Chapter 3

RoCKIn@Home: Domestic Robots Challenge

Luca Iocchi, Gerhard K. Kraetzschmar,
Daniele Nardi, Pedro U. Lima, Pedro Miraldo,
Emanuele Bastianelli and Roberto Capobianco

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70015

Abstract

Service robots performing complex tasks involving people in houses or public environments are becoming more and more common, and there is a huge interest from both the research and the industrial point of view. The RoCKIn@Home challenge has been designed to compare and evaluate different approaches and solutions to tasks related to the development of domestic and service robots. RoCKIn@Home competitions have been designed and executed according to the benchmarking methodology developed during the project and received very positive feedbacks from the participating teams. Tasks and functionality benchmarks are explained in detail.

Keywords: robot competitions, domestic robots, speech understanding, semantic mapping, person and object detection and recognition

1. RoCKIn@Home motivations and rules

With the goal of fostering scientific progress and innovation in cognitive systems and robotics, and to increase the public awareness of the current state-of-the-art of robotics in Europe, the RoCKIn project [1] developed RoCKIn@Home, a competition for domestic service robots. The competition was designed around challenges that are based on easy-to-communicate and convincing user stories, which catch the interest of both the general public and the scientific community. In particular, the latter aims at solving open scientific challenges and to thoroughly assess, compare and evaluate the developed approaches with competing ones.

The RoCKIn@Home competition hence aimed at bolstering research in service robotics for home applications, and to raise future capabilities of robot systems to meet societal challenges, like healthy ageing and longer independent living. To allow this to happen, competitions

were designed to meet the requirements of benchmarking procedures and good experimental methods. The integration of benchmarking technology with the competition concept is one of the main goals of RoCKIn.

Behind the definition of the @Home benchmarks, we considered a scenario in which an elderly person, named 'Granny Annie', lives in an ordinary apartment. Granny Annie is suffering from typical problems of aging:

- She has mobility constraints and she gets tired fast;
- She needs to have some physical exercise;
- She needs to take her medicine regularly;
- She must drink enough;
- She must obey her diet;
- She needs to observe her blood pressure and blood sugar regularly;
- She needs to take care of her pets;
- She wants to have a vivid social life and welcome friends in her apartment occasionally, but regularly;
- Sometimes she has days not feeling so well and needs to stay in bed; and
- She still enjoys intellectual challenges and reads books, solves puzzles and socializes a lot with friends.

For all these activities, RoCKIn@Home is looking into ways to support Granny Annie in mastering her life. The context for performing such activities by technical systems is set in the subsequent scenario description.

The RoCKIn@Home scenario description is structured into three sections: environment, tasks and robots.

- The environment section specifies the environment in which tasks have to be performed. This information is also relevant for building testbeds and simulators.
- The tasks section provides details on the tasks the participating teams are expected to solve through the use of one or more robots and possibly additional equipment.
- The robot section specifies some constraints and requirements for participating robots, which mainly arise for practical reasons (size and weight limitations, for example) and/or due to the need to observe safety regulations.

2. The RoCKIn@Home environment

The goal of the RoCKIn@Home environment is to reflect an ordinary European apartment, with all its environmental aspects, like walls, windows, doors, blinds, etc., as well

http://dx.doi.org/10.5772/intechopen.70015

as common household items, furniture, decoration and so on. The apartment depicted in **Figure 1** serves as a guideline. More detailed specifications are given in the rule book. The following embedded devices are installed and are accessible within the apartment's WLAN:

- A networkable, camera-based intercom at the front door. It allows to see who is in front of the door;
- The lamps in the bedroom (e.g. on the bed stand) are accessible and controllable via network; and
- The shutters on the bedroom or living room window are accessible and controllable via network.

Figure 1. Model of the apartment used in the competitions.

3. Task benchmarks

Based on the user story described above, we defined three task and three functionality benchmarks. The latter represent basic functionalities that every robot should have, in order to successfully complete the tasks.

3.1. TBM1. Task benchmark '*Getting to know my home*'

The robot is told to learn about a new environment. It is supposed to generate a semantic map of the apartment within a limited time frame. How exactly to approach this task is left to the teams. For example, a team member may 'demonstrate' the apartment by guiding the robot through the apartment, pointing to objects and speaking aloud their names. Alternatively, a robot may explore the environment completely autonomously. The robot may also interrogate a team member about the names of objects or places. At the end of the environment learning phase, the robot must show through a behaviour the understanding of the environment.

The expected robot behaviour in this task is:

- *Phase 1: knowledge acquisition*. The robot in any way (through human-robot interaction (HRI) or autonomously or mixed) has to detect changes,[1] which may include: open or close doors connecting two rooms; moved pieces of furniture; or moved objects, possible objects are shown in **Figure 2**. In case of a HRI-based approach, a team member can guide the robot in the environment and show the changes with only natural interactions (speech and gesture). No input devices are allowed (e.g. touch screens, tablets, mouse, keyboard, etc.). At any time, teams can decide to move to Phase 2, even if not all the changes have been detected. However, the task in Phase 2 can refer only to objects acquired during Phase 1.
- *Phase 2: knowledge use*. The robot has to show the use of the new acquired knowledge. This phase is accomplished by executing a user command mentioning one of the items affected by the change. The user command must be given to the robot in a natural way. The preferred way is using speech interaction.

During *Phase 1*, the robot can move around in the environment for up to the maximum time limit of this task, possibly accompanied by the user (a team member) and interacting with him/her. The robot has to detect changes, and then it must represent them in an explicit format. In *Phase 2*, the robot is asked (e.g. by receiving a voice command) to move one of the changed objects recognized in *Phase 1* to a piece of furniture, also recognized in *Phase 1*. The accomplishment of the behaviour in *Phase 2* will be rewarded only if it refers to an object/piece of furniture that has been correctly reported in the output of *Phase 1*.

