Chapter 4
RAWSEEDS: Building a Benchmarking Toolkit for Autonomous Robotics

Giulio Fontana, Matteo Matteucci and Domenico G. Sorrenti

Abstract Within computer science, autonomous robotics takes the uneasy role of a discipline where the features of *both* systems (i.e., robots) and their operating environment (i.e., the physical world) conspire to make the application of the experimental scientific method most difficult. This is the reason why much experimental work in robotics is, from the methodological point of view, built on shaky grounds. In particular, scientifically sound benchmarking tools are still largely missing. This chapter starts from RAWSEEDS, a project focused precisely on benchmarking in robotics, to highlight the reasons for these difficulties and to propose strategies for overcoming some of them. The main result of RAWSEEDS is a Benchmarking Toolkit: a readily usable instrument to assess and compare algorithms for SLAM, localization, and mapping. Its most innovative aspects include a set of high-quality, validated, multisensor datasets, collected both in indoor and in outdoor locations and complemented by ground truth data, and the explicit definition of a set of quantitative performance metrics for the evaluation of algorithms.

Keywords Robotic datasets · Benchmarking · SLAM · Ground truth

G. Fontana (✉) · M. Matteucci
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
e-mail: giulio.fontana@polimi.it

M. Matteucci
e-mail: matteo.matteucci@polimi.it

D.G. Sorrenti
Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy
e-mail: domenico.sorrenti@unimib.it

F. Amigoni and V. Schiaffonati (eds.), *Methods and Experimental Techniques in Computer Engineering*, PoliMI SpringerBriefs,
DOI: 10.1007/978-3-319-00272-9_4,

4.1 Introduction

Autonomous robotics is a peculiar discipline. While computing is certainly a key ingredient to it, the added element of physical interaction with the environment changes everything. The presence of large uncertainties (in the outcome of the interactions between robot and environment) and errors (in the perception of the world by the robot) are not the exception, but the rule. Moreover, an autonomous robot determines the course of its actions according to its own assessment of the world: thus, the very behavior of the robot is subject to considerable uncertainty.

This translates into a methodological problem. While the importance of the experimental scientific method to autonomous robotics is as large as it is to other scientific fields associated to computing, the practical difficulties associated to performing accurate experimentation are not. For this reason, methodological soundness often takes a secondary role in robotics papers; while, in the absence of sufficient guarantees of repeatability and/or reproducibility, even the best experimental work tends to take the diminutive role of a mere "proof of concept".

The task of assessing the performance of a robot system (or subsystem) is never trivial; yet, a whole new level of complexity (and cost) must be added to make the results of such assessment usable outside the group which designed the system. For this reason, it is often impossible to quantitatively compare the experimental results obtained with different solutions and/or by different research teams. Over the years, these problems have become increasingly critical to the development of robotics, to the point that today there exists a widespread drive to define reliable benchmarking tools and methodologies [1, 2]. The fact that such tools and methodologies are still largely missing is not due to lack of effort: instead, it is a side effect of strongly heterogeneous experimental conditions, of the weak repeatability of most experimental results, and of the use of subjective and/or insufficiently general performance metrics.

This situation has an even greater impact on industrial research policies. Companies are wary of entering a technological field where marketable applications abound, but heavy investments are needed even to perform preliminary testing of an idea: especially where the lack of established benchmarks also makes technological progess difficult to prove to prospective clients.

The RAWSEEDS project[1] is a benchmarking effort focused on sensor fusion, self-localization, mapping and SLAM in autonomous mobile robots.[2] This chapter describes the development of the RAWSEEDS *Benchmarking Toolkit*, illustrating how the choice of building it on an explicit methodological foundation had an impact on

[1] Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets (RAWSEEDS) [4] has been financed by the European Commission, within the 6th Framework Programme.

[2] Sensor fusion is the joint processing of more than one sensor datastreams. Self-localization is the process of finding one's position on a map of the environment. Mapping is the set of operations required to build such a map, usually involving exploration. Finally, Self-Localization And Mapping (SLAM) requires to autonomously build a map of the environment and to keep track of one's location on it.

each stage of the real-world experimental effort of the project. As we will see, the aforementioned impact was (as anticipated in the planning stage) strong in terms of cost, effort and complexity.

