of times that a single experiment can be repeated during the contest. However, these limitations are usually perceived as an acceptable price to pay to obtain access to otherwise unapproachable types of experimental infrastructure and setting.

II. Benchmarking Competitions

Even if the potential advantages of competitions as tools to encourage innovation are clear, exploiting these advantages in the context of scientific research (i.e., designing *scientific competitions*) is not a straightforward process. Adding the requirement – specific to RoCKIn – that such scientific competitions must also act as benchmarking tools complicates the matter further, while bringing its own rewards as well. Hereafter, we will use the term *benchmarking competitions* to identify this type of scientific robot competition, i.e., the one that aim at being considered as scientific experiment, and that RoCKIn aims at setting up.

Support for benchmarking cannot be an afterthought: it must be designed into a competi-

[3]Please note that, while the metrics are required to be objective, the *data* that they take as their input can include subjective elements. This is necessary, for instance, when evaluating some of the issues associated to Human-Robot Interaction were human judgement is sometimes the only viable option.

tion right from the start. This is why RoCKIn will take inspiration from existing competitions, but it will define new ones instead of modifying the originals. More specifically, RoCKIn is dedicated to the design, setup, execution, and promotion of two RoCKIn benchmarking competitions: *RoCKIn@Home* and *RoCKIn@Work*. Their names are intentionally similar to those of the established RoboCup@Home and RoboCup@Work robot competitions, for three reasons: to underline the contribution that they have given to progress in robotics, to acknowledge their role as the initial inspiration for RoCKIn, and, finally, to point out that RoCKIn@Home and RoCKIn@Work are dedicated to similar scenarios. RoCKIn builds on the strenghts of RoboCup@Home and RoboCup@Work aiming at the broadening of their scope, both in terms of scientific validity, generality, and impact on the state the art. The stated goal of RoCKIn is to provide a way forward that starts from RoboCup@Home and RoboCup@Work towards a more capable, more scientifically sound, and more powerful combination of competition and benchmarking experiment.

A. Reproducibility and Repeatability

On methodological grounds, the work of RoCKIn will be founded on two concepts that are crucial to the scientific experimental method: those of *reproducibility* and *repeatability* [4]. Reproducibility is the possibility to verify, in an independent way, the results of a given experiment. Repeatability concerns the fact that a single result is not sufficient to ensure the success of an experiment. To guarantee that the result has not been achieved by chance, but is systematic, a successful experiment must be the outcome of a number of trials. Ensuring reproducibility and repeatability in the context of robot competitions is an extremely complex task (and possibly one that cannot be fully accomplished while retaining the attractive properties of competitions). For this reason, existing robot competitions tend to avoid tackling the issue at all. On the contrary, RoCKIn will openly face these problems.

B. Benchmarking Modules and Systems

One of the key limitations of available robot competitions and benchmarks is that they are focused either on *integrated systems* or on *specific modules*. For instance, RoboCup@Home and RoboCup@Work assess the performance of integrated robot systems executing specific tasks in domestic or factory environments, while the Rawseeds Benchmarking Toolkit [5] is dedicated to benchmarking software modules for self-localization, mapping, and SLAM. Unfortunately, focusing on only one of these two aspects (system or module) strongly limits the possibility to gain useful insight about the limitations and shortcomings of a robot. For this reason, one of the objectives for RoCKIn's benchmarking competitions it that of targeting both aspects and, crucially, to allow a deeper analysis of a robot by *combining system-level* and *module-level benchmarking*.

System-level and module-level tests do not investigate the same properties of a robot, and the insights they provide about system performance overlap only partially. Module-level benchmarking has the benefit of focusing only on the specific functionality that a module is devoted to, removing interferences due to the performance of other modules which are intrinsically connected at the task level. For instance, if the grasping performance of a mobile manipulator is tested by having it autonomously navigate to the grasping position, visually identify the item to be picked up, and finally grasp it, the efficacy of the grasping functionality is affected by the actual position where the navigation module stopped the robot and by the precision of the vision module in retrieving the pose and shape of the item. If, conversely, the grasping test is executed by locating the robot in a predefined position and by feeding it with predefined information about the item to be picked up, the final result will be almost exclusively due to the performance of the grasping module itself. By setting up the latter (module-level) test, the performance of the grasping module can be assessed accurately and with a high repeatability.