For scoring and ranking, we consider the following items. The set A of achievements for this task are:

- The robot detects the door with changed state;

[1] Before each task run, some random changes in the environment are made with respect to the nominal configuration given to the teams during the set-up days.

http://dx.doi.org/10.5772/intechopen.70015

- The robot detects each piece of moved furniture;
- The robot detects each changed object; and
- The robot correctly executes the command given in *Phase 2*.

The set PB of penalized behaviours for this task are:

- The robot requires multiple repetitions of human gesture/speech;
- The robot bumps into the furniture;
- The robot stops working; and
- The robot was helped to manipulate an object.

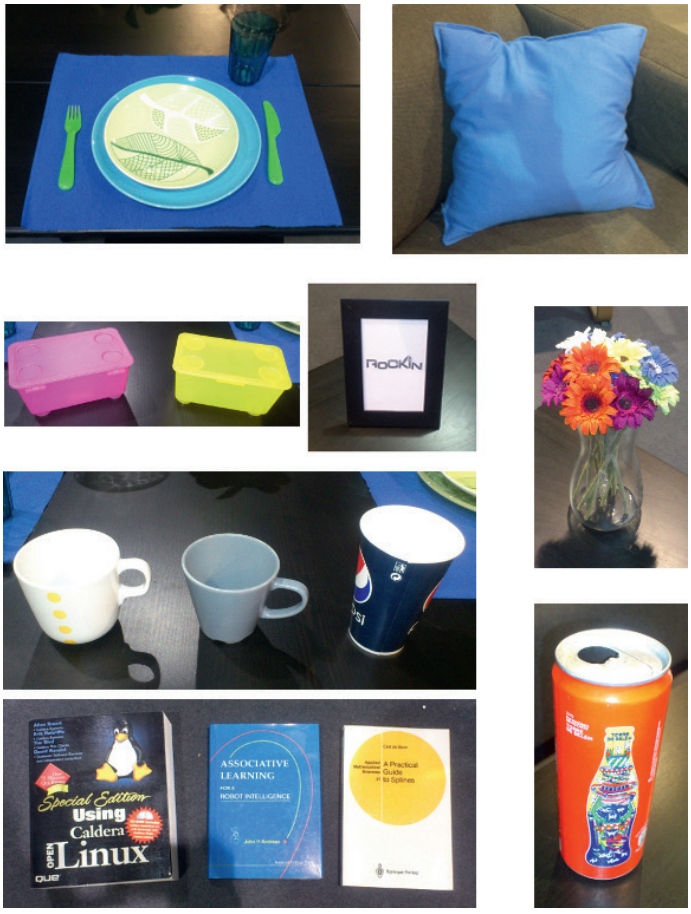

Figure 2. Objects used in TBM1.

Additional penalized behaviours may be identified and added to this list if deemed necessary. The set DB of disqualifying behaviours for this task are:

- The robot hits Annie or another person in the environment; and
- The robot damages the testbed.

Additional disqualifying behaviours may be identified and added to this list if deemed necessary. These sets will be completed in later rule revisions.

3.2. TBM2. Task benchmark '*Welcoming visitors*'

This task assesses the robot's capability to interact effectively with humans and to demonstrate different behaviours when dealing with known and unknown people.

Granny Annie stays in bed because she is not feeling well. The robot will handle visitors, who arrive and ring the doorbell, as described in Chapter 4.

In all runs of this task, the four persons indicated above will ring the doorbell. The robot is thus required to deal with all the situations described above. However, the order in which the people will appear will be randomized for each run. Every visit will terminate before the next one. Pictures of Dr. Kimble are available, and images of the uniforms of both the Deli Man and the Postman are also given to the teams (see **Figure 3**).

The task involves handling several visitors arriving in any sequence, but separately from each other. The robot must be able to handle/interact with an outside camera. If a visitor has been admitted, the robot should guide him out after the visit.

The expected robot behaviour in this task is:

- *Phase 1: detection and recognition of the visitor*. Whenever a person rings the doorbell, the robot can use its own on-board audio system or the signal from the home automation devices to detect the bell ring(s). The robot has to understand who the person is asking for a visit, using the external camera. If the robot does not detect the ring call after three times, then

Figure 3. Visitor uniforms for TBM2.

http://dx.doi.org/10.5772/intechopen.70015

the person will leave and the task will continue with the next person after a while. The robot can choose any way of opening the door, either using its manipulator or requesting a referee, a team member or the visitor to open the door (e.g. using speech).

- *Phase 2: greeting of the visitor*. For each detected visitor, the robot has to greet the visitor. In this spoken sentence, the robot has to demonstrate that it understood the category of the person.
- *Phase 3: executing the visitor-specific behaviour*. Depending on the visitor, the following behaviours are expected:
- Dr. Kimble: the robot allows the Doctor to enter and guides him/her to Annie's bedroom. Then, it waits until the Doctor exits the bedroom, follows him/her to the entrance door and allows the Doctor to exit;
- Deli Man: the robot allows the Deli Man to enter, guides the Deli Man to the kitchen, asking him/her to deliver the breakfast box on the table. Then, it guides the Deli Man back to the entrance door, and allows him/her to exit;
- Postman: the robot allows the Postman to enter, receives the postal mail (or ask the Postman to put it in the table in the hall) and allows him/her to exit; and
- Unknown person: do nothing.

After the execution of the visitor-specific behaviour, the robot should return to the initial position where it can receive the next visit.

For scoring and ranking, the seta of achievements, penalized and disqualifying behaviours for this task are those listed in Chapter 4.

3.3. TBM3. Task benchmark '*Catering for Granny Annie's comfort*'

This benchmark aims at assessing the robot's performance of executing requests about Granny Annie's comfort in the apartment.

