The RAWSEEDS Proposal for Representation-Independent Benchmarking of SLAM

Giulio Fontana and Matteo Matteucci and Domenico G. Sorrenti

Abstract— **Current SLAM algorithms output the world map in different formats, and the maps generated by different algorithms are not easily interchangeable in format. This is mainly due to the different inner functioning, including the sensing, of each algorithm. In our opinion this might affect the evaluation of SLAM algorithms. In order to make the evaluation criterion independent on the representation used for the world modeling, we propose to consider the SLAM outcome as finalized to executing other robot activities. The map per sè has therefore a limited value, its main quality, in our view, is its effectiveness in the contest of another robot task. At the same time, it has to be noted that the adoption of a particular SLAM algorithm for a given robotic application cannot base on a single metric. In comparing SLAM algorithms we need to take into account several aspects, e.g., the available sensor suite, the required accuracy, etc. If we consider the classical localization and navigation tasks, provided the rest remains unchanged, a SLAM algorithm is better than another as much as these tasks can be (measurably) better achieved. In the paper we describe the proposal developed in the FP6 RAWSEEDS project, giving also some methodological insights about the required ground truth and the score computation.**

I. Introduction

Progress in the field of autonomous mobile robotics requires that robotic systems gain the ability to operate with less and less direct human control, without detriment to their performance and, most importantly, to the safety of the people interacting with them. We are convinced that we will witness the birth of a new phase in the industrial development of the world, when robots will be able to safely navigate through environments designed for human beings, and to effectively execute tasks in those environments safely co-existing and cooperating with people.

A key factor for a rapid progress towards this "robotic spread" is a substantial advancement in the performances of moving in the working environment without collisions, while being able to reach a goal location; we see these as the basic abilities that a robot must necessarily possess to autonomously operate. This requires, in particular, the robot to be able to localize itself in the environment; this is usually achieved by introducing in some form an internal representation of the environment, i.e., the map. Both the position of the robot and of its goal position are then located on the map.

This work has been supported by the European Commission, Sixth Framework Programme, Information Society Technologies; EC Contract Number FP6-045144 (RAWSEEDS).

G. Fontana and M. Matteucci are with the Dept. of Electronics and Information, Politecnico di Milano, I-20133, Milan, Italy. {fontana,matteucci}@elet.polimi.it

D. G. Sorrenti is with the Dip. Informatica Sistemistica e Comunicazione, Università degli Studi di Milano - Bicocca, I-20126, Milan, Italy. sorrenti@disco.unimib.it

While applications only require the simpler ability to perform self-localization in known map, others might require the Simultaneous Localization And Mapping (SLAM) functionality [1], [2], [3]). Although these abilities are not sufficient to ensure the robot to be also able to execute a task, they can be thought of as necessary conditions for a mobile robot to be capable of effective autonomous behavior.

A huge amount of work has been done in SLAM, nevertheless little has been done for establishing a good methodology for its benchmarking. Some work, oriented to the more general issue of following Good Experimental Methodologies, took place in a Special Interest Group [4] of the EURON2 [5] EEC Network Of Excellence. Beside that effort, what used to happen is that research groups (mainly universities) collected their own data, to test the performances of their own algorithm, and then shared these datasets with the community [6], to foster the advancement of the research. In these cases, the produced datasets and tools have been widely accepted and used by the community, although they are limited in their usefulness: no ground truth is present, hence no grounded comparison is possible, and no uniform methodology is used.

Benchmarking and performance measurement are a key factor for the industrial development of robotics. The study, design, engineering and marketing of autonomous robotic systems and solutions relies on the fact that the actors involved (mainly, research groups and companies) possess, or can easily acquire, the tools for developing and testing sophisticated localization, mapping or SLAM algorithms. Such tools can be subdivided into the following categories:

- sensor datasets (or real facilities) for the testing of systems on real-world environments;
- benchmarks and methodologies for the quantitative evaluation and comparison of algorithm performance;
- proven algorithms, i.e., which have already demonstrated a successful performances, to be used both as starting points, to develop new solutions, and for comparison.

