Chapter 5

RoCKIn Benchmarking and Scoring System

Giulio Fontana, Matteo Matteucci,

Francesco Amigoni, Viola Schiaffonati,

Andrea Bonarini and Pedro U. Lima

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70013

Abstract

The main innovation brought forth by the European Project RoCKIn is the definition, implementation and application to an actual robot competition of the novel paradigm of *benchmarking through competitions*. By doing so, RoCKIn set in motion an evolutionary process to transform robot competitions from successful showcases with limited scientific impact into *benchmarking tools* for the consistent and objective evaluation of the performance of autonomous robot systems. Our work began by revisiting, in the light of the features and limitations of a competition setting, the very foundations of the scientific method; then we built on these by designing a novel type of competitions where the concepts of benchmark and objective performance metrics are the key points; finally, we arrived to the implementation of such concepts in the form of a real-world robot competition. This chapter describes the above process, explaining how each of its several aspects (theoretical, technical, procedural) has been tackled by RoCKIn. Special attention will be devoted to the problems of defining performance metrics and of capturing the *ground truth* needed to reliably assess robot perceptions and actions.

Keywords: benchmarking, robot competition, benchmarking through competitions, performance metrics, ground truth

1. Introduction

The main innovation brought forth by the European Project RoCKIn [1] is the definition, implementation and application, in the form of actual robot competitions, of the novel paradigm of **benchmarking through competitions** [2].

This paradigm is an evolution of the concept of robot competition. The idea is that—alongside the already established roles of demonstration towards the general public and networking event for researchers—competitions for autonomous robots can and should evolve to become benchmarking tools for robot systems. This is why the tests that robots are subjected to, during the RoCKIn competitions, are called *benchmarks*: their aim is in fact to act as reference tasks and activities in which robot performance is evaluated according to well-specified and quantitative metrics. The pioneering work of RoCKIn publicly demonstrated the feasibility of this approach, paving the way to further developments. One of such developments is the *European Robotics League* (*ERL*), an on-going robot competition set up by the European Project RockEU2 [3].

Robot competitions do not easily lend themselves to act as benchmarks, intended as rigorous evaluation procedures to assess the capabilities, reliability, dependability and performance of robot systems in precisely defined settings [4]. First, because the (often frantic) setting of a competition is badly suited to the execution of procedures that require accuracy and care. Second, because concepts that are crucial to experimentation and to benchmarking (such as *repeatability* and *replicability*, defined in Section 2) are difficult to reconcile with the necessary spectacular element of a public competition. As the final objective of RoCKIn was to obtain results that could be transferred into other competitions, such difficulties have been carefully taken into account, and viable solutions have been devised. Both the technical elements and the procedures required by such solutions had to avoid interfering with the execution of the competition.

A first instance of successful infusion of the RoCKIn legacy into other established robot competitions has occurred at the 2016 RoboCup competition held in Leipzig (Germany). In fact, at RoboCup 2016, the aforementioned ERL has been able to both collect benchmarking data from some of the existing RoboCup tests and to incorporate benchmarks directly based on such tests into the European Robotics League.

The first phase of RoCKIn's work consisted of going back to the foundations of the experimental method to carefully reassess the elements that characterize a *scientific experiment*. A further step has been to define the special case of the **benchmarking experiment**, that is, a comparison test which presents some of the features of a scientific experiment. Finally, this analysis formed the foundation for the design and execution of the benchmarks involved in the two challenges of the RoCKIn Competition (RoCKIn@Home and RoCKIn@Work). As explained in other chapters of this book, RoCKIn@Home focuses on a service robot scenario where a robot has to assist a person in her daily life, while RoCKIn@Work is aimed at the shop-floor scenario.

This chapter is dedicated to providing the reader with a summary of the steps composing the path that leads from the scientific foundations (Sections 2 and 3) to the implementation of the RoCKIn Competition, focusing on the methodologies (Sections 4 and 5) and the infrastructure (Section 6) used by RoCKIn to design and execute the Competition. Special attention is to be devoted to the solutions devised by RoCKIn to the problems of defining reliable performance metrics for robot activities, and of capturing the *ground truth* (GT) necessary to apply such metrics in an objective and consistent way.

2. Robot competitions as benchmarking tools

Project RoCKIn is based on the idea of **benchmarking through competitions**: that is, of transforming competitions into vessels for experiment-based scientific progress. The successful RoCKIn Competition demonstrated the feasibility of this innovative approach. Besides RoCKIn, the point of view that robotic competitions can (under suitable circumstances and despite some essential differences) be considered as experiments has also emerged elsewhere, both within the academic community and at the level of the European Commission. In particular, competitions are now considered as good vehicles for advancing the state-of-the-art in terms of new algorithms and techniques in the context of a common problem [5–8].

While scientific progress is often related to the concept of experiment, in the majority of cases significant differences exist between experiments and competitions [2]. Just to cite the most obvious, an experiment is aimed at evaluating objectively a specific hypothesis, while a competition is aimed at defining a ranking and winners; for this reason, competitions push towards the development of solutions, while experiments aim at exploring phenomena. Notwithstanding these and other differences, there are a number of reasons for recasting robot competitions as experiments, considering traditional experimental principles (comparison, repeatability, reproducibility, justification, etc.) as guidelines. *Comparison* is to know what has been already done in the field, to avoid the repetition of uninteresting experiments and to get hints on promising issues to tackle. *Reproducibility* is the possibility for independent scientists to verify the results of a given experiment by repeating it, while *repeatability* is the property of an experiment that yields the same outcome when performed at different times and/or in different places. *Justification* and *explanation* deal with the necessity of interpreting experimental data in order to derive correct implications.

Competitions usually provide controlled environments where approaches to solve specific problems can be compared. Furthermore, they require integrated implementations of complete robotic systems, suggesting a new experimental paradigm trying to complement the rigorous evaluation of specific modules in isolation (typical of most laboratory research). RoCKIn set out to prove that an experiment-oriented perspective on competitions can reach the aims of both research and demonstration, while providing a common ground for comparison of different solutions. By reframing robot competitions as experiments via the RoCKIn Competition, the project aimed at increasing their scientific rigour while trying to maintain their distinctive aspects, which are significant and valuable. For instance, competitions are appealing to the participants (people like to compete) and to the general public, in a way that laboratory experiments could never achieve. Competitions are excellent showcases of the current state-of-the-art in research and industry. Competitions push their participants to their creative limits, coordinating to solve difficult problems while doing better than their competitors, ultimately leading to the development and sharing of novel solutions. Competitions promote critical analysis of system performance out of labs. Finally, competitions are a way to share the cost and effort of setting up complex installations among a multitude of participants, making costly experimental setups feasible.

3. Benchmarking experiments

Although competitions can be considered as a way of comparing the performance of robots in a partially controlled environment, their character of being, to some degree, unique events, puts serious limits on the generalizability and replicability of their results. As it has been already noticed [9], robot competitions are not necessarily experimental procedures: on the contrary, some of their features may not fit an assessed experimental methodology. A competition can be considered as a kind of experiment only if its settings and scoring are properly defined.

To define what we intend for experiments, we turn to experimenting practice in computing, which can be intended as the empirical practice to gain and check knowledge about a computing system. It is worth noticing that in this context there are at least five different ways in which the notion of experiment is used [10]. These are ranked below, ordered by increasing complexity of execution and, more importantly, of general scientific significance of the results.

- *Feasibility experiment*. It is the loosest use of the term 'experiment' that can be found in many works reporting and describing new techniques and tools. Typically, the term 'experiment' is used in this case with the meaning of empirical demonstration, intended as the existence of proof of the ability to build a tool or a system.
- *Trial experiment*. This requires the evaluation of various aspects of a system using some predefined variables, which are often measured in laboratory, but can occur also in real contexts of use, possibly given some limitations.
- *Field experiment*. It is similar to trial experiment in its aim of evaluating the performance of a system against some measures, but it takes place outside the laboratory in complex socio-technical contexts of use. The system under investigation is thus tested in a live environment, and features such as performance, usability or robustness are measured.
- *Comparison experiment*. In this case, the term experiment refers to comparing different solutions with the goal of looking for the best solution for a specific problem. Typically, comparison is made in some setup and is based on some measures and criteria to assess the performance. Thus, alternative systems are compared and, to make this comparison as rigorous as possible, standard tests and publicly available data are introduced.
- *Controlled experiment*. It is the golden standard of experimentation of traditional scientific disciplines and refers to the original idea of experiment as controlled experience, where the activity of rigorously controlling (by implementing experimental principles such as reproducibility or repeatability) the factors that are under investigation is central, while eliminating the confounding factors, and allowing for generalization and prediction.

Many existing robot competitions are designed in such a way that their position within the above experimental hierarchy is not higher than field experiments. This cannot be considered as a flaw of such competitions, since they are usually not aimed at being recognized as scientific experiments. On the contrary, the aspiration of the RoCKIn Project has been to define a competition based on tests acting as **benchmarking experiments** [11].

In RoCKIn, benchmarking experiments are defined as *a way of performing experimental evaluation, of comparing different systems on a common, predefined, setting and of providing a set of metrics (together with a proper interpretation) to perform an objective evaluation, with the goal of enabling the reproducibility and repeatability of experiments*. The goal of RoCKIn is to devise benchmarking experiments that—according to the rank presented before—can be classified as comparison experiments or even, possibly, controlled experiments.

It is important to point out that the concept of 'objective evaluation' does not rule out human judgement of robot performance, which is often a key tool for performance evaluation (e.g. whenever human-robot interaction is involved). For RoCKIn benchmarks whose performance metrics include evaluation by humans, 'objective evaluation' means the setup of a suitable framework to ensure that human judgement is done according to clearly defined criteria, and that the elements of such judgement are separated and visible (instead of being lumped together in a single score). Several RoCKIn benchmarks (such as the one illustrated in **Figure 1**) involve human-robot interaction.

Additional information about how human judgement is managed in the context of RoCKIn benchmarks is available in Section 5.