The robot helps Granny Annie with her daily tasks throughout the day. After waking up in the morning, Granny Annie calls the attention of her service robot by touching a button on her tablet computer.[2] When the robot approaches her, Granny Annie uses spoken commands to ask the robot to operate on several home-automated devices, for instance, lifting the shutters, switching on a light, etc. Besides operating on home-automated devices, Granny can also ask the robot to further provide comfort, by looking for several of her belongings and bringing them back to her (see examples in **Figure 4**). There is no specific amount for the number of requests that Granny Annie has for the robot and the requests do not follow any specific order.

In the context of this task, a subtask is considered to be the resulting behaviour taken by the robot to accomplish something that Granny asked it to. In practical terms, if she asks the robot, for instance, to get her a cup, the resulting subtask is the process of looking for and

[2]An application for this purpose is provided by the @Home organization committee.

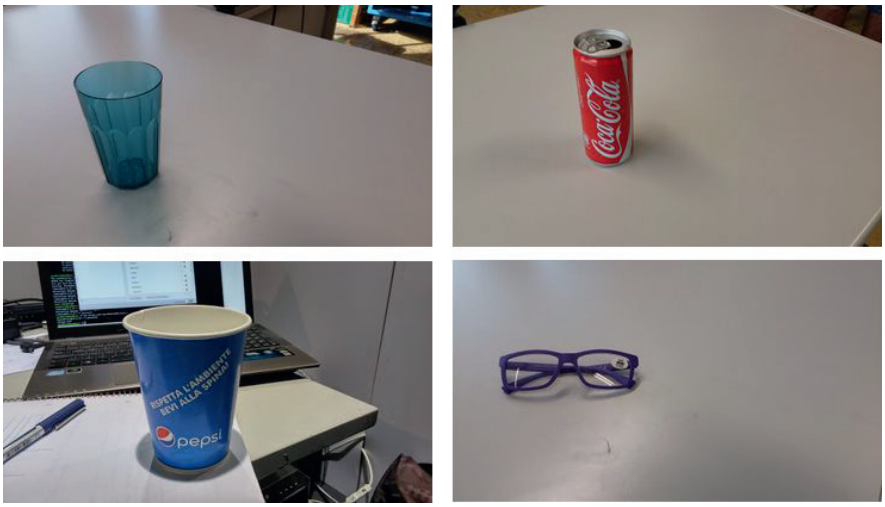

Figure 4. Objects used in TBM3.

bringing the cup back to her. In each run of this task, the robot will be asked to perform several subtasks. Granny Annie may only give one command at a time, and only after the robot executes the corresponding subtask another one may be given.

For each run of this task, in no specific order, the robot will be asked to operate the home devices and to find and bring back an object:

- Regarding the device operation, each team can choose whether the robot operates the devices with its manipulator, or over the home automation devices. The networked communication follows a pre-established common protocol which is specified by the organization committee; and
- A list of possible objects to be used is given to the teams, in advance. In addition, to ease the searching for objects, the likelihood of the position of the objects is also provided to the teams.

Afterwards, the robot will be given a finalizing command.

The expected robot behaviour in this task is:

- To reach the room where Granny Annie is located when she calls upon its service, approaching her in such a way that spoken communication is possible;
- The robot should then state its readiness to receive orders of subtasks to execute;
- When given a command, it should be confirmed in an appropriate way (e.g. by repeating it back to Granny Annie and asking if it was correctly understood). If the robot fails to understand a certain command after three tries, Granny Annie will move onto the next one; and
- The subtask corresponding to the given command should then be executed, and the robot should return to where Granny Annie is located.

http://dx.doi.org/10.5772/intechopen.70015

This procedure should be repeated until Granny orders the robot to return to its idling position, concluding the task.

For scoring and ranking, we consider the following items. The set A of achievements for this task are:

- The robot enters the room where Granny Annie is waiting;
- The robot understands Annie's command(s);
- The robot operates correctly the right device(s);
- The robot finds the right object(s); and
- The robot brings to Annie the right object(s).

The set PB of penalized behaviours for this task are:

- The robot bumps into the furniture;
- The robot drops an object; and
- The robot stops working.

Additional penalized behaviours may be identified and added to this list if deemed necessary. The set DB of disqualifying behaviours for this task are:

- The robot hits Annie or another person in the environment;
- The robot damages or destroys the objects requested to manipulate; and
- The robot damages the testbed.

Additional disqualifying behaviours may be identified and added to this list if deemed necessary.

3.4. FBM1. Functionality benchmark '*Object perception*'

This functionality benchmark has the goal of assessing the capabilities of a robot in processing sensor data, in order to extract information about observed objects. All objects presented to the robot in this task benchmark are commonplace items that can be found in a domestic environment. Teams are provided with a list of individual objects (instances), subdivided in classes. The benchmark requires that the robot, when presented with objects from such list, detects their presence and estimates their class, instance and location. For example, when presented with a bottle of milk, the robot should detect a bottle (class) of milk (instance) and estimate its pose w.r.t. a known reference frame.