To be fully and readily useful, these elements should be integrated into a coherent benchmarking toolkit. This in turn requires: common and well-documented interfaces, immediate interoperability, extensive documentation, and accompanying support services. Presently, neither a toolkit of the kind described above nor its constituents are available to a generic potential users.

In the line of developing effective benchmarking toolkits, RAWSEEDS is a project funded by the European Commission, as part of the VI EU Framework Program, with the aim of defining a SLAM benchmarking toolkit. In a previous paper we described the structure of the RAWSEEDS toolkit [7],

in this paper we focus on the evaluation criteria. The paper presents an introduction to SLAM benchmarking in the next section, it then moves to the specific of our proposal, and then some conclusions are drawn.

II. The RAWSEEDS Benchmarking Toolkit

Advancement in any scientific and technical discipline relies on two basic mechanisms: competition between research groups, and exchange and dissemination of results among the research community. Both require that one research group could evaluate the results obtained by another research group in a quantitative way; this means also that the results obtained by the groups could be compared, in order to find the best solutions. In the context of RAWSEEDS, i.e., evaluating and comparing SLAM algorithms, this requires that:

1) algorithms are applied to the same data;
2) an evaluation methodology is defined.

As we already discussed, even the first of these two conditions is presently very rarely fulfilled. RAWSEEDS will offer a solution to this problem by providing comprehensive and validated multisensorial datasets. It is worth noticing that the many synchronized sensor streams, from exactly the same situations, allows not only to deal with multisensor fusion, but also, more relevant in our view, to quantitatively evaluate the performance of a sensing suite; which is very interesting for new companies entering the robotic field.

A benchmark can be defined as a standard problem to which any algorithm, in the considered class, can be applied, together with a set of rules to evaluate the output produced. RAWSEEDS will generate and publish the datasets needed to define Benchmark Problems (BPs) and Benchmark Solutions (BSs).

A Benchmark Problem (BP) is defined as the union of:

1) a detailed and unambiguous description of a task;
2) an extensive, detailed and validated collection of multisensorial data, gathered through experimental activity, to be used as the input for the execution of the task;
3) a rating methodology for the evaluation of the results of the task execution.

The application of the given methodology to the output of an algorithm or piece of software, designed to solve a Benchmark Problem, produces a set of scores that can be used to assess the performance of the algorithm or compare it with other algorithms.

A Benchmark Solution (BS) is defined as the union of:

1) a Benchmark Problem;
2) the detailed description of an algorithm for the solution of the BP (possibly including the source code of its implementation and/or executable code);
3) the complete output of the algorithm applied to the BP;
4) the set of scores of this output, obtained with the methodology specified in the BP.

The complete set of BPs and BSs published by RAWSEEDS is what, in this document, we call the "RAWSEEDS Benchmarking Toolkit". For instance, a Benchmark Problem may be a precise description of the task of extracting a map of an environment composed of line segments from the point-based representation of the environment produced by a laser range scanner, plus the complete scanner data recorded on location, plus the rating methodology to be applied to the results. The union of this BP with an algorithm solving the problem (and possibly a software implementation of it), its results, and their rating (obtained with the given methodology) may then be a BS.

The main use of a BP is to test existing (or in the course of development) algorithms. On the other hand, a BS can be very useful in many ways, as it might be used for:

- comparing the rating obtained by the algorithm included in the BS with the rating obtained by another algorithm, applied to the same BP (the rating methodology is defined in the BP itself, and so can be applied to different BSs);
- using the output of the algorithm included in the BS to get pre-processed input data for higher level algorithms, e.g., as path planners;
- using the algorithm included in the BS as a "building block" to design a complete multi-layered system for the processing of sensor data;
- using the algorithm included in the BS (and, if available, the source code of its implementation) as a source for the design of new, more sophisticated algorithms.