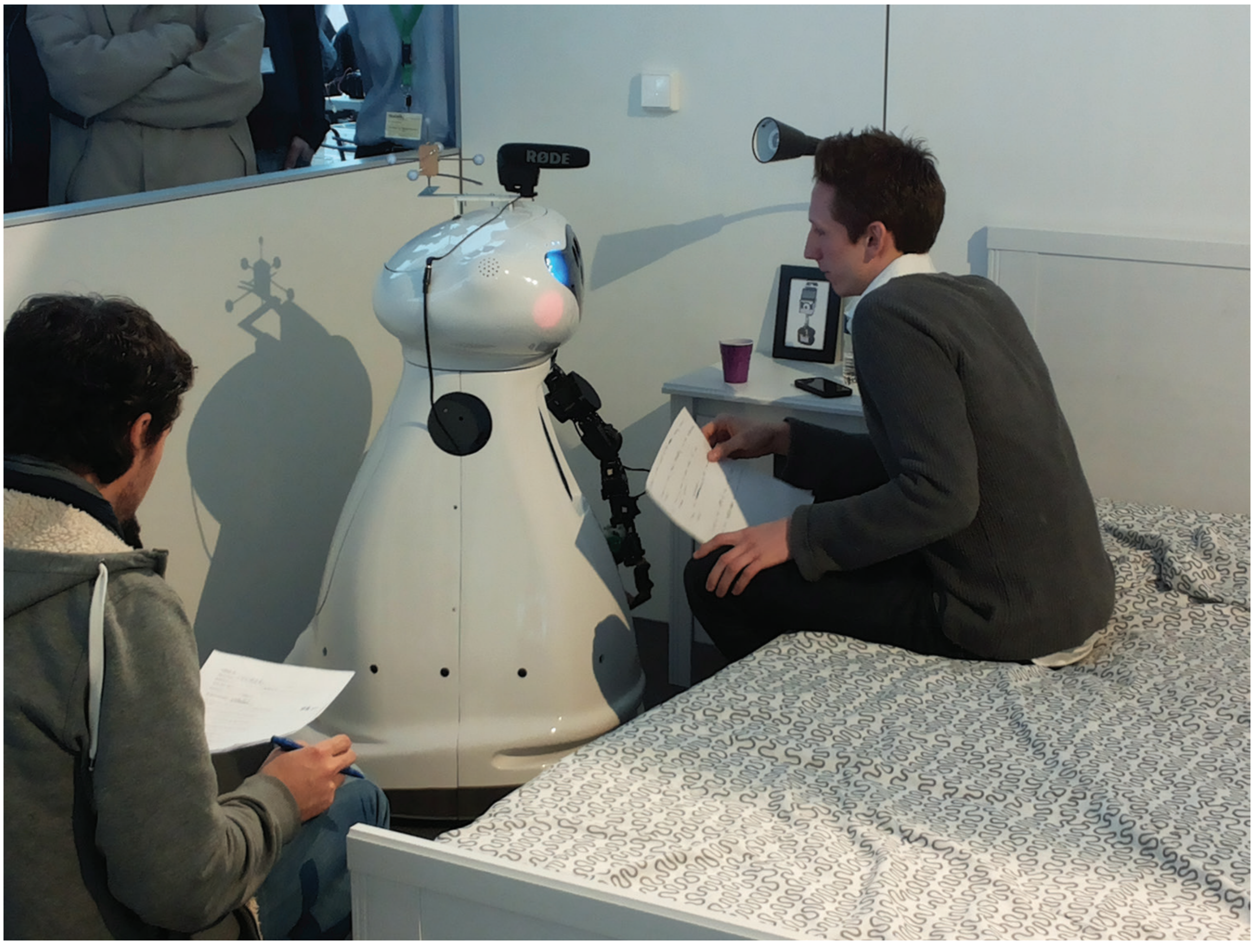


Figure 1. Example of RoCKIn benchmark requiring human robot interaction (at the RoCKIn 2015 Competition in Toulouse).

4. Benchmarking modules and systems

The approach of RoCKIn to benchmarking experiments (in the sense defined in Section 3) is based on the definition of two separate, but interconnected, types of benchmarks [2]:

- **Functionality benchmarks** (**FBM**s), which evaluate the performance of *robot modules* dedicated to specific functionalities, in the context of experiments focused on such functionalities.
- **Task benchmarks** (**TBM**s), which assess the performance of *integrated robot systems* facing complex tasks that require the interaction of different functionalities.

Of the two types, FBMs share more similarities with a scientific experiment. This is due to their stricter control on setting and execution. On the other side, this same feature of functionality benchmarks limits their capability to capture all the important aspects of the overall robot performance in a systemic way.

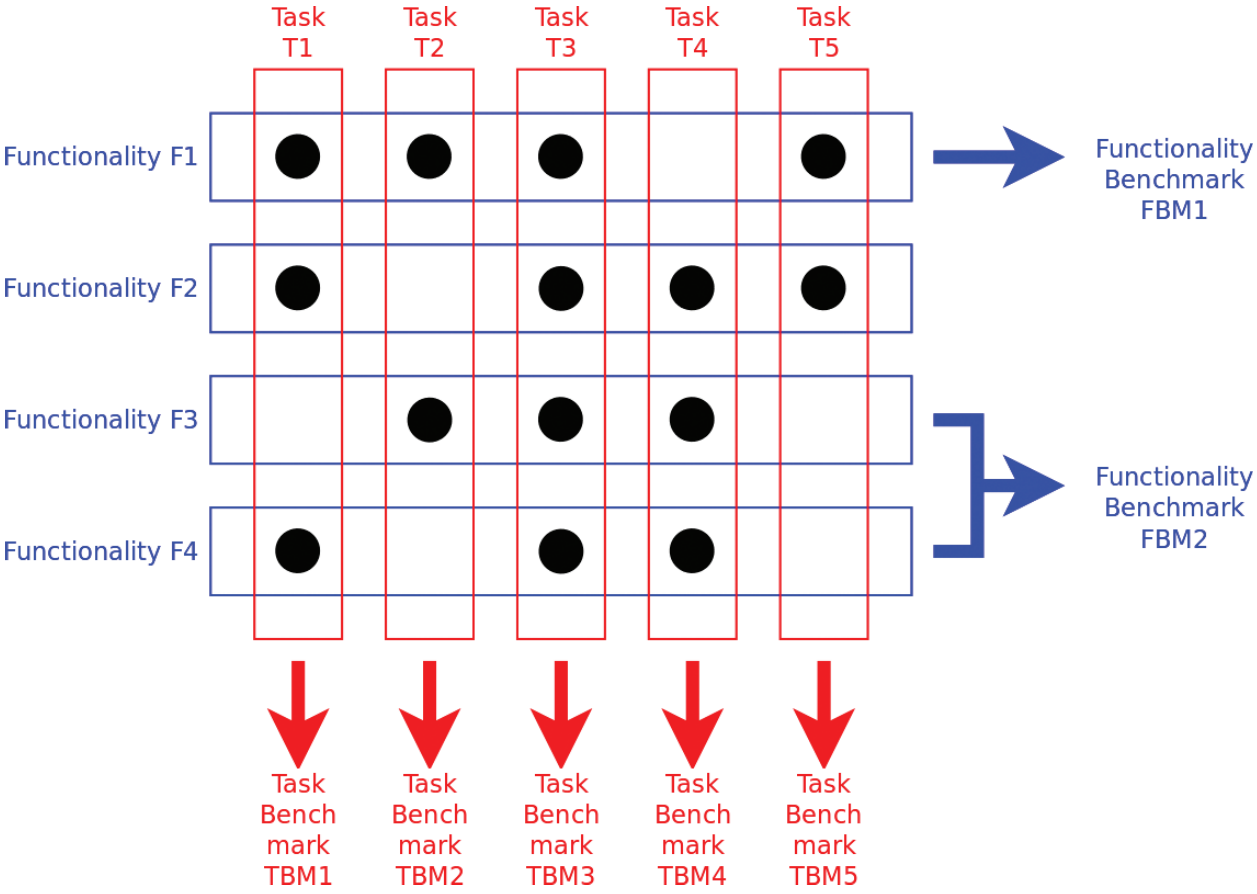
Focusing on either integrated systems or specific modules is a limit of traditional robot competitions and benchmarks. For instance, RoboCup@Home [12] and RoboCup@Work [13] assess the performance of integrated robot systems executing specific tasks in domestic or factory environments, while the Rawseeds Benchmarking Toolkit [14] is dedicated to assessing the performance of software modules that implement specific functionalities such as self-localization, mapping and SLAM (Simultaneous Localization And Mapping). Unfortunately, focusing only on one of these two approaches (system or module analysis) strongly limits the possibility to gain useful insight about the performance, limitations and shortcomings of a robot system. In particular, evaluating only the performance of integrated systems can identify the best performance for a given application, but it does not provide information about how the single modules are contributing to the global performance, and provides no information about where to spend further development effort in order to improve system performance. On the other side, the good performance of a module in isolation does not necessarily mean that it will perform well when inserted in an integrated system.

The RoCKIn Competition targets both aspects, thus enabling a deeper analysis of a robot system by combining system-level (TBM) and module-level (FBM) benchmarking [15]. Module-level testing has the benefit of focusing only on the specific functionality that a module is devoted to, removing interference due to the performance of other modules which interact with it at the system level. For instance, if the grasping performance of a mobile manipulator is tested by having it to autonomously navigate to the grasping position, visually identify the item to be picked up and finally grasp it, the effectiveness 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. On the other side, if the grasping test is executed by placing the robot in a predefined position and by configuring it with precise information about the item to be picked up, the final result will be almost exclusively due to the performance of the grasping module itself. The first test can be considered as a ‘system-level’ benchmark, because it involves more than one functionality of the robot; on the contrary, the second test can assess the performance of the grasping module with minimal interference from other modules and a high repeatability, and can thus be classified as ‘module-level’ benchmark.

It must be stressed that there are issues that module-level testing can neither identify nor assess, and nonetheless have a major impact on real-world robot performance. For instance, the interactions among the navigation, vision and grasping modules, which act as disturbance factors in evaluating the performance of the grasping module alone, take a crucial role in defining the actual performance of a robot system in a real setting where grasping is needed. Performing an experiment that excludes such interactions (such as a FBM focused on grasping) implies a major loss of useful information. Here lies the specific worth of system-level robot testing: it is the only way of making system-level properties apparent. We already cited the most obvious of such properties (i.e. direct interactions among modules), but subtler ones exist. One of the most important one is the quality of the integration between modules: experience shows that this is indeed crucial for the capability of a robot to achieve its goal.