The RAWSEEDS Benchmarking Toolkit is composed of:

1. several high-quality multisensor *datasets*, with associated ground truth (GT), gathered by exploring real-world environments (indoor and outdoor) with a mobile robot equipped with a wide set of sensors;
2. a set of *Benchmark Problems* (BPs) built on such datasets, i.e., well-defined problems that also include quantitative criteria to assess their solutions;
3. example solutions to the BPs, called *Benchmark Solutions* (BSs), based on state-of-the-art algorithms and evaluated according to the criteria defined by the associated BPs.

RAWSEEDS is not the only effort towards the definition of benchmarks in SLAM. Other projects with a similar objective exist, the most known of which is the Robotics Data Set Repository (Radish) [3]. However, the RAWSEEDS Benchmarking Toolkit sports novel features, many of which explicitly focus on methodological issues. These include:

- presence of ground truth: each dataset is associated to ground truth data generated independently from the robot sensors;
- strong multisensoriality: the multiple data streams that compose each dataset have been produced at the same time by several sensor systems mounted on the same robot base, thus presenting fully coherent data to subsequent sensor fusion processes;
- data completeness: all data produced by the sensors have been logged in raw form, without performing any data reduction or lossy data compression;
- data synchronization: efforts were devoted to ensure that all data—including those produced by the ground truth collection system—were accurately timestamped according to a single reference clock source;
- data validation and certification: a separate, specific activity was devoted to assess and certify the quality of the datasets and their fitness for their intended purpose;
- wide set of scenarios: datasets have been captured (with the same hardware) indoors and outdoors, under natural and artificial light, in static and dynamic conditions, thus giving to the user the possibility to perform comparisons and verifications;
- explicit evaluation metrics: each Benchmark Problem includes the methodologies needed to evaluate objectively any solution to it, allowing comparison of different solutions independently from the actual choices of implementation and/or representation.

By providing high-quality benchmarking tools to researchers working in SLAM and associated fields, RAWSEEDS empowers them with a way to objectively measure their progress, as well as to compare it to available state-of-the-art algorithms. This is expected to lead to smoother progress (dead-ends are identified earlier), faster recognition by the community of outstanding solutions, and quicker and wider adoption

of successful approaches. For what concerns industry, RAWSEEDS aims at providing companies with ready-made tools to evaluate and compare the performance of available algorithms, increasing their confidence towards incorporating them into innovative robotic products. Moreover, the output of the same tools could also be used by companies to demonstrate the quality of their products.

Section 4.2 of this chapter describes the experimental setup, environments and procedures for RAWSEEDS data collection activities. Given the importance of ground truth data within the methodological background of RAWSEEDS, Sect. 4.3 will be dedicated to sketching how they were collected. Section 4.4 focuses on the elements of RAWSEEDS Benchmarking Toolkit, briefly outlining their features. Finally, Sect. 4.5 closes this chapter and outlines the limits of RAWSEEDS.

4.2 Setup and Data Acquisition

As previously said, the RAWSEEDS project focused on methodological issues right from the planning stage. Indeed, many of the methodology-grounded features outlined in Sect. 4.1 had a direct impact both on the design of the robot used to collect sensor data and on the data collection activities. Namely, such features are: strong multisensoriality, data completeness, data synchronization and wide set of scenarios. This section outlines the technological choices made to meet the requirements posed by these features.

4.2.1 Robot Setup

The robot used to acquire RAWSEEDS datasets was composed of two elements: a robot platform and a "sensor frame" module affixed to it, which comprised all the sensors. The robot platform was a custom (i.e., non-commercial) design called *Robocom*, designed for high payload, small dimensions and very good maneuverability. The last two qualities proved crucial for indoor data acquisitions, while the first was necessary due to the significant weight of the sensor frame. During RAWSEEDS data acquisitions, the robot was teleoperated; autonomous navigation would not have added to the quality of the datasets, while requiring additional effort to be set up.

The sensor systems used by RAWSEEDS were chosen to cover a wide range of devices, selected among those which are more frequently used for SLAM. They are:

- Black-and-white binocular and trinocular camera systems.
- Color monocular camera.
- Color omnidirectional camera.
- Four laser range scanners (two high-performance units and two low-cost ones).
- Inertial Measurement Unit.
- Sonar belt comprising 12 ultrasonic transceivers.
- Odometry system based on wheel encoders.