It must be noted that different BSs can be constructed for a single BP, so the number of BPs is not a limiting factor for the number of BSs that can be defined. Additionally, it is important to stress that the ratings of all the BSs based on the same BP will be directly comparable. The BSs defined as part of the RAWSEEDS Benchmarking Toolkit will use state-of-the-art, well-proven algorithms that will constitute a corpus of "standard solutions" for the BPs and for similar problems.

For the construction of the BPs, typical instances of different indoor and outdoor environments will be used, in both static (i.e., excluding moving elements such as people) and dynamic conditions. Each multisensor dataset will be collected moving the test robot, see Figure 1, through the environment on a complex exploratory path.

Each environment will be covered by multiple datasets, generated by performing exploration sessions on different paths with the same test robot; in this way it will be possible e.g., to use multiple datasets associated to the same environment to simulate a multi-robot dataset. In our view the inclusion of outdoor locations is particularly significant, since many research groups do not own robot platforms capable to navigate through unstructured terrain and thus research results in this field are very scarce, even if many possible scientific and commercial applications can be envisaged.

The raw collection high-resolution sensorial data is not sufficient to guarantee their precision and consistency, i.e. the fact that the data obtained from different sensor devices are coherent with each other, with the (logged) actions performed by the robot and with the physical environment explored. Moreover, advanced robotics applications require time coherence between different sensor data streams, which usually is

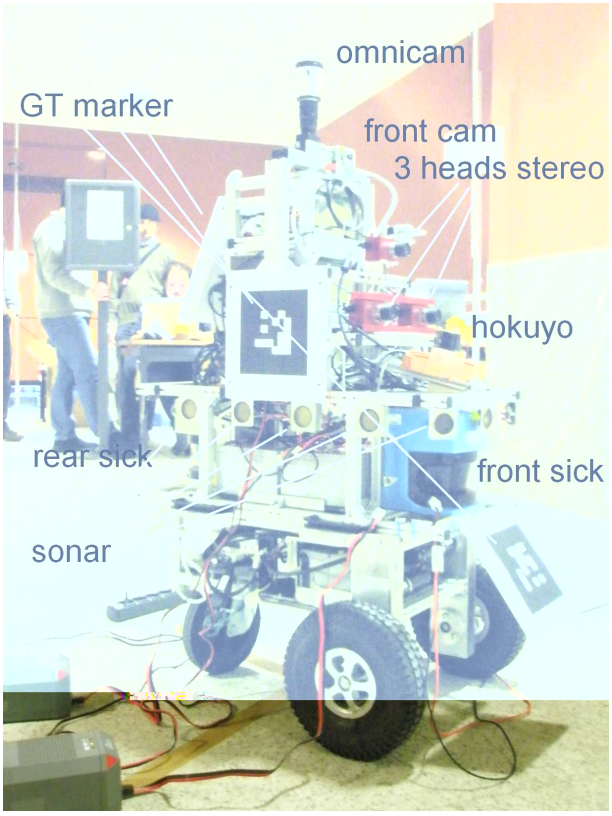

Fig. 1. robocomduringdatacollectionsv2.eps; the 3 b/w cameras can constitute different stereoheads

neither guaranteed nor verifiable. To overcome these problems RAWSEEDS will include an extensive data validation phase, with the aim of verifying and certifying the consistency of the data produced by each sensor and their coherence with the ground truth. Statistical analysis of the data against the ground truth will be performed during the validation process, and its results (e.g. noise levels, and distribution) added to the BPs.

It is extremely important to note that such rating is arbitrary: it depends on the specific dataset included in the BP and on the choice of rating methodology. However, it gives a rough way to compare different algorithms, which is something that has always been very difficult in robotics. We would like to stress that the purpose of RAWSEEDS is not that of compiling and publishing a ”hit parade” of the more successful algorithms for mobile robotics (which we hope people will submit to RAWSEEDS for publication in the form of BSs), nor to certificate the performance of algorithms, as the published ratings will always be measured by the authors of the BSs themselves. On the contrary, RAWSEEDS wants to contribute to the progress of robotics by publishing a set of instruments - the benchmarking toolkit - useful to develop, evaluate and perfect algorithms for mobile robotics. The fact that the evaluation ratings can, with attention to all the pitfalls associated to such an operation, be used to compare the performance of different algorithms when applied to the same data is certainly useful for research and development but is not, in any measure, the focus of RAWSEEDS' activity.