Autonomous robots are systems of sufficiently high complexity to make loosely defined emerging properties (such as the aforementioned 'system integration') an important factor in the overall performance of the integrated system. As a consequence, even perfect knowledge of the performance of each robot module does not provide reliable, or sufficient, information to predict the performance of the complete robot once these modules are put together.

The considerations reported in this section can be represented in a matrix form, as shown in **Figure 2**.



	Task T1	Task T2	Task T3	Task T4	Task T5	
Functionality F1	●	●	●		●	Functionality Benchmark FBM1
Functionality F2	●		●	●	●	
Functionality F3		●	●	●		Functionality Benchmark FBM2
Functionality F4	●		●	●		
	Task Bench mark TBM1	Task Bench mark TBM2	Task Bench mark TBM3	Task Bench mark TBM4	Task Bench mark TBM5	

Figure 2. Benchmarking along the horizontal (functionality or module-level) and vertical (task or system-level) directions.

Let us consider an imaginary version of the RoCKIn Competition composed of five tasks (T1, T2, …, T5). **Figure 2** describes such competition as a matrix, showing the tasks as columns while the rows correspond to the functionalities for successfully executing the tasks. For the execution of the whole set of tasks, four different functionalities (F1, F2, …, F4) are required; however, a single task usually requires only a subset of these functionalities. In **Figure 2**, task Tx requires functionality Fy if a black dot is present at the crossing between column x and row y. For instance, task T2 does not require functionalities F2 and F4, while task T4 does not require functionality F1.

Two final observations conclude this section. First of all, while the robot tasks explored by the RoCKIn Competition correspond to the TBMs, the Competition does not necessarily include a FBM for each of the functionalities required by such TBMs. Second, it is conceivable that a functionality benchmark tests more than one functionality at the same time while still allowing to separate their contributions. In **Figure 2**, this happens to FBM2, which tests functionalities F2 and F4.

The reader is invited to compare theory with practice by consulting the descriptions of the specific functionality and task benchmarks composing the RoCKIn Competition, as provided by the Rulebooks of the RoCKIn Competition [16, 17].

5. Performance metrics

Performance metrics are an important element of the task benchmarks (TBMs) and functionality benchmarks (FBMs) presented in Section 4. In particular, their definition has a key role in enabling the benchmarks to act as *benchmarking experiments*, a concept defined in Section 3.

It is not possible to define useful general-purpose benchmark metrics: to be relevant, performance metrics need to be closely related to the specific robot activities under test. For some activities, it is reasonably easy to define suitably objective metrics: this mostly happens when the scope of the activity is limited and very well defined. In the context of RoCKIn, this mainly applies to FBMs. Other times, defining objective metrics is not easy. This happens especially when the activity that the robot is required to perform is complex, composed of different parts and with multiple objectives. For RoCKIn, this typically applies to TBMs.

As already pointed out in Section 3, an especially critical problem is that of defining performance metrics for robot activities that require human robot interaction (for instance, all the task benchmarks of RoCKIn@Home require HRI). In this case, subjective judgement by humans cannot be expunged from performance evaluation; on the contrary, it is crucial to it and must necessarily be part of the metrics. However, this introduces the necessity to establish a framework to guide and interpret subjective judgements, in order to limit their arbitrary elements and to enable their use as elements of consistent performance metrics. Defining such framework has been one of the tasks of RoCKIn; its results will be described in the following of this section. More precisely, Section 5.1 deals with the problem of collecting ground truth data, while Sections 5.2 and 5.3 explain how RoCKIn manages the problem of defining meaningful performance metrics, respectively, for task benchmarks and functionality benchmarks.

5.1. Ground truth

While discussing metrics, a key issue is that of **ground truth** (GT). In fact, in order to define objective performance metrics, it is necessary to get reliable data about the actual activity of the robot. Once available, these data can be compared (depending on the benchmark considered) either with the expected goals, or with robot perceptions. The accuracy of GT data should in any case be sufficiently high that any residual error is much lower than the accuracy required from the robot. In this way, with correctly designed performance metrics, errors in GT data have a negligible effect on the benchmark score.

Some types of GT data are suitable for collection by human referees: for instance, in FBM1 of RoCKIn@Home and RoCKIn@Work (object perception: the robot is required to identify an object presented to it and provide its pose), the actual identity of the object is ascertained by the referee. Other types of data, instead, can only be determined with sufficient precision using special machines: in FBM1, an example of such data is the pose of the object.

For robotics, the pose of an object (either a part of a robot such as the base or the end effector, or an external object such as those used for FBM1) is indeed an especially important type of ground truth. As a consequence, the capability of accurately measuring such data is a key enabler in the development of robot benchmarks. For this reason, RoCKIn collects pose data using a specialized machine [18], the main component of which is a *motion capture* (*mocap*) *system*. While RoCKIn does not specify the type of mocap system but only its performance, the current setup (which will be described in Section 6) is based on a commercial product. The mocap system uses IR-sensing cameras observing the volume of space where the objects to be tracked move, and special reconstruction software fed with the output of the cameras. The system is not capable of localizing objects on their own; instead, it localizes IR-reflecting *markers* affixed to the object. For the mocap system, a set composed of three or more rigidly connected markers can be used to define a *rigid body*, to which a 3-axes reference system is associated. When the system perceives a set of markers corresponding to a known rigid body, it computes and outputs the pose of the rigid body. This output pose, read by software developed for RoCKIn, is the ground truth used to compute the benchmark metrics. In the RoCKIn Competition, the rigid bodies tracked using the motion capture system are associated to **marker sets**, that is, special objects fitted with configurations of markers chosen to maximize tracking accuracy. Examples of RoCKIn marker sets are shown by **Figures 3** and **7**. During the execution of RoCKIn benchmarks that require pose measurements, marker sets are affixed to the objects to track and the mocap system used to track the associated rigid bodies. Tracking data from the mocap system are then used for online localization of the object. For instance, FBM2 of RoCKIn@Home is a navigation benchmark where the robot is required to reach a sequence of poses; for this benchmark, then, a marker set is affixed to the base of the robot in order to measure the differences between assigned and actual robot poses.

5.2. TBM metrics: achievements and penalties

The scoring framework for the evaluation of the task benchmarks in the RoCKIn Competition is based on the concept of **performance classes**. The performance class of a robot is determined by the number of **achievements** (or goals) that the robot collects during its execution of

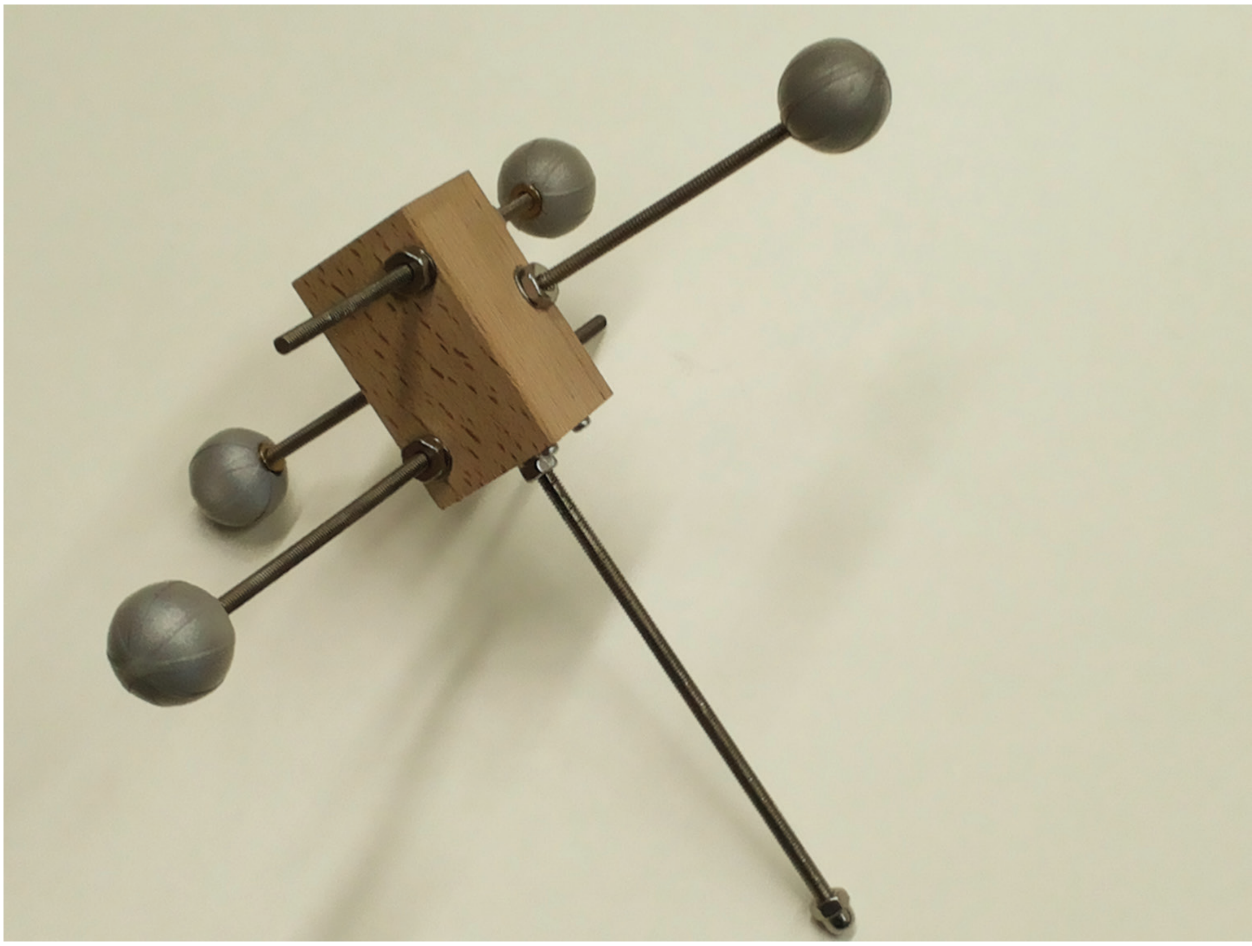


Figure 3. Marker set used for functionality benchmark 'Control' of RoCKIn@Work at the 2015 RoCKIn Competition.

the assigned task. Within each class, ranking is defined according to the number of **penalties** collected by the robots belonging to the class. Penalties are assigned to robots that, in the process of executing the assigned task, make one or more of the errors (which correspond generally to unwanted behaviours) defined by a list which is part of the specifications of the TBM.