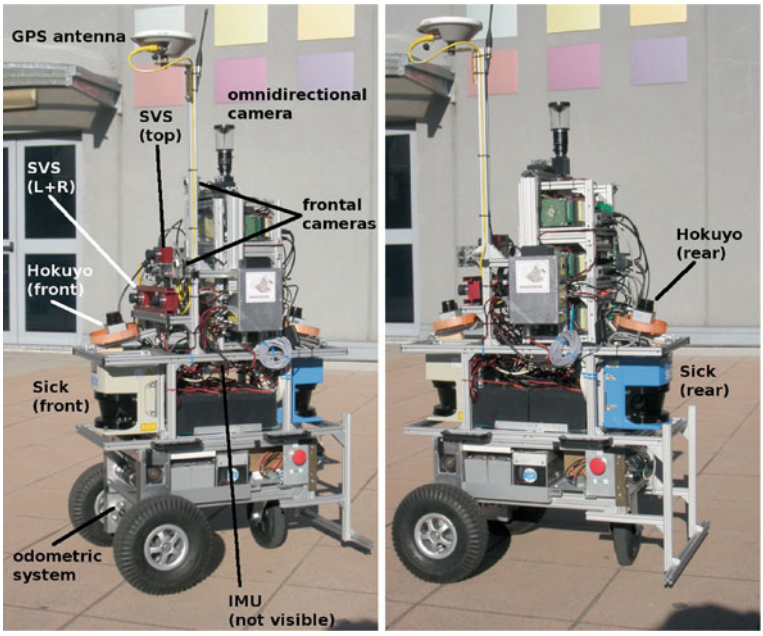

Fig. 4.1 RAWSEEDS data acquisition robot (outdoor setup). SVS is the trinocular camera system, Hokuyo and Sick are the laser range scanners

For outdoor acquisition sessions, the sonar belt was not used (its limited range made it useless), while a GPS receiver and its satellite antenna were added to the robot. Though physically contiguous, the GPS unit was not conceptually part of the sensor suite: it belonged, instead, to the ground truth collection system which will be described in Sect. 4.3. This preserved the independence of the GT collection system from the sensors used for data collection: an important methodological point.

For what concerns calibration, RAWSEEDS provides, as part of the datasets, a full description of the calibration procedures used for each sensor. In the case of cameras, the test images used for calibration are included as well, so that camera calibration can be verified and/or redone by the user. Figure 4.1 shows the RAWSEEDS robot, set up for outdoor data acquisition.

4.2.2 Locations

To avoid resource dispersion, we decided from the beginning of the RAWSEEDS project to perform acquisitions in urban environments only. Within this scope, however, we tried to include a wide set of locations, covering different kinds of environments. RAWSEEDS Benchmarking Toolkit includes datasets recorded in indoor,

Fig. 4.2 A typical view of the Bovisa location (*left*), and an aerial view showing the path followed by the robot during the collection of one of the mixed (i.e., indoor + outdoor) datasets (*right*)

outdoor and mixed (i.e., partially indoor and partially outdoor) urban locations; in natural and artificial light conditions; and in static and dynamic (i.e., with moving objects and people) conditions. Here follows a brief description of the locations featured in RAWSEEDS datasets.

Bovisa. This location is a refurbished factory site, and was used for outdoor and mixed (i.e., outdoor + indoor) datasets. It has a very composite nature, which closely mimics that of a small town or factory area, with buildings separated by roads with sidewalks, complete with parked (and occasionally moving) cars. The dynamic datasets include large quantities of people walking, standing and sitting. The Bovisa location comprises buildings of different kind and style, as well as a wide range of features such as slopes, passages of various widths, external stairs and so on. Robot explorations covered both the outside and the inside of the buildings. Figure 4.2 illustrates the location.

Bicocca. This location was used for indoor datasets only. Corresponding floors of two buildings were involved: an office area in the first building and a library area in the second. The two buildings are interconnected by two glass-walled, roofed bridges, each about 20 m long. The main features of the Bicocca location include: corridors with doors on their sides (some of the doors are deeply recessed within the walls, so corridor boundaries are far from planar); hallways, sporting features such as tables and chairs, columns, staircases and escalators; the two bridges already described; a rather large and architecturally varied library; various kinds of doors and passages. Terrain is very smooth and exclusively horizontal, the only exception being short ramps at one end of the bridges. Figure 4.3 illustrates the location.