III. The Ground Truth Issue

One of the RAWSEEDS aims is to enable the evaluation of the performance of different algorithms; for this reason we need a joint collection of the datasets and the appropriate Ground Truth (GT in the following). Collection of the GT means collecting the real value for the variables to be estimated by the algorithms, i.e., position of walls, that will be then evaluated. In the cases where such values change in time, i.e., the robot pose with respect to the world reference, the collection has to take place at the same time of the sensor data collection. The GT, together with the data collected by the robot sensors, constitutes a Benchmark Problem (BP).

Of course, no device is available to measure “the real ground truth”, i.e. real position with zero error; instead, the best accurate ground truth estimate suitable for common robotics requirement will be provided. This estimate will be integrated with error bounds and/or confidence intervals to be properly compared with the accuracy of the proposed Benchmark Solutions (BSs). In the unfortunate case that the accuracy of the independent GT measuring device is (or in the time will become) not high enough, the ground truth will be built basing on the output of the best known algorithm.

Which independent GT collecting device can we devise, for the robot poses? In the indoor scenarios, a potentially interesting devices like D-GPS does not work properly and we therefore need to base on a different technology. Since we are currently considering continuous acquisition, the manual measurement approach is also not admissible, beside being error prone and very cumbersome. We designed and implemented a system, independent on the onboard sensors, based on a network of cameras. This system exploits a public available software tool, to locate the robot by locating some markers carried by the robot (visible in Figure 1), in known positions on the robot itself. We believe that this approach, which can be in short described as “GT just in a few places along the path”, i.e, where the cameras are, is realistic and good enough for many years forward (of GT usage).

For the map, on the other hand, we might rely on executive drawings, possibly integrated by hand measurements for those items (e.g., furniture), which are not in the executive drawings. Another option is to base on maps collected, independently from the robot sensors suite, by a human operator. On these hand-made maps the GT will be computed only on the relative position between pairs of well-defined environment landmarks, e.g., vertical edges, etc. This will allow comparison of maps obtained by the algorithm under evaluation w.r.t. the GT, in terms of reconstruction error of relative distances. A last option is to base the GT on maps produced by manually registering the data output by the most accurate sensors available on the robot, e.g., laser range scanners.

It has to be noted that some performance evaluation figures might not require GT, e.g., the pre loop-closure error. A question that might arise concerns the need for the GT to

be absolute or relative. Relative means that both the GT map and poses are referred to some previous robot frame. As the evaluation can be performed between pairs of map elements and/or poses, we deduce we do not need an absolute GT.

IV. Evaluation of SLAM

The output of a SLAM algorithm is a map of an environment. The most frequently mentioned evaluation methodology bases on the comparison of the reconstructed map with a "reference" map of the environment, which is included in the associated ground truth. This comparison is usually performed by comparing the position in the maps of specific landmarks, i.e., features that are both important for navigation and easy to identify, such as corners or borders of the walls. This (predefined) set of landmarks is chosen on the reference map, and then the same landmarks are searched in the reconstructed map: the ratings of the algorithm are then defined in terms of presence and correct positioning of the landmarks in the reconstructed map. Examples of such ratings might be the percentage of landmarks that can actually be identified in the reconstructed map, or the mean error obtained when comparing the distances between pairs of landmarks in the reconstructed map with the same distances evaluated in the reference map.

However, suppose that we have a Ground Truth accurate enough; in order to evaluate the performance of a SLAM approach we need to associate parts of the reconstructed map to the elements in the Ground Truth. Many different kind of maps exist: line segment maps, occupancy grid maps, and so on; a few examples, for the same environment, are presented in Figure 2, 3, 4. Therefore, when associating map elements to Ground Truth elements, it might turn out that some SLAM approach, just because of the map representation used, appears less performing than others.