More formally, in order to establish the ranking of the robots that execute a specific TBM, the elements of three sets have to be defined. While the contents of these sets are specific to the specific TBM considered, the general semantics is common to all TBMs. The three sets are:

- set **A** = **achievements or goals**, that is, things that the robot *is required to do*: during the execution of the benchmark, an **achievement** is assigned to the robot for each of these;
- set **PB** = **penalizing behaviours**, that is, things that the robot *is required to avoid doing*: during the execution of the benchmark, a **penalty** is assigned to the robot for each of these;
- set **DB** = **disqualifying behaviours**, that is, things that the robot *absolutely must not do*.

The content of each of the sets above must be specified as part of the specifications of the TBM. Then, the ranking of the robots that executed the same TBM is defined according to the following rules:

- The performance class of a robot corresponds to the number of achievements collected by the robot during the execution of the benchmark. Class 0 is the lowest performance class.
- A robot belonging to performance class *N* is considered as higher in rank than a robot belonging to performance class *M* whenever $M < N$.
- Among robots belonging to the same performance class, ranking is defined by the number of penalties collected by the robots: if robot R1 has less penalties than robot R2, then R1 is considered as higher in rank than R2.
- Among robots belonging to the same performance class and with the same number of penalties, the ranking of the robot which accomplished the task in a shorter time is considered as higher.

To apply the RoCKIn scoring framework for task benchmarks, the following three-step sorting algorithm is used:

1. if one or more of the disqualifying behaviours of set DB occurred during task execution, the robot gets disqualified, that is, it is assigned at performance class 0 and no further scoring procedures are performed for it;
2. the robot is assigned to performance class *X*, where *X* corresponds to the number of goals of set A accomplished by the robot (these sets do not contain repetitions, thus if a given achievement has to be accomplished multiple times, there will be as many distinctive instances of that achievement as required by the task; for instance, if the task requires to serve four guests during dinner, there will be four items in set A, one for each guest); and
3. a penalization is assigned to the robot for each behaviour belonging to set PB that occurred during the execution of the task. Unless clearly specified, it is sufficient that a penalized behaviour occurs once to assign a penalty, and further repetitions of the same behaviour do not lead to additional penalties.

One key property of this scoring system is that a robot that executes a larger part of the task associated to the TBM will always be placed into a higher performance class than a robot that executes a smaller part of the task. The measure of 'what part of the task' a robot accomplished is the subset of set A composed of the achievements assigned to the robot; the metric used to evaluate how large is the 'part of the tasks' accomplished by a robot is the number of elements of such subset, that is, the number of achievements assigned to the robot. Penalties do not change the performance class assigned to a robot and only influence intra-class ranking.

So far, for RoCKIn task benchmarks the assignment of achievements and penalties and the detection of disqualifying behaviours has been performed by human referees. This is an example of how human judgement, if correctly employed, can be part of a ranking procedure without compromising the objectivity of such procedure. In the case of RoCKIn's task benchmarks, the key to such objectivity lies in the precise definition of the elements of the aforementioned sets A, PB and DB, and in the training of the referees. Printed forms prepared with suitable boxes are provided to the referees, in order to guide their work and reduce the probability of mistakes.

It is possible that future benchmarks based on RoCKIn's framework (such as those developed by the on-going European Project RockEU2 [3]), or future implementations of existing benchmarks, will make use of methods different from human judgement to detect goals, penalized behaviours and/or disqualifying behaviours. This will not require any change to the scoring framework described in this section.

As a real-world example of TBM metrics, the remainder of this section describes one of the task benchmarks of the 2015 RoCKIn Competition (Lisbon, Portugal). Interested readers can find a complete description of the benchmark (including much more detail) in the RoCKIn@ Home Rulebook [16].

5.2.1. Example: task benchmark 'Welcoming Visitors'

A person takes the role of *Granny Annie*, a fictional character corresponding to an elderly woman. Granny Annie is helped in her daily activities by her service robot (i.e. the robot under test). In this TBM the robot is required to handle several visitors who arrive at Annie's home and ring the doorbell. The robot has to treat each visitor appropriately, according to the following scenarios:

- *Dr Kimble* is Annie's doctor stopping by to see after her. He is a known acquaintance; the robot lets him in and guides him to the bedroom.
- The *Deli Man* delivers the breakfast; the actual person is changing almost daily, but they all have a Deli Man uniform. The robot guides the Deli Man to the kitchen, and then guides him out again. The robot is supposed to constantly observe the visitor.
- The *Postman* rings the doorbell and delivers mail and a parcel; the actual person is changing almost daily, but they all have a Postman uniform. The robot just receives the deliveries and bids farewell to him.
- An *unknown person*, trying to sell magazine subscriptions, is ringing. The robot will tell him goodbye without letting the person in.

The robot must recognize the visitor by comparing the images from a camera located outside the door to known faces and/or uniforms. Interaction between people and robot is done vocally. Performance evaluation for this TBM is done as follows.

The set A of the achievements is composed by the following elements:

- The robot opens the door when the doorbell is rung by Dr Kimble and correctly identifies him.
- The robot opens the door when the doorbell is rung by the Deli Man and correctly identifies him.
- The robot opens the door when the doorbell is rung by the Postman and correctly identifies him.
- The robot opens the door when the doorbell is rung by an unknown person and correctly identifies the person as such.

- The robot exhibits the expected behaviour for interacting with Dr Kimble.
- The robot exhibits the expected behaviour for interacting with the Deli Man.
- The robot exhibits the expected behaviour for interacting with the Postman.
- The robot exhibits the expected behaviour for interacting with an unknown person.

The set PB of *penalized behaviours* is composed by the following elements:

- The robot fails in making the visitor respect the proper rights.
- The robot generates false alarms.
- The robot fails in maintaining the original state of the environment.
- The robot requires extra repetitions of speech.
- The robot bumps into the furniture.
- The robot stops working.

Finally, the set DB of *disqualifying behaviours* is composed by the following elements:

- The robot hits Annie or one of the visitors.
- The robot damages the testbed.

5.3. FBM metrics: benchmark-specific measurements

As explained in Section 2, among the RoCKIn benchmarks, FBMs are those that can be more easily designed to act as *benchmarking experiments* (i.e. *a way of performing experimental evaluation, of comparing different systems on a common, predefined, setting, and of providing a set of metrics—together with a proper interpretation—to perform an objective evaluation, with the goal of enabling the reproducibility and repeatability of experiments*). The reason for this is that FBMs are focused on one (or a very small subset) of the functionalities of a robot, which allows a much more precise definition of the activity that the robot is required to perform with respect to what happens in TBMs.

An important consequence of focusing the benchmarking action towards specific functionalities is that it sometimes enables the benchmark designer to completely eschew evaluation by human referees, thus making the definition of objective metrics easier. As observed in Section 5.2, devising objective performance metrics which include human evaluation is possible, but requires special care. On the other hand, a performance metric based on a well-specified algorithm applied to instrumental measurements of physical quantities is objective by definition. Of course, an objective metric—if badly designed—can nonetheless be a bad indicator of robot performance: however, this is a problem common to any metric.

An example of the 'objective by design' performance metrics described above are those used by RoCKIn's FBMs assessing the physical movements of the robot (or parts of it) through space. These metrics are based on comparisons (according to specific criteria) of the expected motion of the robot (or robot part) with the ground truth pose data produced by the motion capture system introduced in Section 5. Section 6 will show how such system is set up and used in practice.

A consequence of the very specificity of the functionality benchmarks is that it is impossible to define a general scoring framework for FBMs. In fact, the close link between FBMs and a single functionality requires that performance metrics are based on the features of such functionality. For this reason, a general methodology suitable for all FBMs cannot be defined. This differs markedly from what has been done for task benchmarks in Section 5.2, where a common framework for the performance metrics for TBMs, based on the concept of *performance classes*, was presented.

As a real-world example of FBM metrics, the following of this section describes one of the functionality benchmarks used in the 2015 RoCKIn Competition (Lisbon, Portugal). Interested readers can find a complete description of the benchmark (including much more detail) in the RoCKIn@Work Rulebook [17].

5.3.1. Example: functionality benchmark 'Control'

This functionality benchmark assesses the capability of a robot to control the manipulator's (and the mobile platform's) motion in a continuous manner. The robot has to follow a given path in the Cartesian space using the tip of a *marker set*, that is, a special object (shown in **Figure 3**) which can be precisely localized in space using RoCKIn's motion capture system.

More precisely, the mocap system is used to measure the deviation between the assigned path and the path actually followed by the tip of the marker set due to the movements of the robot's end effector. In the 2015 RoCKIn Competition, the given path could be a segment of a straight line or a portion of a sine function.

Without going into the procedural details of the benchmark, we focus here on the accuracy metric used to assess control performance. Let us define:

- $r(l) = (x_r(l), y_r(l))$ the parametric representation of the actual robot path,
- $t(l) = (x_t(l), y_t(l))$ the parametric representation of the given (target) path,

where l is a parameter ranging from 0 to 1. Then, the accuracy metric is

$$\frac{1}{N}\sum_{l\in L_{sampled}} d(r(l\,)\,,\,t(l\,)\,) \qquad (1)$$

where $L_{sampled}$ is a subset of the set L_{gt} of values of l in correspondence to which is available a location measurement from the ground truth system, $N = |L_{sampled}|$ and $d()$ represents the Euclidean distance between two points.

As anticipated, for this FBM the process of collecting the necessary data and computing the accuracy metric is entirely performed by machines; no human intervention is required.

5.4. RoCKIn benchmarking system

The RoCKIn benchmarking system is the infrastructure supporting the activities of the RoCKIn Competition that are directly related to benchmarking. The setup described in this section corresponds to the one used by at the 2015 RoCKIn Competition held in Lisbon, Portugal.