Each RAWSEEDS dataset is composed of data collected by the robot while following multiple paths through the environment. Such paths partially overlap and are organized into loops, to trigger the "loop closure" feature of SLAM algorithms. One of the key performance elements for such algorithms is, in fact, the ability to correctly detect that the zone presently explored has been visited before, and to update the map accordingly.

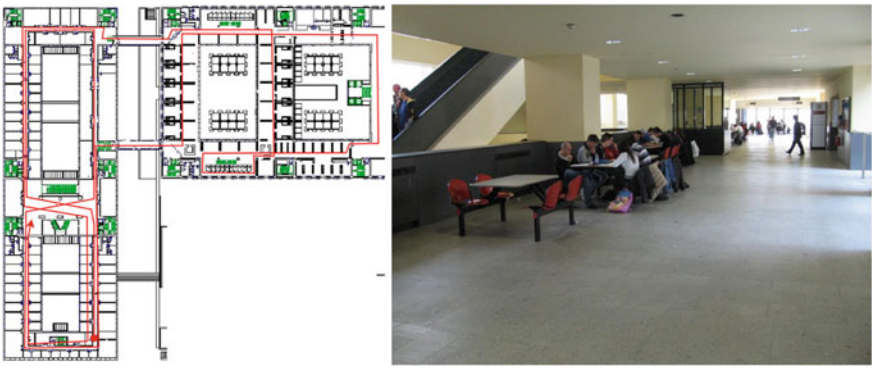

Fig. 4.3 A map of the Bicocca location (*left*), and a typical view (*right*). The map shows the path followed by the RAWSEEDS robot during the collection of one of the datasets

4.2.3 Data Validation and Distribution

All the datasets were subjected to a thorough validation process before making them public, to ensure their validity as the basis for high-quality benchmarking tools. Such process was deemed necessary, notwithstanding the care devoted to the data collection setup and organization, due to the high complexity of the data acquisition hardware and the wide range of environmental conditions. Of all the datasets acquired, those which were considered to be of insufficient quality were discarded. The availability of analyses and certifications of data quality, according to well-specified public criteria, is one of the methodological qualifying points of RAWSEEDS when compared to other available datasets for robotics, which were collected in relation to specific experimental papers.

The data validation process took into consideration several aspects. Specific attention was given to the suitability of the datasets for SLAM algorithms, to ensure the actual usability of the data in the research context to which they were addressed. Specifically, the validation process was aimed at certifying the quality of the following properties of the data: file format; file readability; presence and reliability of timestamp on each data sample; correct time synchronization between data streams; sufficient overlap between successive samples from the same sensor (necessary for tracking of environmental features); density and quality of the data. To ensure that the verification criteria were fully consistent with the intended use of the data, they were checked by actually using them as input for suitable algorithms (feature extraction algorithms for visual data, scan matching algorithms for laser data).

RAWSEEDS Benchmarking Toolkit is freely available trough the project's website http://www.rawseeds.org/, along with the additional documentation needed to make full use of it. Due to the overall mass of downloadable data (around 1.5 TBytes) and to the presence of very large single files (up to 50 GB), the distribution method chosen by RAWSEEDS is very unusual for a scientific dataset: i.e., a

peer-to-peer architecture based on the BitTorrent protocol. This ensures robustness and short download times, even under heavy load, without requiring a complex and costly server structure.

4.3 Ground Truth

The availability (or not) of ground truth data is one of the most important features of a dataset intended for benchmarking. In general, providing the ground truth (GT) requires the availability of data that describe the performance of the system without depending on the aspects of the system which are under evaluation. In the case of robots, this means that ground truth must not rely on the same sensor data used by the system for the task under test: therefore separate sensor systems must be deployed. For what concerns errors in the GT, if they are not negligible, they must be known: in this way, their effect on the assessment of system performance can be evaluated and discounted.

Unfortunately, GT collection, especially in autonomous robotics, is difficult and costly, and it was even more in 2006 (when RAWSEEDS collected its datasets).[3] This is the reason why, in robotics, experimental data are seldom accompanied by some form of ground truth. In the case of RAWSEEDS, GT collection accounted for a significant part of the overall effort devoted to setting up and performing data acquisitions.