An alternative might be to specify, in each BP, which kind of map is required from the solution algorithm, and of course it will be possible to define multiple BPs, differing only for the kind of map they require. We do not think this to be a convenient approach for the advancement of the research, as fixing the representations in the BPs stops the competition between the different representations. A BP should represent a real problem, without any bias, like fixing the map representation, the class of the algorithm, etc.

It is also possible to define, on such maps, SLAM-specific methodologies; an example is the evaluation of loop-closure error. When the dataset includes a loop, i.e., the trajectory of the robot returns to a previously visited point, a SLAM algorithm, which updates the map as the robot proceeds, has a means of correcting the errors on the estimated pose of the robot: in fact it possesses two different estimates of the robot's pose, in two different time instants, knowing that they must be coincident. Forcing this coincidence gives additional constraints on the trajectory of the robot and greatly reduces the errors due to imperfect odometry.

Moreover, as the features of the reconstructed map have an estimated position calculated simultaneously with the estimated trajectory of the robot, when the trajectory is corrected by "closing the loop", the map is subject to correction too, and its precision rises. Loop-closure error is the error between the estimated pose of the robot (or the position of some feature of the environment) when reaching the end of a loop, and the modified pose of the robot (or position of feature) after the correction due to the closure.

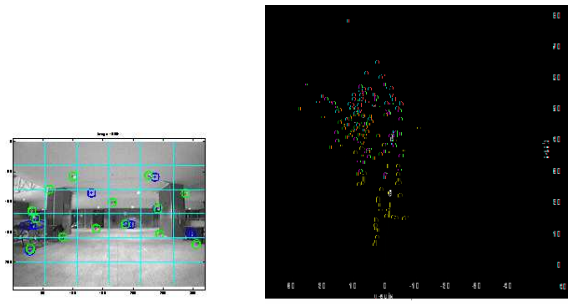

Fig. 2. Left: example of image features used in MonoSLAM. Right: Bird's view of a MonoSLAM map. Pictures courtesy of Univ. of Zaragoza, Spain

Fig. 3. A map obtained by integrating scans from a LRF. Picture courtesy of Univ. of Freiburg, Germany

As it can be easily observed, there are currently many possible ways for assessing the performance of a SLAM algorithm, among these we can mention:

1) Quantitative measures of path quality, w.r.t. GT;
2) Quantitative measures of map quality, w.r.t. GT;
3) Performance changes as the map size grows;
4) Quantitative measure of the estimation error, before loop closure;
5) Loop detection performance (false positives, false negatives, etc.).

It seems clear to us that there is not a single and agreed way to evaluate a SLAM algorithm, instead, like for many other performance evaluation efforts, see e.g., [8, 9], we will need

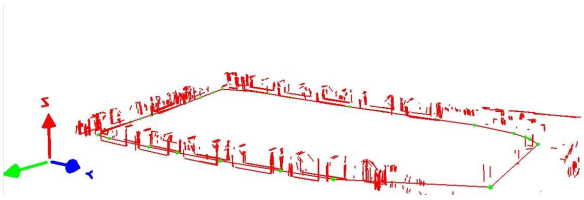

Fig. 4. A map obtained integrating 3D segments from a segment-based 3D stereo reconstruction system

to consider different criteria. Nevertheless, we believe there is still some missing point in SLAM performance evaluation so far: no one of the previous mentioned metrics does take in to consideration the task the robot should pursuit.

V. Task level benchmarking

In the end of the previous section we introduced an element of doubt about the fact that the current approaches to benchmarking are really optimal. In fact geometry-based methodologies measure the geometric quality of the map, when compared to a reference (i.e., correct and complete) map; in other words, they measure the capability of the mapping algorithm to produce good maps. It has to be noted, though, that many SLAM algorithm exist, and a relevant difference preventing comparison is in the way the map is represented; we call this the map representation issue.