The set of individual objects, which will actually be presented to the robot during the execution of the functionality benchmark, is a subset of a larger set of available objects, here denoted as 'object instances' (examples of object instances, and their respective coordinates systems are shown in **Figure 5**). Object instances are subdivided into classes of objects that have one or more properties in common, here denoted as 'object classes'. Objects of the same

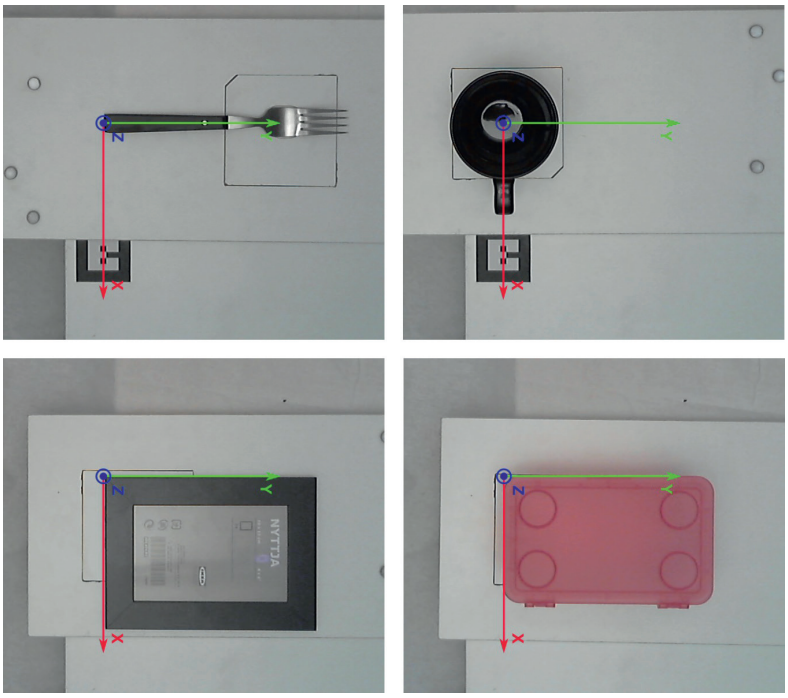

Figure 5. Object instances for FBM1.

class share one or more properties, not necessarily related to their geometry (for instance, a class may include objects that share their application domain). Each object instance and each object class is assigned a unique ID.

All object instances and classes are known to the team before the benchmark, but the team does not know which object instances will actually be presented to the robot during the benchmark. More precisely, the team will be provided with the following information:

- Descriptions of all the object instances;
- Subdivision of the object instances into object classes (for instance: boxes, mugs, cutlery); and
- Reference systems to each object instance (to be used to express object poses).

Regarding the expected robot behaviour, the objects it is required to perceive are positioned, one at the time, on a table (benchmark setup area) located directly in front of the robot. The actual pose of the objects presented to the robot is unknown before they are set on the table. For each presented object, the robot must perform:

- Object detection (i.e. class recognition): perception of the presence of an object on the table and the association between the perceived object and one of the object classes;

- Object recognition (i.e. instance recognition): association between the perceived object and one of the object instances belonging to the selected class; and
- Object localization (i.e. pose estimation): estimation of the 3D pose of the perceived object, with respect to the benchmark setup reference frame (given *a priori*).

These steps are repeated until the time runs out, or the maximum number of objects has been processed.

The evaluation of the performance of a robot according to this functionality benchmark is based on:

1. The number and percentage of correctly classified objects;
2. The number and percentage of correctly identified objects;
3. Pose error for all correctly identified objects; and
4. Execution time (if less than the maximum allowed for the benchmark).

These criteria are in order of importance (since this functionality benchmark is primarily focused on object recognition). The first criterion is applied first and teams will be scored according to the common accuracy metrics. The ties are broken by using the second criterion, again applying accuracy metrics. Finally, if needed, the position error is evaluated as well.

3.5. FBM2. Functionality benchmark *'Navigation'*

This functionality benchmark aims at assessing the capabilities of a robot to autonomously navigate in a typical apartment, containing furniture and objects spread through the apartment's rooms. From a predefined starting position, the robot will receive a list of waypoints that it must visit, before reaching a goal position.

Teams will have to take into account the following changes between different runs:

- Distinct starting points, waypoints and goal positions;
- Different number of waypoints to reach the goal; and
- Different number of obstacles blocking the path.

Teams are required to set their robot on a specific starting position (given to the teams before each run). Then, the robot should behave as follows. It receives the start signal, as well as an ordered list of waypoints that it must reach. The robot must then follow the order in which the waypoints are sent, sending back a signal each time it reaches a waypoint. The evaluation of the navigation will take into account the following three items:

- The distance between the robot's position and the respective position of the waypoint. It will be accounted both the Euclidean distance between the waypoint and the robot, and the difference in the orientation;

- The time spent by the robot to go from each waypoint to the next waypoint; and
- The number of times that the robot hits each obstacle. If the robot hits the same obstacle more than once, it will count as multiple hits.

The functionality benchmark ends as soon as the robot reaches the last waypoint, the time available for the functionality benchmark expires or if the robot hard-hits an obstacle.

The objects that can be in the robot's path are divided as follows:

- *Static and previously mapped:* hardware already present in the house such as furniture, doors and walls. The teams should already have these obstacles mapped from set-up days. These items will not change during this functionality benchmark;
- *Static:* items Granny Annie left lying on the ground. The obstacles may be of different shapes and sizes, are not previously known by the teams and may be different in between runs; and
- *Dynamic:* Granny Annie's visitors. People moving inside the house. Obviously, the movement people will do is unpredictable.

Regarding the scoring and ranking, at each run and for each team, three metrics will be used to score the performance:

- Accuracy scoring will be based on the distance and the orientation errors. The mean of the distances between the robot and the target waypoint is computed and stored in A, while the difference in orientations computed and stored in B. After the computation of these accuracy scorings, they will be discretized and fitted in one of the following groups:
  - 1: $A < 10$ cm AND $B < 20°$;
  - 2: $A < 30$ cm AND $B < 45°$;
  - 3: $A < 50$ cm AND $B < 90°$; and
  - 4: $A < 80$ cm AND $B > 90°$;

A lower group number corresponds to the better performance. Therefore, teams will be ranked starting from group 1. Note that for a team to be placed in any of the groups, it must respect the limits for A and B. If a team has a score that does not fit any of the groups defined above (e.g. mean of the error above 80 cm), it will not receive scoring in the respective functionality benchmark run;

- If more than one team falls inside each of the previously defined group, the number of obstacle hits will be used as a tie breaker, where the team with less hits will be ranked first and so on. Note that hits will only be considered as a tie breaker, i.e. a team in group 2 will never be ranked before any team in group 1, despite of the number of hits; and
- If teams are still tied, time will be the decisive tie breaker.

http://dx.doi.org/10.5772/intechopen.70015

3.6. FBM3. Functionality benchmark '*Speech understanding*'

This functionality benchmark aims at evaluating the ability of a robot to understand speech commands that a user gives in a home environment. A list of commands will be selected among the set of predefined recognizable commands (i.e. commands that the robot should be able to recognize within the tasks of the competition or in similar situations).