The ground truth attached to RAWSEEDS datasets takes the forms of maps and (portions of) trajectories. As described by the following of this section, for outdoor datasets GT trajectories were generated with a two-unit RTK GPS system, while for indoor datasets a completely different approach was used (GPS signals are blocked by walls and roofs). For what concerns GT maps, they consist of executive drawings (in the form of CAD files) both for indoor and outdoor environments.

4.3.1 Ground Truth Collection in Outdoor Environments

For outdoor GT collection, RAWSEEDS chose to rely on the established GPS system to gather GT data. Single-receiver GPS systems are insufficiently precise, so a two-unit GPS system (belonging to the class of differential GPS, or DGPS) was required. In our setup, a spatially fixed receiver (*base station*) was used to generate correction signals that are then relayed to the GPS unit to be localized (mounted on the robot and called *rover*) using a dedicated radio link. Conventional DGPS systems estimate location using time shifts between the payloads carried by satellite radio signals: unfortunately, the accuracy afforded by this approach is still insuf-

[3] Nowadays the availability of systems for *motion capture* provides a means to precisely acquire the trajectory of a robot. This is still an expensive technology, especially when the capture area is large; however, it is readily available on the market and prices are getting lower.

ficient for an accurate description of robot trajectories. For this reason RAWSEEDS chose a more advanced RTK (Real Time Kinematic) GPS system (several tutorials on RTK GPS systems are available online, such as [5]). These systems use the disalignment between satellite radio carrier signals to estimate timing error, thus yielding greater precision. The price for better precision is paid in terms of increased complexity, difficulty of setup, and price. Under optimal conditions, RTK GPS systems provide centimetre-level positioning errors, which is fully satisfactory for the needs of RAWSEEDS. In the following we will give an account of how frequently these conditions occur in practice.

Our practical experience with the RTK GPS system highlighted some critical issues, many of which are associated to the fact that we were operating in urban environments. In particular, satellite reception was critical, and antenna positioning was a key factor. In theory, a single GPS receiver only requires visibility of 3 GPS satellites to be able to provide a calculated position (called a "GPS fix"). In real-world conditions, due to limitations in the precision of time synchronization between receiver and satellites, visibility of 4 or more satellites is needed. This figure refers to "well-positioned" satellites, i.e., those that have a significant elevation above the horizon, which reduces the number of actually usable satellites. Figure 4.4 shows the RAWSEEDS base station and illustrates RTK GPS performance.

The most damaging effect to reception is due to buildings (and other high obstructions such as trees). As these block GPS signals, obtaining a reliable GPS fix in urban environments can be very difficult. For this reason RAWSEEDS base station was always placed on the top of tall buildings. With this setup, the typical number of (well-positioned) satellites visible from the base station ranged from 6 to 9. Unfortunately, the mobile GPS antenna of the rover was forcibly at ground level, which led

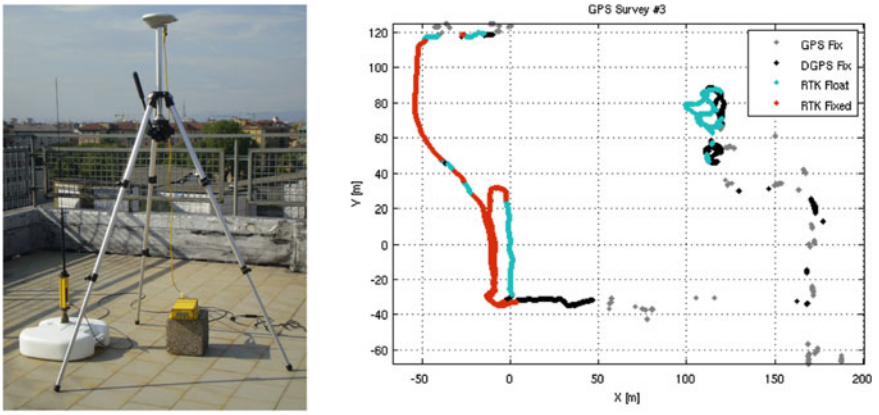

Fig. 4.4 *Left:* RTK GPS base station used by RAWSEEDS to capture ground truth data in outdoor environments. *Right*: a typical example of GPS data acquisition in urban environment. The portions where a best-quality GPS fix is available are the red ones; *light blue* corresponds to a *lower-quality* (but still usable for RAWSEEDS) fix. Building and trees in the immediate vicinity of the robot account for the insufficient quality of GPS data over significant portions of the trajectory

to frequent loss of satellite signals. Situations where the rover received less than the 4 satellites required for the most basic form of GPS fix were very frequent. Things were even more critical for RTK GPS operation, which requires that a minimum of 5 satellites must be available to the base station above the elevation threshold; and, even more critically, the *same* satellites must be visible to the rover. By placing the base station in such a way that all the sky was visible to it, the limiting factor was confined to satellite visibility by the rover.