If we stick to a pure geometry-based approach we need to solve the problem of matching different representations to the existing GT map. Finding these matches means: to identify a set of landmarks in the reference maps (e.g., corners), to find those landmarks by hand in the reconstructed representation, and to compute the errors. Here, just reverting the order and selecting the landmarks in the reconstructed representation instead of in the reference map, might turn into a different error score, and the same applies to the selection of the landmarks. Anyway, if there are enough landmarks, the score becomes representative.

It might appear that the problem with the different representations is that the process of matching to the GT might be not fair, with respect to the different approaches. For instance, we could obtain an occupancy grid from the SLAM system, to be matched to a 2D sparse map of points for GT, we have the grid quantization playing against. Similarly, if we receive 3D segments from the SLAM system, again to be matched to a 2D sparse map of points for GT, we might associate, after a 2D projection, a 3D segment endpoint with the wrong 2D point. However, it would be better if we could figure out a benchmarking metric that could provide natively an evaluation of the effectiveness of the map, without any need such always suboptimal matches.

Such representation independent benchmark would also be in agreement with the aim to ease the access to mobile robotic technologies from subjects, like SMEs, active in other fields. In such case, it is required to be able not only to compare the different SLAM algorithms, but also the sensing suites or, more interesting, the combined effect of each algorithm on a given sensor stream.

These observations motivate our apparently controversial point, i.e., that it is not the purpose of a mapping algorithm for robotics to obtain the best map, per sè. In our view, the purpose of such an algorithm is, instead, the creation of a map that, when used by a robot to navigate into the real environment allows the best performance of the robot. This requires the map to be good, for example, as an instrument for path planning and/or for obstacle avoidance. It is absolutely possible that, given a map with good geometric accuracy and one with a worse geometric accuracy, the second will lead to a much better performance of the robot in the real environment. For example, in planning, it is more important to know which passages (such as doors) exist between the rooms of an office than to know the exact position of each passage: a map which has perfect geometry, but representing a wall instead of the doors, could well lead to disastrous results, up to the incapacity to perform a given task. We therefore advocate the need to define new metrics for the evaluation of maps and of mapping algorithms, more closely related to the real objectives of those maps, i.e., their usage in robotic applications.

So our proposal for an alternative solution to SLAM benchmarking is the quantitative measure of the effectiveness in performing a certain (set of) mobile robotics task(s) based on the reconstructed map. In other words, we are not really interested in the amount of accuracy w.r.t. ground truth, provided we can plan, navigate, and localize in our map. Moreover, any representation is good, if it allows to a good performance in these task; similar considerations apply to the sensor suite: provided the robot can plan, navigate, and localize itself, it is irrelevant which sensor is used or better, the relevant dimension moves to other sensor features, e.g., power consumption, cost, etc. Summarizing: the definitive solution to SLAM benchmarking lays in the benchmarking of Planning, Navigation and Localization. The idea of task-level benchmarking has been mentioned already, by the authors in a previous paper [7], and partially also by Collins et al. in [10].

VI. A preliminary implementation, based on Localization

To be more practical we focus now on the localization task. We will evaluate the map built with a SLAM algorithm by measuring the quality of robot localization, within it, once the robot moves along a new path, in the same environment.

In this benchmark the first issue is that the dataset should comprise sensor data to perform SLAM, i.e., the robot should have moved adequately, revisiting enough times the same places, allowing ample gathering of the environment features. Different datasets can be envisaged here, by shortening the path and/or its informativeness, to change the task difficulty, and providing a better benchmark measure.

The path used for localization should include sensor streams, collected from the same robot in the very same environment, but during different paths. These paths might be gathered under different conditions, e.g., lighting, dynamic obstacles, etc. Note that ground truth for the robot pose is needed only for these paths. Different scores can be built, such as: how long it takes to get globally localized, basic statistics of the localization error under the different changing conditions, etc.