Each implemented system should be able to capture audio from an on-board microphone, to record the captured audio in a file and to interpret the corresponding utterance. A standard format for audio files will be chosen (e.g. WAV) and communicated to the teams in advance before the competition. The system should produce an output according to a final representation defined below. Such a representation will have to respect a command/arguments structure, where each argument is instantiated according to the command evoking the verb. It is referred to as Command Frame Representation (CFR) (e.g. 'go to the living room' will correspond to MOTION (goal:"living room")). Summarizing, for each interpreted command the following relevant information will be collected: an audio file, its correct transcription and the corresponding correct CFR.

Variations between different runs can be:

- Different complexity in the syntactic structures of the spoken commands;
- The use of complex grammatical features, as pronouns;
- The use of synonyms for referring to objects; and
- The use of sentences where more than one action is expressed, resulting in a composed command (e.g. 'take the bottle and bring it to me').

Furthermore, variation in the quality of the audio corresponding to the user utterances can be considered, as for representing more or less noisy conditions.

Some information about the lexicon (verbs and nouns of objects) used in the benchmark is made available to the teams before the competition. In order to evaluate the correct understanding of a command expressed in natural language (e.g. through a sentence), a semantic representation formalism based on semantic frames has been selected. Each frame corresponds to an action, namely, a robot command. A set of arguments is associated to each frame, specifying part of the command playing a particular role with respect to the action expressed by the frame. For example, in the command 'go to the dining room' the motion frame is expressed by the verb go, while the part of the sentence 'to the dining room' corresponds to the goal argument, indicating the destination of the motion action. The set of frames defined and selected for this benchmark are given to the times before the competition.

Composition of actions is also possible in the CFR, corresponding to more complex action as the 'pick and place' action, represented by a sequence of taking frame followed by a bringing frame (e.g. for the command 'take the box and bring it to the kitchen'). The grammar specifying the correct syntax for a CFR is also provided.

Regarding the expecting robot behaviour, it should be able to understand a command starting from the speech input. The robot should correctly transcribe the user utterance and recognize the action to perform, resulting in the correct command frame (e.g. MOTION for a motion command) and the arguments involved (e.g. the goal of a motion command). The output of the robot should provide the CFR format for each command. For each command uttered or for each audio file directly provided during the speech understanding functionality benchmark, the system should generate the corresponding transcription and the interpretation in the CFR format.

All the teams are evaluated on the same set of spoken sentences. These spoken sentences are divided in two groups: a first group is formed by pre-recorded audio files, and a second group by voice commands uttered by a user during the benchmark. The robots are disposed in a circle, and the audio are broadcast using a 360° speaker (or an equivalent structure of speakers) with high fidelity performance placed in the centre. In this way, all the robots receive the same audio at the same time.

All teams are required to perform this functionality benchmark according to the steps mentioned below:

1. Each team receives the audio files randomly selected among the predefined set. This subset is the same for each team in order to reproduce fair conditions in the evaluation. Only one button can be pressed (either a button in a graphical user interface (GUI) or a key in the keyboard) to start the benchmark;
2. For each audio file, the system should generate the corresponding interpretation in the CFR format, together with the correct transcription of the corresponding utterance. The time for this processing will be restricted to an amount that is communicated in advance by the organization committee; and
3. After a proper communication, a member of the organization committee pronounces some commands using a microphone. The audio is instantly reproduced using a loudspeaker, conveniently positioned to be equally distant from each robot involved in the benchmark. Each command will be given after an interval of about 15 s of silence from the previous one. During this second part of the test, a designated member of the team will be allowed to press a button of the robot PC once for each sentence uttered by the speaker.

After the test is completed, only one button can be pressed to stop the processing.

During the execution of the benchmark, the following data are collected:

- Sensor data (in the form of audio files) used by the robot to perform speech recognition;
- The set of all possible transcription for each user utterance;
- The final command produced during the natural language analysis process; and
- Intermediate information produced or used by the natural language understanding system during the analysis as, for example, syntactic information.

Regarding the scoring and ranking, different aspects of the speech understanding process are assessed:

http://dx.doi.org/10.5772/intechopen.70015

- The word error rate on the transcription of the user utterances, in order to evaluate the performance of the speech recognition process.
- For the generated CFR, the performance of the system will be evaluated against the provided gold standard version of the CFR, which is conveniently paired with the analysed audio file and transcription. Two different performance metrics will be evaluated at this step. One measuring the ability of the system in recognizing the main action, called Action Classification (AcC), and one related to the classification of the action arguments, called Argument Classification (AgC). In both cases, the evaluations will be carried out in term of Precision, Recall and F-Measure. This process is inspired by the Semantic Role Labeling evaluation scheme proposed in [24]. For the AcC, this measures will be defined as follow:
  - Precision: the percentage of correctly tagged frames among all the frames tagged by the system;
  - Recall: the percentage of correctly tagged frames with respect to all the gold standard frames; and
  - F-Measure: the harmonic mean between Precision and Recall.

  Similarly, for the AgC, Precision, Recall and F-Measure will be evaluated, given an action *f*, as:
  - Precision: the percentage of correctly tagged arguments of *f* with respect to all the arguments tagged by the system for *f*.
  - Recall: the percentage of correctly tagged arguments of *f* with respect to all the gold standard arguments for *f*.
  - F-Measure: the harmonic mean between Precision and Recall.
  - Time utilized (if less than the maximum allowed for the benchmark).