The system is composed of two interconnected but separate subsystems: one dedicated to the RoCKIn@Home benchmarks, and the other to the RoCKIn@Work benchmarks. This is due to the fact that in Lisbon the benchmarks of these two challenges were running in parallel due to the time constraints of the competition. In a less demanding setting, it would be possible to lower the number of components of the RoCKIn benchmarking system by relaxing the constraint of being capable of managing one RoCKIn@Home benchmark and one RoCKIn@Work benchmark at the same time.

6. System architecture

The architecture of the RoCKin benchmarking system is shown in **Figure 4**.

As shown in **Figure 4**, for each of the two challenges (RoCKIn@Home and RoCKIn@Work) the system includes three main computers. These are:

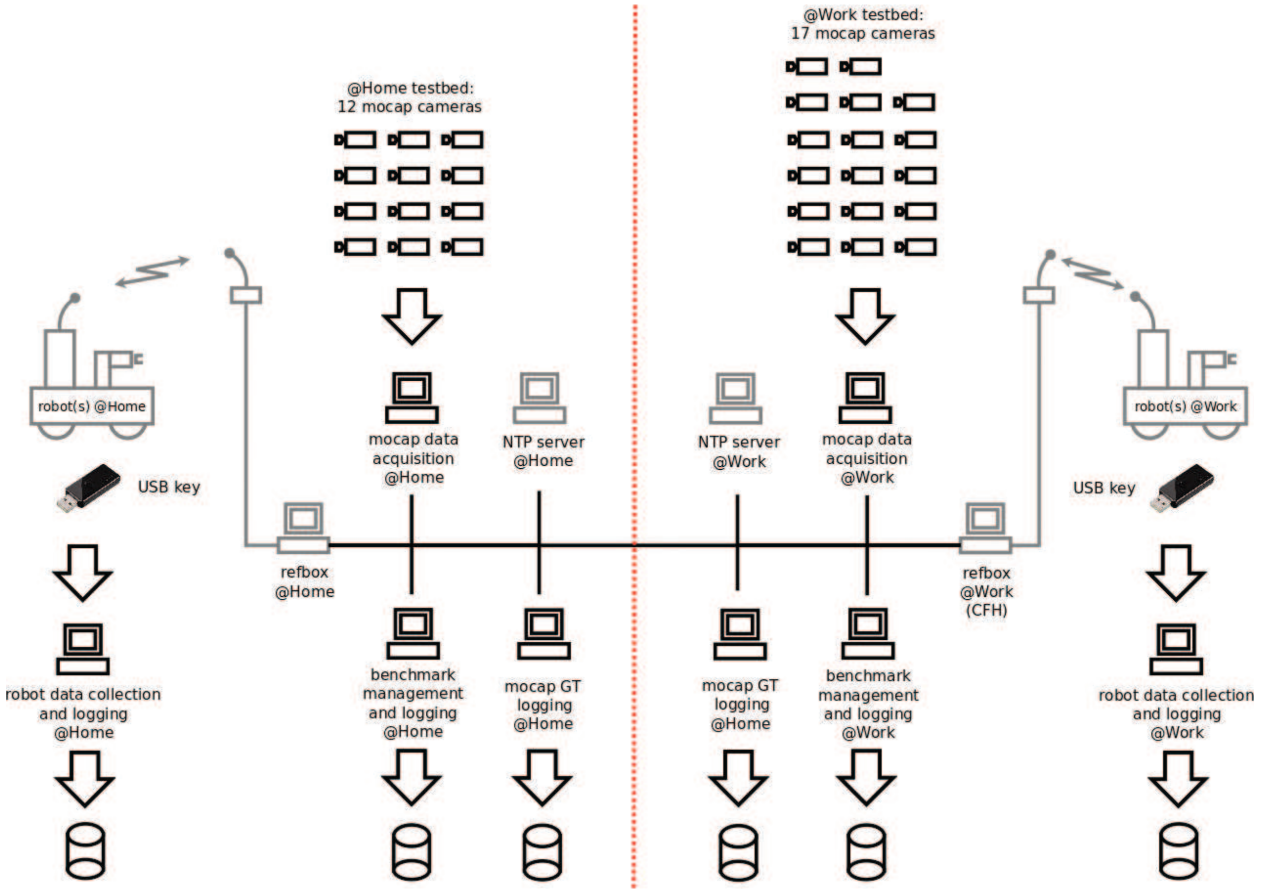

@Home testbed:
12 mocap cameras
@Work testbed:
17 mocap cameras
robot(s) @Home
USB key
mocap data acquisition @Home
NTP server @Home
NTP server @Work
mocap data acquisition @Work
USB key
robot(s) @Work
refbox @Home
refbox @Work (CFH)
robot data collection and logging @Home
benchmark management and logging @Home
mocap GT logging @Home
mocap GT logging @Home
benchmark management and logging @Work
robot data collection and logging @Work

Figure 4. Architecture of the RoCKIn benchmarking system used at the 2015 RoCKIn Competition held in Lisbon, Portugal.

1. one computer acquiring motion capture data from the special cameras of the mocap system and streaming it to the other machines;
2. one computer processing the mocap data streamed by the former to extract and log ground truth (GT) pose data; and
3. one computer managing the benchmarks (which also logs the data related to their execution).

One additional computer is used to collect and save robot-generated data, logged by the robots on USB keys during the execution of the benchmark. These USB keys are physically brought by the teams to the referees immediately after each benchmark.

In the end, the number of machines involved and the complexity of their interconnections is fairly high. This is due to several factors, including the fact that the software of the mocap system requires the Windows operating system, while all other machines are Linux-based, and the fact that the various subsystems have been developed (and physically brought to the Competition) by different partners of project RoCKIn. **Figure 5** shows the 'benchmarking table' hosting part of the PCs used for benchmarking for RoCKIn@Home at the 2015 RoCKIn Competition. A similar working area was used for RoCKIn@Work.

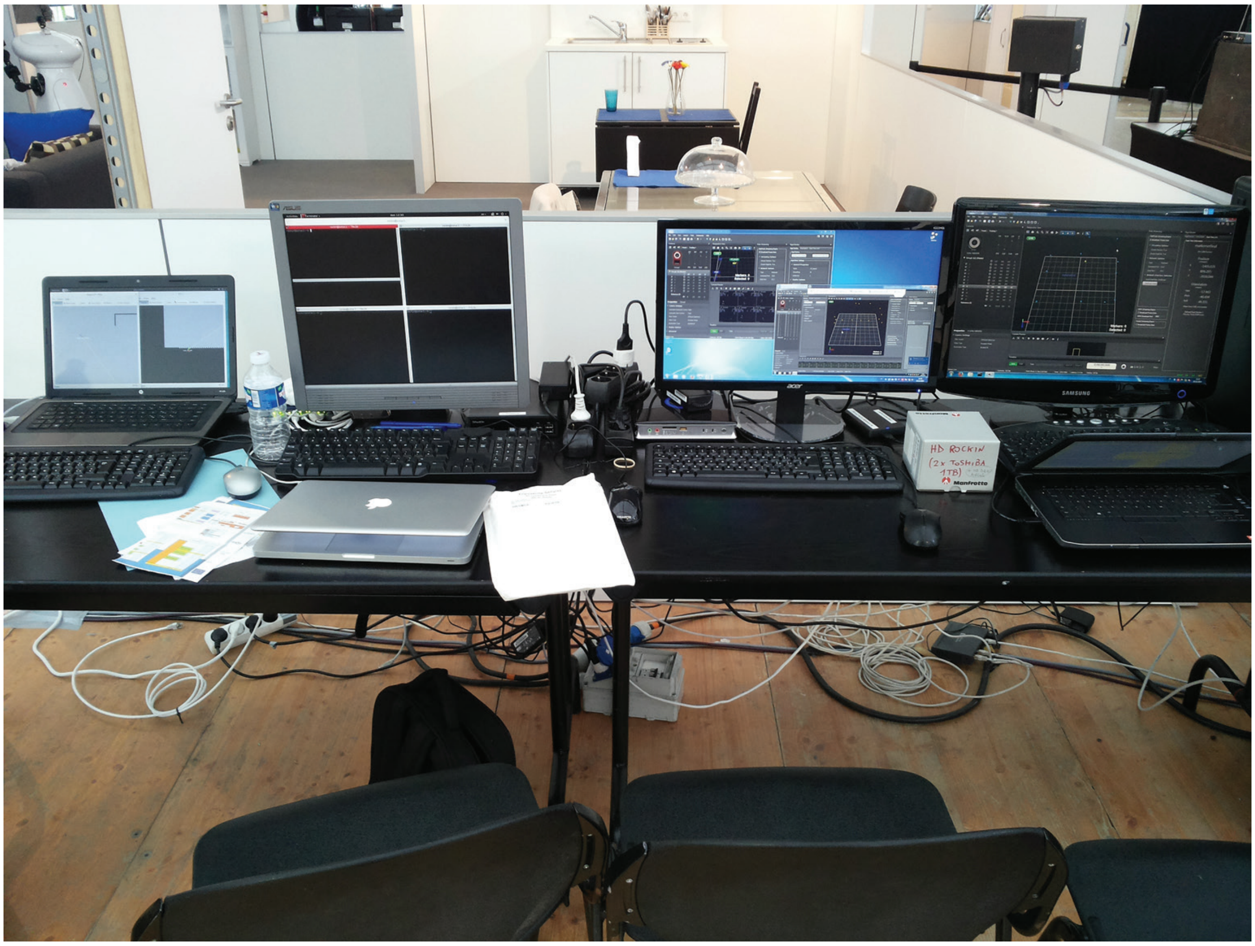


Figure 5. PCs used for RoCKIn@Home benchmarks at the 2015 RoCKIn Competition.

To operate, the RoCKIn benchmarking system also needs to interact with external systems, shown in grey in **Figure 4**. Interactions occur over TCP/IP networks, which include wireless segments. The systems external to the RoCKIn benchmarking system shown (in grey) in **Figure 4** are:

1. the robot under test;
2. the Referee Box (also called Central Factory Hub or CFH in RoCKIn@Work), which organizes competition activities, interfaces with devices belonging to the testbed and interacts with human referees; and
3. the NTP (Network Time Protocol) server with which all the PCs in **Figure 4** (including those on board of the robot) have to synchronize.