We realize that a few question opens: which localization algorithm? How can we trust this measure for being a measure of the SLAM algorithm performance? Is the representation affecting this measure? Our answers stem from the underlying idea of measuring the task-level effectiveness:

- the author(s) of the SLAM algorithm are the most appropriate person(s) for performing the selection of the localization algorithm. Of course they are given the extra-burden of this selection and implementation, but their selection cannot be accused of unfairness for the map representation, for the implementation of the localization algorithm itself, etc.
- The authors are well aware of the pros and cons of their SLAM algorithm and are in the best position for selecting the optimal combination with a localization algorithm, and this is in agreement with reaching the best task-level performance.
- The performance will measure the joint performance of the SLAM and the localization algorithm. But this is perfectly in agreement with measuring the task-level performance: the latter is the only relevant measure, our thesis is "the map quality per sé is not relevant".

There are two positive side-effects, the first is that no more map representation issue has to be faced, and secondly that we can compare different sensing suites and not only the algorithms.

On the other hand, a for us correct criticism is that the usage of localization, as the only mobile robotic task exploiting the map, is unfair; this is the reason for calling the proposal preliminary. We believe that each task should be scored. Then, anyone interested in the performance of SLAM algorithms should look at the task-oriented score that best suites his specific application.

VII. Conclusions

In this paper we propose to base the evaluation of SLAM algorithms on the measured performance of some algorithm implementing another mobile robotic task, e.g., localization. To avoid any unfairness, e.g., related to the map representation, in the selection of the specific algorithm for performing that robotic task, we propose to give the selection to the author(s) of the SLAM algorithm under evaluation.

This proposal has the advantage of allowing direct comparison between SLAM algorithms, something that is obviously interesting, although such interest is perhaps mostly academic. On the other hand, the proposal has also the advantage of allowing direct comparison between different sensor suites; this is, in our opinion, something currently missing for reaching a of widespread usage of mobile robotic technologies in the industry.

VIII. Acknowledgments

This work has been supported by the European Commission, Sixth Framework Programme, Information Society Technologies; EC Contract Number FP6-045144 (RAWSEEDS).

References

[1] F. Lu and E. Milios, *Globally consistent range scan alignment for environment mapping*, Autonomous Robots,1997.

[2] S. Thrun and M. Montemerlo and D. Koller and B. Wegbreit and J. Nieto and E. Nebot, *FastSLAM: An Efficient Solution to the Simultaneous Localization And Mapping Problem with Unknown Data Association*, Journal of Machine Learning Research, 2004.

[3] J. Folkesson and H. I. Christensen, *Robust SLAM*, 5th IFAC Symp. on Intelligent Autonomous Vehicles, 2004.

[4] website of the *Good Experimental Methodologies* SIG of EURON2, available at http://www.heronrobots.com/EuronGEMSig.

[5] website of the EEC-funded Network of Excellence *EURON* and *EURON2*, available at http://www.euron.org.

[6] A. Howard and N. Roy, *The Robotics Data Set Repository (Radish)*, available at http://radish.sourceforge.net.

[7] A. Bonarini and W. Burgard and G. Fontana and M. Matteucci and D. G. Sorrenti and J. D. Tardòs, *Rawseeds a project on SLAM benchmarking*, proceedings of the IROS'06 Workshop on Benchmarks in Robotics Research, available online at http://www.robot.uji.es/EURON/pdfs/Lecture_Notes_IROS06.pdf.

[8] D. Scharstein and R. Szeliski, *Stereo Vision Research Page* http://www.middlebury.edu/stereo, 2002.

[9] L. Prechelt, *Proben1 – A Set of Neural Network Benchmark Problems and Benchmarking Rules*, Technical Report 21/94 Universitat Karlsruhe, 1994

[10] T. Collins, J. J. Collins, C. Ryan, *Occupancy Grid Mapping: An Empirical Evaluation*, in Proceedings of Mediterranean Conference on Control and Automation, 2007