The final score is evaluated considering both the AcC and the AgC. Only the F-Measure is considered for both measures, each one contributing for 50% of the score. The AgC F-Measure is evaluated for each argument, and the final F-Measure for the AgC is the sum of the single F-Measure of the single arguments divided by the number of arguments. This final score has to be considered as an equivalence class. If this score will be the same for two or more teams, the *WER* will be used as penalty to evaluate the final ranking. This means that a team belonging to an equivalence class cannot be ranked lower than one belonging to a lower one, even though the final score, considering the *WER* of the first is lower than the score of the second.

4. Robots and teams

The purpose of this section is twofold:

1. It specifies information about various robot features that can be derived from the environment and the targeted tasks. These features are to be considered at least as desirable, if not

required, for a proper solution of the task. Nevertheless, we will try to leave the design space for solutions as large as possible and to avoid premature and unjustified constraints.

2. The robot features specified here should be supplied in detail for any robot participating in the competition. This is necessary in order to allow better assessment of competition and benchmark results later on.

4.1. General specifications and constraints on robots and teams

A competition entry may use a single robot or multiple robots acting as a team. At least one of the robots entered by a team must be mobile, and able to visit different task-relevant locations by autonomous navigation. Teleoperation (using touch screens, tablets, mouse, keyboard, etc.) of robots for navigation is not permitted (except when otherwise specified, e.g. in particular instances of task and functionality benchmarks). The robot mobility must work in the kind of environments specified for RoCKIn@Home, and on the kind of floors defined in the RoCKIn@Home environment specifications.

Any robot used by a team may use any kind of on-board sensor subsystems, provided that the sensor system is admitted for use in the general public, its operation is safe at all times and it does not interfere with other teams or the environment infrastructure. A team may use any kind of sensor system provided as part of the environment, by correctly using a wireless communication protocol specified for such purpose and provided as part of the scenario.

Any robot used by a team may internally use any kind of communication subsystem, provided that the communication system is admitted for use in the general public, its operation is safe at all times and it does not interfere with other teams or the environment infrastructure. A robot team must be able to use the communication system provided as part of the environment by correctly using a protocol specified for such purpose and provided as part of the scenario.

Any mobile device (especially robots) must be designed to be usable with an on-board power supply (e.g. a battery). The power supply should be sufficient to guarantee electrical autonomy for a duration exceeding the periods foreseen in the various benchmarks, before recharging of batteries is necessary. Charging of robot batteries must be done outside of the competition environment.

Any robot or device used by a team as part of their solution approach must be suitably equipped with computational devices (such as on-board PCs, microcontrollers or similar) with sufficient computational power to ensure safe autonomous operation. Robots and other devices may use external computational facilities, including Internet services and cloud computing to provide richer functionalities, but the safe operation of robots and devices may not depend on the availability of communication bandwidth and the status of external services.

All robots are checked by the organization committee for compliance with the specifications and constraints described in the rulebook. Teams will be asked to show the safety mechanisms of their robots and to demonstrate their use. A live demonstration is necessary: for example, pushing an emergency stop button while the robot is moving and verifying that the robot immediately stops. If the robot has other mechanical devices (e.g. a manipulator), their safety must be demonstrated as well. This inspection is done before the competition.

5. RoCKIn@Home research challenges and solutions

The development of the functionalities required by the tasks described in the previous section was very challenging for the teams, since it required not only realizing and testing robust solutions for each component, but also to properly integrate them in a fully working system. In this section, we briefly summarize the main research challenges that inspired the competition tasks and provide some comments about the adopted solutions.

For other features not described here, such as navigation and mapping, standard off-the-shelf components have been used by the teams.

5.1. Person and object detection and recognition

Person and object detection and recognition are important basic functionalities for service robots. In RoCKIn, TBM2 and FBM1 focussed on these topics.

TBM2 was designed to assess the ability of robots to properly understand the user with whom they are interacting and to provide the adequate behaviour according to the situation.

Many techniques are available in computer vision for face detection [2], face recognition [3], person modelling and people tracking [4]. However, their application on a robotic platform with limited on-board computation, real-time constraints and limited Internet connection for using cloud services makes this functionality very challenging.

During RoCKin competitions, person recognition was addressed in TBM2, where the robot was required to distinguish among four different kinds of people and to act accordingly. Images to be processed came from a fixed external camera (the same for all the teams) through a wireless link to the robot. Moreover, the robot can also decide to open the door and further examine the person with its on-board sensors.

This setup allowed teams to use some calibration procedure to identify the visitors according to some known features. For example, Dr. Kimble can be recognized through face recognition, while the Deli Man and the Postman by their uniforms.

Although this component may be considered quite straightforward and easy to implement, the integration in the entire system and some practical difficulties of the competition environment (e.g. acquiring images through a wireless channel in real time) required a very robust implementation.

Object recognition was specifically assessed in FBM1. Also, this test is significantly different from standard computer vision benchmark, since (1) the robot can move its sensors in order to reach a desirable viewpoint or integrate several views over time, and (2) position and orientation of recognized objects must also be estimated. Items to be recognized were available to teams during the set-up days before the test and, also in this case, the teams could benefit from calibration procedures. However, the test takes place in a physical environment (not through image dataset) and thus a variability introduced by different lighting conditions between calibration time and testing time must be considered and robustness to this variability is required to keep a high score.