Synchronization is important for the correct execution of the RoCKIn benchmarks, for two reasons. First, because benchmark execution requires to associate and compare data generated by different sources, which can only be done if such data is correctly time-stamped. Secondly, because RoCKIn records datasets comprising both robot-generated data (e.g. sensor streams) and data generated externally to the robot (e.g. ground truth): for the datasets to be usable, all such data streams must therefore share the same time base.

A consequence of the synchronization constraints described above is that, in order to execute one of RoCKIn's benchmarks, a robot must precisely align its own internal clock to the clock of the RoCKIn NTP server. The Referee Box checks for misalignments and only starts the benchmark when these have been reduced below a predefined threshold. To help participating teams to perform such adjustment automatically, RoCKIn recommended installation on the robots of a software package called *Chrony* and provided a suitable configuration file for it.

6.1. Motion capture setup

To be able to benchmark actual robot performance, RoCKIn needs to collect ground truth data. For RoCKIn an especially important category of GT data is that describing the pose of objects and robots in space. As anticipated in Section 5.1, RoCKIn captures such data using a custom hardware and software system based on a commercial motion capture (mocap) system. The mocap system used at the 2015 Competition is called OptiTrack and is manufactured by Natural Point. OptiTrack relies on special infrared 'smart' cameras and proprietary software (running partly on the cameras and partly on a PC) to detect the location of IR-reflective markers. **Figure 6** shows an example of how RoCKIn used such cameras for its activities.

A set of at least three markers having fixed distance between each other can be defined as a *rigid body* in the OptiTrack mocap system. The system can then track the 6DOF pose of all defined rigid bodies and stream the data, which are subsequently collected by special software running on one of the machines in **Figure 4** (*mocap GT logging*). RoCKIn benchmarks make use of this to track the pose of special objects called *marker sets*. When the marker set is rigidly affixed to a robot component, it is possible to reconstruct the pose of the component from tracking data.

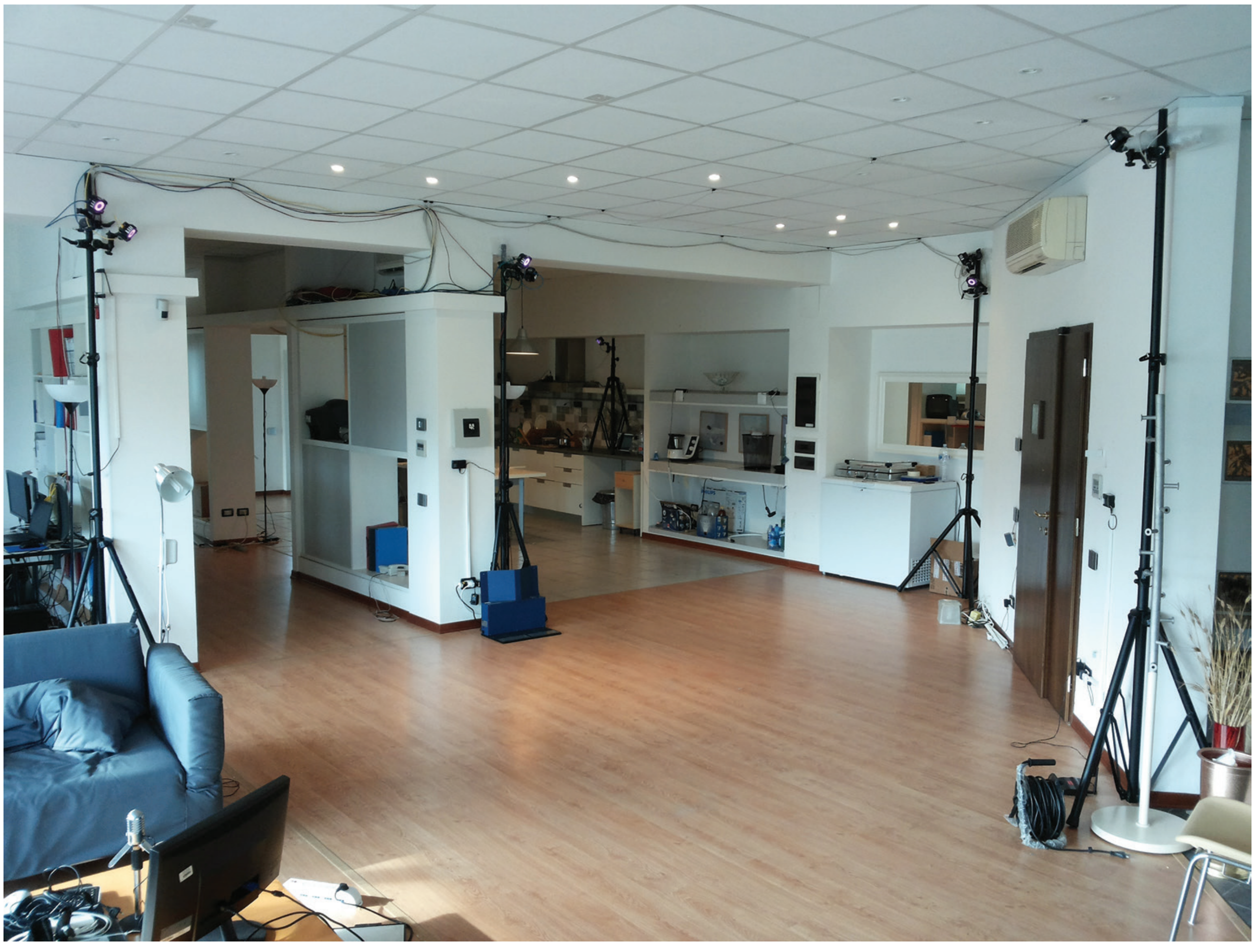


Figure 6. Part of the motion capture cameras used to cover the RoCKIn@Home area at the RoCKIn Camp held in Peccioli (Italy) in 2015.

For instance, to track robots, RoCKIn uses marker sets composed of a planar base (made of 4-mm-thick plywood) fitted with five spherical markers, shown in **Figure 7**.

The locations of the markers of the marker set maximize the distances between markers while keeping the marker set reasonably compact. Most importantly, such locations have been carefully chosen to ensure that inter-marker distances are all significantly different. This is necessary to prevent ambiguity, which may cause severe fluctuations in reconstructed pose. The marker set of **Figure 7** is used, for instance, during the execution of the functional benchmark 'Navigation' of RoCKIn@Home. This FBM (already presented in Section 5.1) requires that the robot navigates through the environment to reach, in order, a series of waypoints specified in terms of position and heading.

Practical experience at RoCKIn events (Camps and Competitions) showed that, unfortunately, obtaining a good setup of the motion capture system requires significant experience. Especially critical are the choice of camera locations and the tuning of the system for optimum performance, also keeping in mind the effect of local lighting. These aspects become less and less critical as the number of cameras increase; however, given the considerable cost of each camera, RoCKIn tried to keep their number as low as possible (though, as shown in **Figure 4**, still not very low in absolute terms; this high number is a direct consequence of the large observed

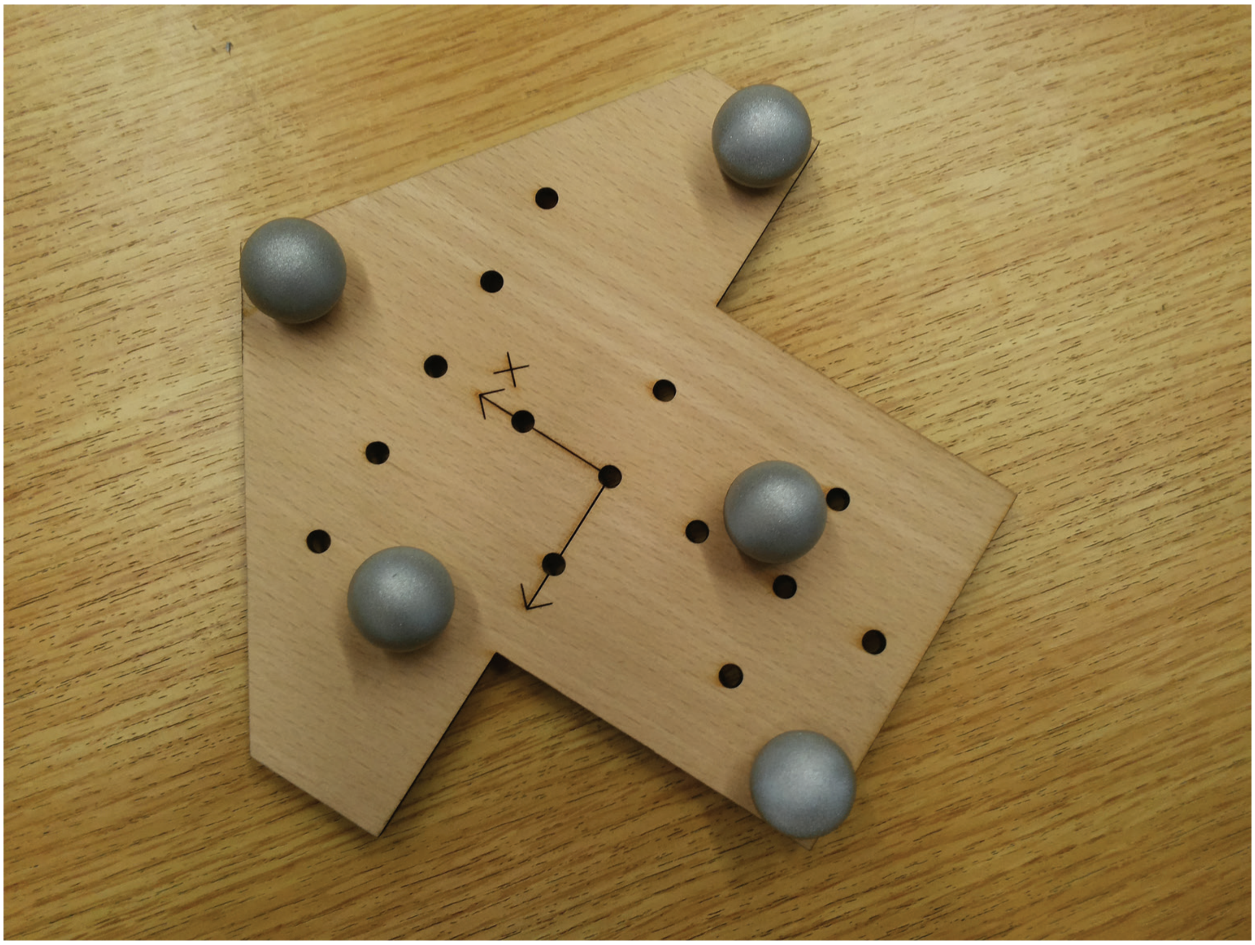


Figure 7. Marker set used to track robots at the 2015 RoCKIn Competition. To get an idea of the its dimensions, the reader can consider that each of the five spherical markers has a diameter of 19 mm, and that the base of the marker set fits within a circle with a diameter of 170 mm.

volumes). In the end, RoCKIn always operated close to the edge of the performance envelope of the OptiTrack mocap system, thus minimizing the cost of the system but paying a price in terms of difficulty of setup and expertise required for successful installation and parameter setting.