On the other hand, there are issues that module-level testing cannot assess, though they have a major impact on robot performance. For instance, the interactions among the navigation, vision, and grasping modules that were previously highlighted as disturbance factors in evaluating the performance of the grasping module take a crucial role in defining the real-world performance of a complete robot. Performing an experiment that excludes such interactions implies, therefore, a major loss of useful information. Here lies the specific worth of task-level robot testing: it is, in fact, the only way to make

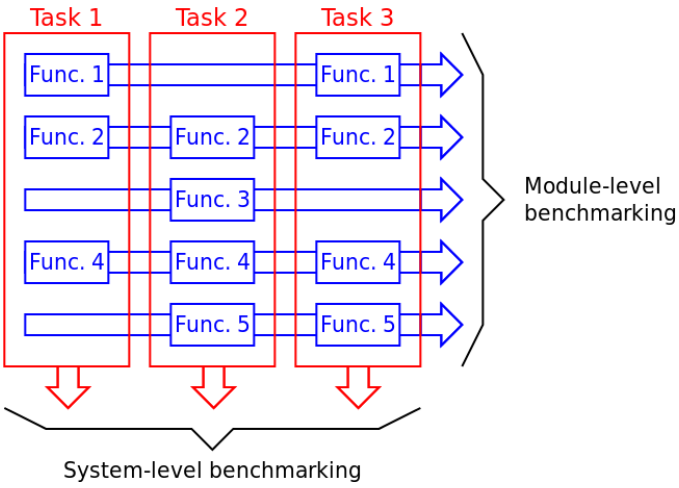

Figure 1. Module-level (i.e., functionality-level) and system-level (i.e., task-level) benchmarks. By jointly analysing their results, it is possible to acquire information about higher-level properties of the robot system, such as quality of system integration or interaction issues among modules.

system-level properties apparent. We already cited the most obvious of them (i.e., direct interactions among modules), but more subtle ones exist. One of the most important of these system-level properties on performance grounds, though it is very difficult to measure, is the quality of the *integration between modules*. Indeed, autonomous robots are systems of sufficiently high complexity to give emerging properties an important role in defining the overall performance of the integrated system.

For the above reasons, devising benchmarks capable of characterizing emerging system-level properties (such as integration issues) is, itself, a worthwhile goal. In our opinion, one path to reach this goal is that of applying to the same robot system(s) *both* system-level (i.e., task-level) benchmarks, and module-level (i.e., functionality-level) benchmarks, and then *jointly process their outputs* according to suitable evaluation criteria. This is the inspiration for the RoCKIn benchmarking competitions design.

Figure 1 describes, using an example competition comprising three tasks, how the two types of benchmarks explore different aspects of a single robot system. By associating to each task the set of functionalities that its execution requires, the two types of benchmarks can be referred to two directions: horizontal, for functionality-level/module-level benchmarks; vertical, for task-level/system-level benchmarks.

Functionality-level benchmarks should be able to investigate the performance of a specific module in a deeper and more general way with respect to task-level benchmarks. To achieve this, they should be aimed at testing (only) one functionality under a range of different conditions, within the chosen scenario(s). On the other hand, task-level banchmarks should evaluate whole-system functionality over a limited set of situations/tasks, taking into account all system modules as well as their interaction.

RoCKIn Competitions will not be the first to address both module-level and system-level aspects. For instance, the rules for RoboCup@Work 2012 defined two module-level tests (Basic Navigation and Basic Manipulation) and one task-level test (Basic Transportation). The contribution of RoCKIn will be related to two innovative elements: first, in RoCKIn the two types of competitions are strictly linked, both in their definition and (more importantly) in the processing of their outcomes; second, in RoCKIn competitions also act as benchmarks.

III. Conclusion

Over the years, robot competitions proved their worth as tools to explore, assess, demonstrate, and promote the state of the art in robotics. RoCKIn is a project dedicated to enhancing the scope and impact of robot competitions by designing and setting up (appealing) *benchmarking competitions* for robot systems, where tests rely on rigorous methodological foundations and outcomes assume benchmarking value, also valid outside the specific competition event where they are generated.

References

[1] M. Anderson, O. C. Jenkins, and S. Osentoski, "Recasting robotics challenges as experiments," *IEEE Robotics and Automation Magazine*, vol. 18, no. 2, pp. 10–11, Jun 2011.

[2] A. G. Cohn, R. Dechter, and G. Lakemeyer, "Editorial: The competition section: a new paper category," *Artif. Intell.*, vol. 175, no. 9-10, pp. iii–, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(11)00060-9

[3] F. Bonsignorio, A. Del Pobil, and J. Hallam, "Defining the requisites of a replicable robotics experiment," in *Workshop on Good Experimental Methodology in Robotics - RSS 2009*, 2009.

[4] F. Amigoni, M. Reggiani, and V. Schiaffonati, "An insightful comparison between experiments in mobile robotics and in science," *Autonomous Robots*, vol. 27, pp. 313–325.

[5] A. Bonarini, W. Burgard, G. Fontana, M. Matteucci, D. G. Sorrenti, and J. D. Tardos, "Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets," in *In proceedings of IROS'06 Workshop on Benchmarks in Robotics Research*, vol. On line, 2006. [Online]. Available: http://www.robot.uji.es/EURON/en/iros06.htm