5.2. Speech understanding

Speech understanding is also a fundamental feature of service and domestic robots, since spoken language is the most natural human-human communication means. Robots capable of understanding human language become accessible to a wider range of users, especially non-experts. This task is composed by two sub-tasks, namely, automatic speech recognition (ASR), that is the process of translating an audio signal into a written text, and (spoken) natural language understanding (NLU), that is the process of assigning a semantic interpretation to the transcribed text [5]. Many techniques are available to tackle ASR and NLU. For the first sub-task, it is possible either to rely on grammar-based method [6], or free-form methods [7]. For NLU, it is possible to rely on features embedded in the grammar framework [8, 9], or rely on data-driven methods [10, 11], where several machine learning techniques can be applied. Gold standards (i.e. ground truths) are necessary to evaluate the performances of both tasks. One of the most used metrics to evaluate ASR systems is word error rate [12], which measures the distance between a transcription hypothesis and the correct transcription. The NLU task instead is often evaluated using metrics derived from information retrieval, namely, Precision (P), Recall (R) and F-Measure (F1), over the semantic annotations.

FBM3 has been designed specifically to assess speech understanding capabilities of robotic platforms. In general, the task was to acquire a set of audio inputs of spoken commands, transcribe them and finally provide a semantic interpretation for each input, representing the actions and the related arguments of the intended command. Such interpretation had to be given according to a formalism inspired by frame semantics [13], specifically as it is represented in FrameNet [14]. Apart from this formalism, no further constraints have been given on the task, so that every team could develop its own system, either relying on a grammar-based method, or on data-driven ones. The benchmark was organized in two phases. In the first one, the audio input was presented to the team as audio files, bypassing the microphone acquisition. In the second phase, which was less controlled and more realistic, a live audio coming from a speaker needed to be acquired and analysed. Given the composite nature of the speech understanding task, it has been necessary to measure the performance of the two aforementioned sub-tasks to eventually evaluate the FBM3. WER has been used for ASR. Two factors have been instead measured for the understanding step: the action recognition (AcC), that is the ability of recognizing the sole actions (without arguments) expressed in a sentence, and the argument recognition (AgR), which takes into account also the action arguments. P, R and F1 have been evaluated for both AcC and AgC.

In order to provide a resource for designing, training and testing speech understanding systems, a corpus of spoken commands has been collected [11, 15]. Such resource has been incrementally build before and during the RoCKIn events (camps and competitions), through simulated or real interaction with robotic platforms. It is a collection of audio files of spoken commands gathered in diverse environmental conditions. Each command transcription is tagged with different levels of linguistic information, like morphology, part-of-speech tags and syntactic dependency trees. On top of that, semantic information is provided in terms of frame semantics. This semantic layer encodes the action intended in a command, together with its parameters. Although resources to evaluate either speech recognition [16] or natural

language understanding [17, 18] for robotics have been developed in the past, this resource differs from them in many aspects. First of all, the provided linguistic information is made explicit and given according to linguistically supported theories (e.g. POS-tags, syntactic dependencies and semantic frames). Secondly, it covers all the linguistic processing steps, providing both audio files and annotations over the corresponding transcriptions. It can be thus used to train or design general linguistic modules of a natural language processing pipeline for robotics. Thirdly, it has been gathered in different phases, and thus it presents a high variability in terms of background noise, complexity of language structures and cardinality of the lexicon. These peculiarities were transferred inside the FBM3, making it definitely different from other benchmarks, specifically for the variability of the language, and the specificity of the adopted semantic formalism. Teams had to devise systems capable of dealing with complex syntactic structures, as well as unseen words. Moreover, the live acquisition phase put additional challenges in setting up suitable microphone configurations. Such difficulties led to poor performance during the first runs of the FBM3, which improved sensitively while going further in the competition, reaching final convincing performance from more than one team at the very end. Although some promising results have been achieved along the whole FBM3, there are still some aspects to explore, and issues to be tackled. An important feature of spoken interaction is dialogue. Robots should be able to deal with longer and more complex spoken interactions to appear more natural, being able, for example, to manage anaphora phenomena that may arise during longer interactions. Another crucial aspect is the acquisition of the audio. The audio can come, from several directions, according to the speaker positioning. Reaching a uniform performance on input coming from different points is for sure a challenge to address.

5.3. Semantic mapping

Semantic mapping is the incremental process of associating relevant information of the world (i.e. spatial information, temporal events, agents and actions) to a formal description supported by a reasoning engine [19], with the aim of learning to understand, collaborate and communicate. In particular, a semantic map is a representation that contains, in addition to spatial information about the environment, assignments of mapped features to entities of known classes [20]. Semantic maps should represent knowledge that can be used by a robot for reasoning and behaviour generation, thus enabling additional information to be inferred whenever the representation is associated with a reasoning or planning engine.

Multiple approaches have been proposed in the literature, characterized by an extreme heterogeneity of methodologies for representing learned maps—that prevents comparative evaluations, standard validation and evaluation procedures, and benchmarking strategies. For example, in Ref. [21] environmental knowledge is represented by anchoring sensor data to symbols of a conceptual hierarchy, based on description logic. The authors validate their approach by building their own domestic-like environment and testing the learned model through the execution of navigation commands. A multi-layered representation, ranging from sensor-based maps to a conceptual abstraction (an OWL-DL ontology), is generated in Ref. [22]. Except for individual modules, their experimental evaluation is mainly qualitative.

Instead, in Ref. [23], a conceptual map is represented as a probabilistic chain graph model, and Ref. [23] evaluate their method by comparing the belief of the robot of being in a certain location against the ground truth. In practice, none of the cited works can compare the performance of their semantic mapping method against each other.

For this reason, in Ref. [19] a formalization of a basic general structure for semantic maps is proposed, as the result of a generalization and intersection effort with respect to the representations adopted in the literature. This representation is proposed to play the role of a common interface among all the semantic maps, and can be easily extended or specialized as needed. Given two semantic maps of the same environment that implement this basic representation, it is at least possible to compare both the semantic and the geometrical parts of the representations [24]. In particular, given a ground truth, it is possible to define some error metrics that account for both the lack and inconsistency of stored information.