Another difficulty of using a motion capture system in a temporary setting (such as a robot competition) is that it is difficult to ensure the required consistency over time of relative camera locations. Even small changes in these locations can, in fact, greatly affect the performance of the mocap system. For this reason, for instance, mobile installations such as the tripod-based one shown in **Figure 6** are not acceptable for competitions. The solution chosen by RoCKIn makes use of an overhead truss, which of course is much heavier, larger and more difficult to mount and dismantle. **Figure 8** shows a rendering of the truss mounted above and around the RoCKIn@Home testbed at the 2015 competition; a similar truss was used for the RoCKIn@Work area.

6.2. Motion capture usage

At the 2015 RoCKIn Competition, each participating team was required to mount a marker set of the type shown in **Figure 7** on the top of their robot (fitting the marker sets on top minimizes

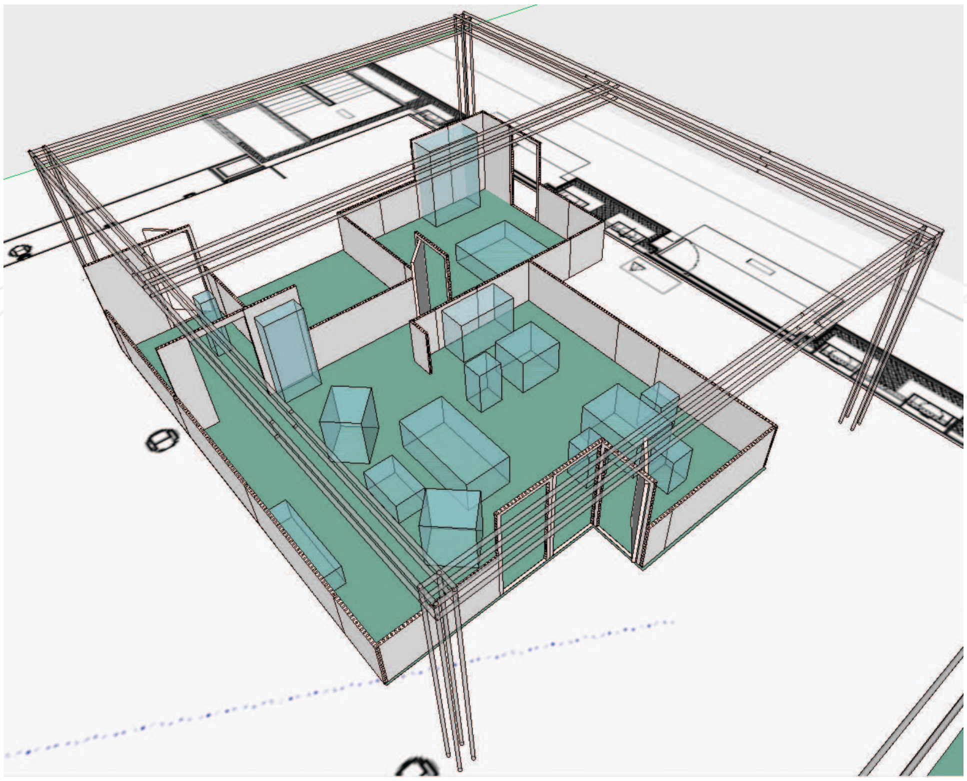


Figure 8. Rendering of the overhead truss used to support motion capture cameras around and above the RoCKIn@ Home testbed of the 2015 Competition.

occlusions due to robot parts), in a roughly horizontal orientation (thus minimizing occlusions between markers, as explained later), with the arrow-shaped marker base pointing forward according to the robot's own odometry reference frame (the shape of the marker set has been chosen to make pointing obvious). To facilitate mounting, marker sets are provided with holes, and CAD models of the marker set base are available.

RoCKIn acquired the transform between the odometry reference frame of each participating robot and the reference frame of the marker set mounted on it. Such transforms have been used, during the execution of the benchmarks, to reconstruct robot pose (according to the robot's own coordinate system) from motion capture data. In this way, ground truth data produced were directly comparable with odometry data logged by the robot, thus facilitating the assessment of the robot's odometric performance. The procedure to acquire the transforms required each team, in turn, to place their robots on the ground in a predefined location, with the X, Y axes of the robot's odometry frame aligned in a predefined way.

Beyond the ones mounted on robots, additional marker sets are used as parts of the setup for specific benchmarks. One of these special purpose marker sets, used for the 'Control' FBM of RoCKIn@Work, is shown in **Figure 3**. Other specialized marker sets are used for the 'Object perception' FBMs of RoCKIn@Home and RoCKIn@Work. This benchmark, already presented in Section 5.1, requires that the robot identifies and localizes a series of objects placed in front of it. **Figure 9** illustrates the elements of the experimental setup for FBM1, while **Figure 10** shows their use during the execution of the benchmark.

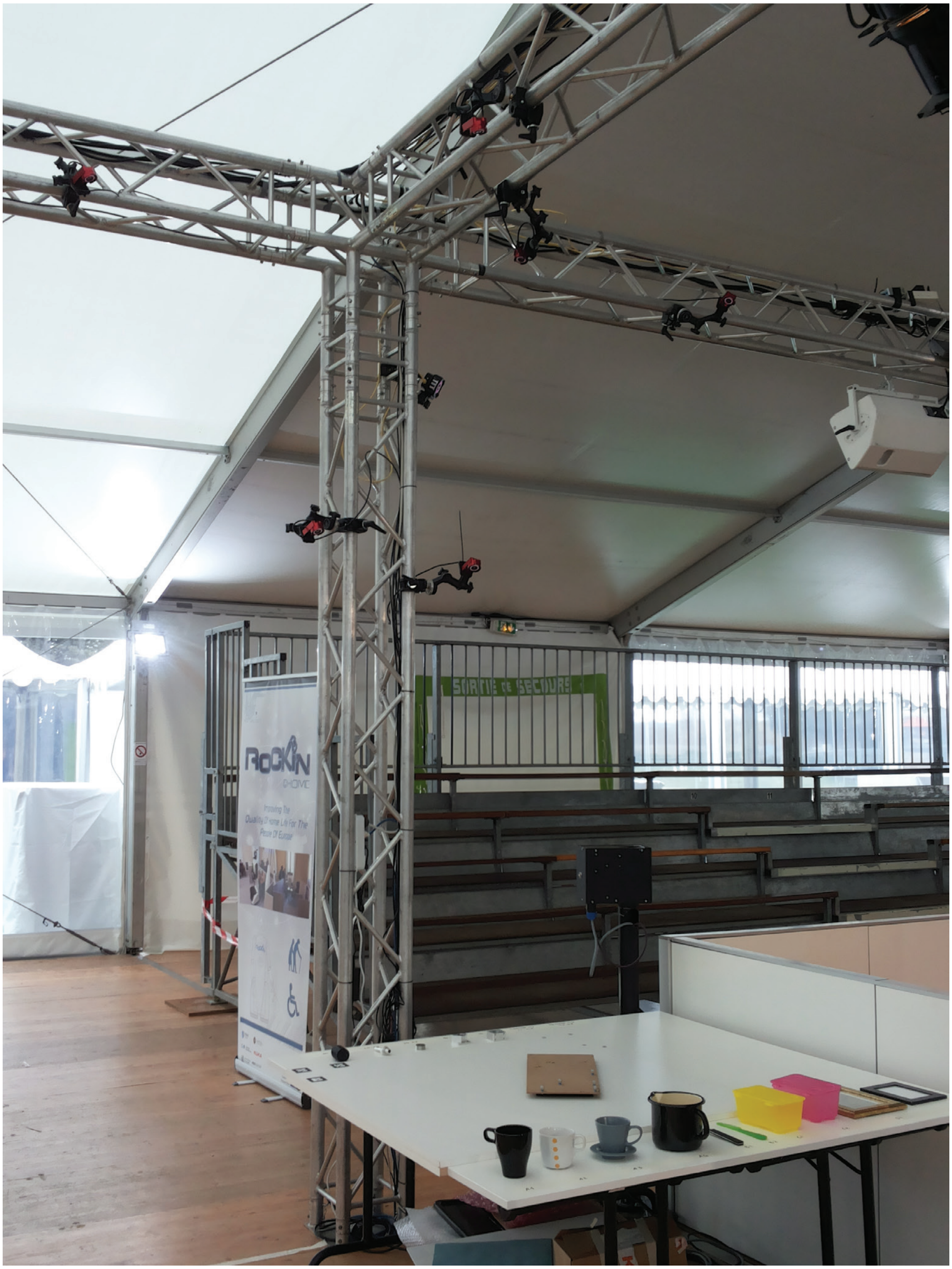


Figure 9. Setup for functional benchmark 1 (Object Perception) at the 2014 RoCKIn Competition in Toulouse, France. The (red) motion capture cameras used to track the objects presented to the robot are mounted on the metal truss adjacent to the table.