Even considering its limitations, for RAWSEEDS the RTK GPS system proved to be an effective way to get precise trajectory data over a subset of the robot's complete trajectory. This was, indeed, how it was meant to be used by the project. On the other hand, our experience showed—confirming our expectations—that in outdoor urban environments the usability of GPS systems for robot localization is subject to heavy limitations, and that the use of GPS as a viable alternative method for odometry is questionable.

4.3.2 Ground Truth Collection in Indoor Environments

The collection of ground truth data indoors could not be based on GPS signals, which are blocked. A different approach was chosen: namely, the deployment, in a predefined area, of suitable systems dedicated to the reconstruction of the trajectory of the robot. Two separate systems for indoor GT collection were used in parallel: one based on cameras, the other based on laser range scanners. Their outputs were combined to allow the generation of better-quality GT data [6]. The main limitation of the indoor ground truth collection systems used by RAWSEEDS is their limited coverage. Ground truth data for robot trajectories is available only over a small portion of the environment explored by the robot; on the other hand, ground truth data for maps is available for the complete area covered by the datasets.

RAWSEEDS indoor datasets also include an *extended ground truth*, covering a much wider area. This has been obtained by applying scan matching algorithms to the output of two of the four laser range scanners on board the robot (namely, the Sick units). Therefore, the extended ground truth acts as fully valid GT for the assessment of any algorithm that does not make use of laser data from such sensors.

The vision-based ground truth collection system was based on a network of 5 cameras, with partially overlapping fields of view, covering an L-shaped portion of a wide hallway. Video from the cameras was processed by specially designed software to extract information about the trajectory of the robot (when visible). The cameras were positioned (at a height of around 3 m from the ground) on high poles. Notwithstanding the fact that the poles were fitted with rotation-blocking systems, rotation due to involuntary touches by passers-by, or even to thermal deformations, proved to be an issue: a fact to be considered by anyone planning to use a similar setup.

To allow reconstruction of the trajectory of the robot, the latter was fitted with visual tags and with a rectangular outer "shell" composed of vertical checkered

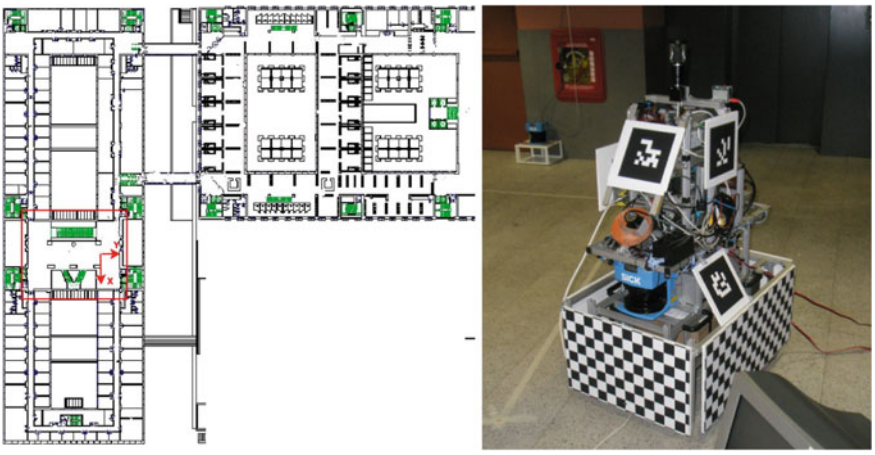

Fig. 4.5 The area of the Bicocca location where ground truth data were collected (*left*) and RAWSEEDS data acquisition robot, set up for indoor data acquisition (*right*). Behind the robot, one of the laser range scanners used by the laser-based GT collection system is visible (a similar device is on board the robot)

boards. This "shell" also had the effect of presenting a more regular shape to the fitting algorithms used by the laser-based GT collection system described below. Figure 4.5 shows Robocom fitted with the tags and "shell" used for indoor GT collection.