This evaluation approach has been applied in the scoring of the TBM1 test. More specifically, the teams have to provide at the end of the run a KB containing the semantic information about the environment acquired during the test. This KB is compared with a ground truth and the score is assigned by considering how many correct semantic labels are reported in the output KB. The use of this scoring methodology was extremely useful to compare different approaches of semantic mapping and, as mentioned above, can be further extended and used outside RoCKIn tasks.

Author details

Luca Iocchi[1]*, Gerhard K. Kraetzschmar[2], Daniele Nardi[1], Pedro U. Lima[3], Pedro Miraldo[3], Emanuele Bastianelli[1] and Roberto Capobianco[1]

*Address all correspondence to: iocchi@dis.uniroma1.it

1 DIAG, Sapienza University of Rome, Italy

2 Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany

3 Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

References

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

[2] Zhang C, Zhang Z. A Survey of Recent Advances in Face Detection, Microsoft Research. Technical Report 2010-66. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/facedetsurvey.pdf

[3] Bowyer KW, Chang K, Flynn P. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Computer Vision and Image Understanding. 2006;**101**(1):1-15. ISSN 1077-3142

[4] Yilmaz A, Javed O, Shah M. Object tracking: A survey. ACM Computing Surveys. 2006;**38**(4):13. ISSN 0360-0300

[5] de Mori R. Spoken language understanding: A survey. In: Furui S, Kawahara T, editors. IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007; December 9-13, 2007; Kyoto, Japan. IEEE; 2007. pp. 365-376

[6] Bos J. Compilation of unification grammars with compositional semantics to speech recognition packages. In Proceedings of the 19th International Conference on Computational Linguistics. Vol. 1. COLING '02; Stroudsburg, PA, USA; Association for Computational Linguistics; 2002. pp. 1-7

[7] Chelba C, Xu P, Pereira F, Richardson T. Large scale distributed acoustic modeling with back-off n-grams. IEEE Transactions on Audio, Speech, and Language Processing. 2013;**21**(6):1158-1169, IEEE Press

[8] Bos J, Oka T. A spoken language interface with a mobile robot. Artificial Life and Robotics. 2007;**11**(1):42-47

[9] Connell J. Extensible grounding of speech for robot instruction. In: Markowitz J, editor. Robots that Talk and Listen. De Gruyter, Germany; 2014

[10] Tellex S, Kollar T, Dickerson S, Walter MR, Banerjee AG, Teller S, Roy N. Approaching the symbol grounding problem with probabilistic graphical models. AI Magazine. 2011;**34**(4):64-76

[11] Bastianelli E, Castellucci G, Croce D, Iocchi L, Basili R, Nardi D. Huric: A human robot interaction corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); Reykjavik, Iceland. European Language Resources Association (ELRA); 2014

[12] Popovic M, Ney H. Word error rates: Decomposition over pos classes and applications for error analysis. In: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07; Stroudsburg, PA, USA. Association for Computational Linguistics; 2007

[13] Fillmore CJ. Frames and the semantics of understanding. Quaderni di Semantica. 1985; **6**(2):222-254

[14] Baker CF, Fillmore CJ, Lowe JB. The berkeley frameNet project. In: Proceedings of ACL and COLING. Association for Computational Linguistics; 1998. pp. 86-90

[15] Bastianelli E, Iocchi L, Nardi D, Castellucci G, Croce D, Basili R. RoboCup@Home spoken corpus: Using robotic competitions for gathering datasets. In: RoboCup 2014: Robot World Cup XVIII [papers from the 18th Annual RoboCup International Symposium; 15 July 2014; Joaõ Pessoa, Brazil. 2014c. pp. 19-30

[16] Bugmann G, Klein E, Lauria S, Kyriacou T. Corpus-based robotics: A route instruction example. In: Proceedings of Intelligent Autonomous Systems (IAS-8); 2004. pp. 96-103

[17] Dukes K. Train robots: A dataset for natural language human-robot spatial interaction through verbal commands. In: ICSR. Embodied Communication of Goals and Intentions Workshop; 2013

[18] MacMahon M, Stankiewicz B, Kuipers B. Walk the talk: Connecting language, knowledge, and action in route instructions. In: Proceedings of the 21st National Conference on Artificial Intelligence. Vol. 2. AAAI '06. AAAI Press; 2006. pp. 1475-1482

[19] Capobianco R, Serafin J, Dichtl J, Grisetti G, Iocchi L, Nardi D. A proposal for semantic map representation and evaluation. In: 2015 European Conference on Mobile Robots. IEEE; 2015. pp. 1-6. DOI: 10.1109/ ECMR.2015.7324198

[20] Nüchter A, Hertzberg J. Towards semantic maps for mobile robots. Robotics and Autonomous Systems. 2008;**56**(11):915-926. DOI: 10.1016/j.robot.2008.08.001

[21] Galindo C, Saffiotti A, Coradeschi S, Buschka P, Fernandez-Madrigal JA, Gonzalez J. Multi-hierarchical semantic maps for mobile robotics. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2005. pp. 2278-2283. DOI: 10.1109/ IROS.2005.1545511

[22] Zender H, Mozos OM, Jensfelt P, Kruijff GJM, Burgard W. Conceptual spatial representations for indoor mobile robots. Robotics and Autonomous Systems. 2008;**56**(6):493-502. DOI: 10.1016/j.robot.2008.03.007

[23] Pronobis A, Jensfelt P. Large-scale semantic mapping and reasoning with heterogeneous modalities. In: 2012 IEEE International Conference on Robotics and Automation. IEEE; 2012. pp. 3515-3522. DOI: 10.1109/ICRA.2012.6224637

[24] Capobianco R. Interactive generation and learning of semantic-driven robot behaviours [Thesis]. Sapienza University of Rome