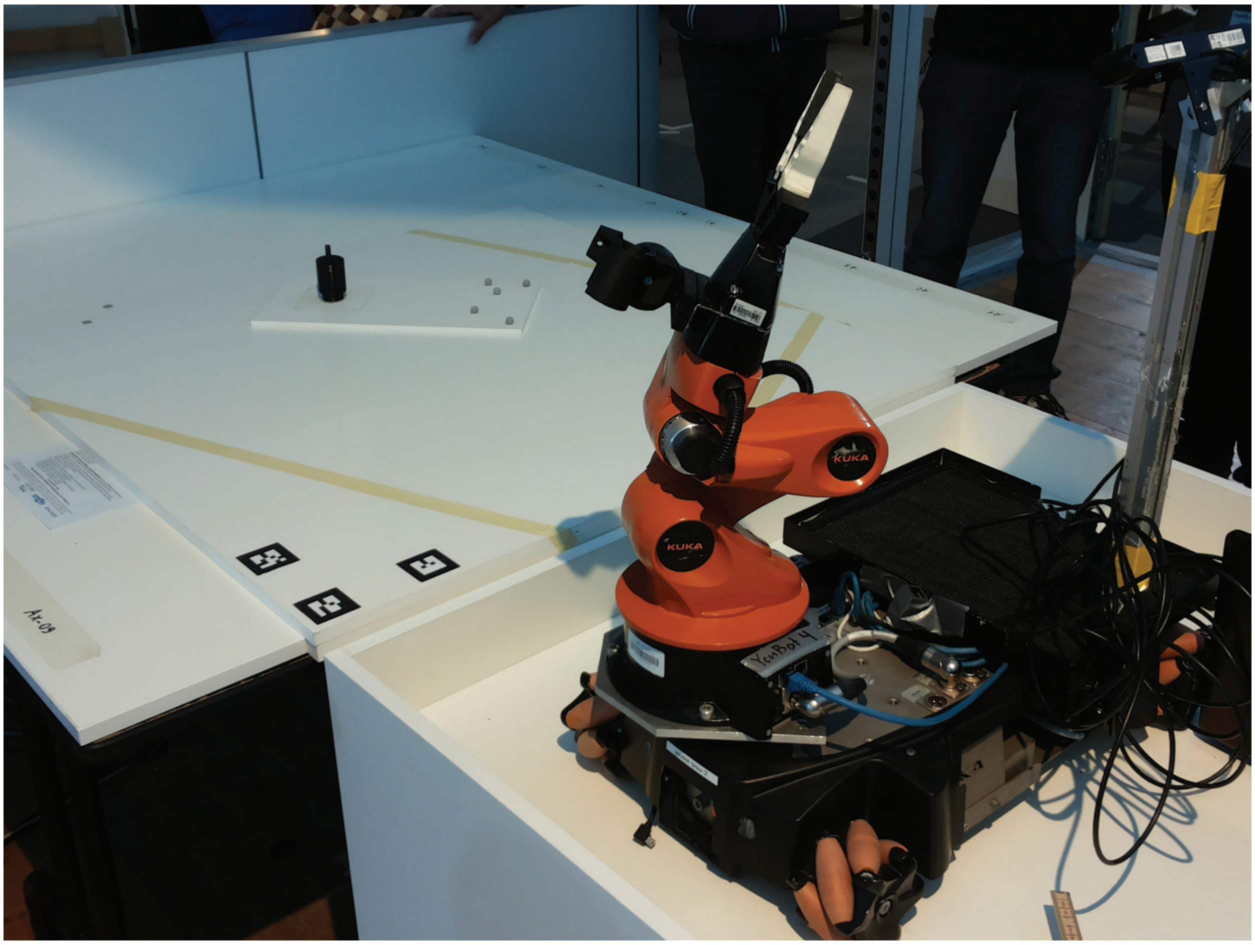


Figure 10. Example of execution of functional benchmark 1 (RoCKIn@Work version) using the setup of **Figure 9**.

In the setup of **Figure 9**, both the table top where the objects are placed for perception and a small wooden tablet supporting the objects (visible in **Figure 10**) are actually marker sets. This way, the mocap system can be used to find the transform between the reference systems associated to them; by combining such transform with the (previously recorded) transform between the object's own reference systems and the tablet's, it is possible to obtain the pose of the object with reference to the table top, to be compared to the reconstructed pose provided by the robot. The AR (augmented reality) markers visible on the table top in **Figure 10** are used to define the 2D reference system that the robot is required to use for reconstructed poses.

It is interesting to point out that the marker set of **Figure 7**, used at the 2015 Competition, is planar and thus significantly simpler to build than the '3D' version used at the 2014 Competition (which was similar to the marker set of **Figure 3**). This change is deliberate, and comes from practical experience. Its goal is to minimize occlusions between markers: in the difficult lighting conditions of the 2014 Competition (a white, partially light-transparent tent, in the open), such occlusions compromised localization performance, requiring 'on the fly' modification of the marker sets. Thus, for the 2015 Competition we designed new marker sets taking better advantage of the known features of the relative positions of mocap cameras and markers. In fact, cameras are significantly higher from the ground than markers, thanks to the mounting points on the overhead truss: therefore, with a marker set with all the markers are on the same horizontal plane, critical occlusions only tend to occur when the markers are perceived by mocap cameras that are already too far from the marker set to provide useful localization data.

7. Conclusions

The differences between scientific experiments and robot competitions are many and significant. Project RoCKIn set out to a difficult task: that of developing methodologies to design novel robot competitions whose tests, without losing the traditional role of technology showcases, could at the same time act as veritable *benchmarking experiments*.

During the life of the project, the above methodologies have been developed; a competition based on them—the *RoCKIn Competition*—has been designed; and two editions of it have been successfully held (in 2014 and 2015). This means that RoCKIn has reached its goal. Most importantly, it means that the way to further, fruitful developments in the field of robot benchmarking is open. Some of these developments are already on-going.

During the course of this chapter, the whole process leading to this result has been retraced; encompassing—without burdening the reader with excessive detail—the range from theoretical foundations to real world implementation. For what concerns the latter, particular attention has been devoted to the key problem of collecting ground truth data.

Author details

Giulio Fontana[1], Matteo Matteucci[1]*, Francesco Amigoni[1], Viola Schiaffonati[1], Andrea Bonarini[1] and Pedro U. Lima[2]

*Address all correspondence to: matteo.matteucci@polimi.it

1 Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

2 Institute for Systems and Robotics, Instituto Superior Técnico, U. Lisboa, Portugal

References

[1] RoCKIn Project. Project Website [Internet]. 2014. Available from: http://rockinrobotchallenge.eu/ [Accessed: 26 May 2017]

[2] Ahmad A, Awaad I, Amigoni F, Berghofer J, Bischoff R, Bonarini A, Dwiputra R, Fontana G, Hegger F, Hochgeschwender N, Iocchi L, Kraetzschmar G, Lima PU, Matteucci M, Nardi D, Schiaffionati V, Schneider S. RoCKIn Project D1.2 “General Evaluation Criteria, Modules and Metrics for Benchmarking through Competitions” [Internet]. 2014. Available from: http://rockinrobotchallenge.eu/rockin_d1.2.pdf [Accessed: 26 May 2017]

[3] RockEU2 Project – European Robotics League. Project Website [Internet]. 2016. Available from: http://sparc-robotics.eu/the-european-robotics-league/ [Accessed: 26 May 2017]

[4] Amigoni F, Schiaffonati V. Models and experiments in robotics. In: Magnani L, Bortolotti T, editors. Handbook of Model-Based Sciences. Springer; 2017. pp. 799-915

[5] Anderson M, Jenkins O, Osentoski S. Recasting robotics challenges as experiments. IEEE Robotics Automation Magazine. 2011;**18**(2):10-11

[6] Cohn AG, Dechter R, Lakemeyer G. The competition section: A new paper category. Artificial Intelligence. 2011;**175**:iii

[7] Holz D, Iocchi L, van der Zant T. Benchmarking intelligent service robots through scientific competitions: The RoboCup@Home approach. In: Proceeding AAAI Spring Symposium on Designing Intelligent Robots: Reintegrating AI II;Palo Alto, USA. 2013. pp. 27-32

[8] Smart B. Competitions, challenges, or journal papers. IEEE Robotics Automation Magazine. 2012;**19**(1):14

[9] Croce D, Castellucci G, Bastianelli E. Structured learning for semantic role labeling. Intelligenza Artificiale. 2012;**6**(2):163-176

[10] Tedre M, Moisseinen N. Experiments in computing: A survey. The Scientific World Journal. 2014;2014:11 p. Article ID 549398. DOI: 10.1155/2014/549398

[11] Amigoni F, Bonarini A, Fontana G, Matteucci M, Schiaffonati V. To what extent are competitions experiments? A critical view. In: Proceedings of the ICRA 2014 Workshop on Epistemological Issues in Robotics Research and Research Result Evaluation; Hong Kong. 2014

[12] RoboCup@Home Competition [Internet]. 2017. Available from: http://www.robocup.org/domains/3 [Accessed: 26 May 2017]

[13] RoboCup@Work Competition [Internet]. 2017. Available from: http://www.robocup.org/leagues/16 [Accessed: 26 May 2017]

[14] RAWSEEDS Project. Project Website [Internet]. 2004. Available from: http://www.rawseeds.org/ [Accessed: 26 May 2017]

[15] Amigoni F, Bastianelli E, Berghofer J, Bonarini A, Fontana G, Hochgeschwender N, Iocchi L, Kraetzschmar G, Lima P, Matteucci M, Miraldo P, Nardi D, Schiaffonati V. Competitions for benchmarking: Task and functionality scoring. IEEE Robotics & Automation Magazine. 2015;**22**(3):53-61

[16] RoCKIn Project. RoCKIn@Home Rulebook [Internet]. 2015. Available from: http://rockinrobotchallenge.eu/rockin_d2.1.3.pdf [Accessed: 26 May 2017]

[17] RoCKIn Project. RoCKIn@Work Rulebook [Internet]. 2015. Available from: http://rockinrobotchallenge.eu/rockin_d2.1.6.pdf [Accessed: 26 May 2017]

[18] RoCKIn Project. "Description of Ground Truth System V2", Deliverable D2.1.8 [Internet]. 2015. Available from: http://rockinrobotchallenge.eu/rockin_d2.1.8.pdf [Accessed: 26 May 2017]