The laser-based ground truth collection system was more straightforward. We used a set of four Sick LMS200 laser range scanners positioned in suitable places along the perimeter of the wide hall where GT acquisition was performed. The sensors were carefully aligned so that their perception plane was horizontal, at the optimal height with respect to the robot's "shell" described above. The positions of the laser range scanners were the result of a tradeoff between the conflicting requirements of covering the largest possible area and ensuring that within such area the robot was always perceived by at least two sensors. The output of the laser scanners was processed by scan matching software.

4.4 Benchmark Problems and Benchmark Solutions

The scope of RAWSEEDS is not limited to providing datasets. While these can be immediately used by the community, the Benchmarking Toolkit also includes carefully chosen problems researchers could test their algorithms on, called Benchmark Problems (BPs). Additionally, ready-made solutions to these, based on state-of-the-art algorithms, are provided to act as references: these are called Benchmark Solutions (BSs).

From the methodological viewpoint, the main issue here is the definition of suitable, scientifically sound, *metrics* to evaluate the performance of an algorithm (e.g., a mapping algorithm). RAWSEEDS made an explicit effort to do that without forcing the algorithm developer to adopt a standardized representation for its output (e.g., a map). Such metrics, as we will see shortly, are an integral part of the BP; they are directly applicable without requiring data conversion, thus widening the appeal of RAWSEEDS Benchmark Problems. Of course, given this strict requirement, retaining physical significance for the metrics has been much more difficult; time will tell if such goal has been attained.

In the following of this section we will define precisely BPs and BSs, and will sketch how RAWSEEDS metrics are designed. For what concerns the metrics, special attention will be given to their relation with the ground truth.

4.4.1 Basic Definitions

A **Benchmark Problem (BP)** is the precise description of a type of problem (e.g., "perform SLAM"). The key element of a BP is that it includes not only the definition of a task, but also a set of well-defined performance metrics to assess the output of solutions. By using such metrics, a quantitative evaluation of the quality of the solutions can be done, and different solutions can be quantitatively compared. More precisely, a Benchmark Problem is defined as the union of: (i) the detailed and unambiguous description of a task; (ii) the specifications for a collection of multisensor data, gathered through experimental activity, to be used as the input for the execution of the task; (iii) a rating methodology (i.e., a set of metrics) for the evaluation of the results of the task execution, based on the use of ground truth (GT) data associated to the sensor data. A **Benchmark Problem instance (BP instance)** is obtained by combining a BP with: (iv) one of RAWSEEDS datasets to be used to execute the task specified by the BP.

A solution to a BP instance is called a **Benchmark Solution (BS)** and is defined as the union of: (i) a BP instance; (ii) the detailed description (and, optionally, code) of an algorithm for the solution of the BP instance; (iii) the output of the algorithm when applied to the BP instance; (iv) the values of the metrics specified by the BP, when applied to such output. A BP instance admits any number of different BSs, while a BS is tied to one specific BP instance.

Depending on the BP, the output of a BS can include the map of an environment, the trajectory of the robot, or both. As anticipated, one important point is that the actual representation of both the map and the trajectory are never specified by the BP: this means that any algorithm capable to solve a BP can be easily turned into a RAWSEEDS BS for that BP. The evaluation metrics, as well, are defined so to be computed independently of the actual representation of the output of the BS. In this way, BSs using different representations for maps and trajectories can be compared on performance grounds, which is a very important feature for a benchmarking tool.

4.4.2 Performance Metrics

As previously said, an integral part of each Benchmark Problem instance is a set of representation-independent performance metrics to be applied to its solutions (Benchmark Solutions). The descriptions of RAWSEEDS' metrics are *operative*, in the sense that they are intended to be directly applied as an algorithm for their calculation.

The metrics designed by RAWSEEDS to be part of the Benchmark Problems are directly connected to the field of application of the RAWSEEDS Benchmarking Toolkit, i.e., self-localization, mapping, and SLAM. Here follows a short description of them.

Mapping Error. The Mapping Error is intended as a measure of the accuracy of a reconstructed map. It requires that the author of the BS selects a set of environmental features of the map produced by the BS, identifies the corresponding features of the ground truth map, and performs distance calculations within pairs of features belonging to the first set of features and (separately) to the second set of features. A comparison between the results of the distance evaluations performed on the first and second set of features yields the value of the Mapping Error.

Loop Closure Error. This metric is intended to capture the localization accuracy of a SLAM algorithm when it cannot rely on the realigning mechanism called "loop closure", which is triggered when a SLAM algorithm detects that the area where the robot is currently located is one of those already explored by it in the past. This allows to evaluate the quality of the internal map produced by a SLAM algorithm. By eliminating the correction effect of loop closure, error mechanisms that otherwise could be masked are highlighted. The evaluation of the Loop Closure Error requires that the author of the Benchmark Solution disables the loop closure mechanism. Fortunately, considering the way SLAM algorithms are usually implemented this is often a trivial task. Actual computation of the Loop Closure Error is done by comparing the reconstructed and actual (i.e., coming from the ground truth trajectory) pose of the robot at a time instant specified by the BP.

Self-Localization Error. This metric is aimed at evaluating the accuracy of the estimate that a robot produces of its own pose within the environment. Given that such estimate usually refers to a map that is itself reconstructed by the robot, computation of the Self-Localization Error requires that such map is aligned with the ground truth map. This can be done manually, or with any other method, provided that a full description of it is given. When the two maps are aligned, Self-Localization Error is computed by processing the distance errors between the estimated pose of the robot and the corresponding pose from the ground truth trajectory, for each time instant where the latter is available.

Integral Trajectory Error. Like the Self-Localization Error, this is a metric that intends to capture the accuracy of pose reconstruction. Computation is similar, but while Self-Localization Error focuses on instantaneous accuracy, Integral Trajectory Error focuses on the overall distance between the reconstructed

and ground truth trajectories (where the latter are available) over the whole path of the robot.

4.5 Conclusion

In a context where the need for benchmarks in robotics is widely perceived but rarely addressed, the RAWSEEDS project is an effort targeted towards fulfilling such need. RAWSEEDS Benchmarking Toolkit is a readily usable instrument to assess and compare algorithms for SLAM, localization, and mapping. Some of its aspects (e.g., strong multisensoriality, focus on ground truth, representation-independent metrics for algorithm assessment) are especially significant from both the methodological and the practical points of view.

Although RAWSEEDS Benchmarking Toolkit has been designed as to be open-ended and extensible, its main limitation lies into its scope. Its components are, in fact, only useful for groups working on the development of algorithms and software which are not involved into the *control* of robots. In fact, the datasets are pre-recorded and therefore not suitable to test algorithms that need to influence the movement of the robot, such as navigation modules. To create a benchmark for this kind of applications, a different approach based on suitable physical *test arenas* [7, 8] seems, at the moment, the only viable alternative.

References

1. Bonsignorio F, Hallam J, del Pobil AP (2007) Good experimental methodology - GEM guidelines. http://www.heronrobots.com/EuronGEMSig/downloads/GemSigGuidelinesBeta.pdf. Accessed July 2013
2. PerMIS (2010) Performance metrics for intelligent systems. http://www.nist.gov/el/isd/ks/permis.cfm, Accessed July 2013
3. Howard A, Roy N (2003) The robotics data set repository (radish). http://www.rawseeds.org/. Accessed July 2013
4. Rawseeds (2006) http://radish.sourceforge.net/. Accessed July 2013
5. Zinas N (2011) GPS network real time kinematic tutorial. Tech. Rep. TEKMON-001, Tekmon Geomatics LLP. http://tekmon.gr/tekmon-research/gps-network-rtk-tutorial/. Accessed July 2013
6. Ceriani S, Fontana G, Giusti A, Marzorati D, Matteucci M, Migliore D, Rizzi D, Sorrenti D, Taddei P (2009) Rawseeds ground truth collection systems for indoor self-localization and mapping. Auton Robot 27(4):353–371
7. Jacoff A, Messina E, Evans J (2002) Experiences in deploying test arenas for autonomous mobile robots. In: Proc 2011 PerMIS, Workshop, pp 87–94
8. Jacoff A, Messina E, Weiss BA, Tadokoro S, Nakagawa Y (2003) Test arenas and performance metrics for urban search and rescue robots. In: Proc 2003 IEEE/RSJ IROS, vol 3, pp 3396–3403