Proposals for benchmarking SLAM

G. Fontana, *M. Matteucci*, J. Neira, D.G. Sorrenti

Euron GEMBENCHForum 2008 - 25/26 March 2008 Prague

Today’s Special!

- GEM vs Benchmarking
- Two Lessons about Benchmarking
- Random thoughts in SLAM Benchmarking
- A commercial about RAWSEEDS (if we have time)
- Conclusions and final remarks
- Discussion … this is up to you!

What’s a Benchmark

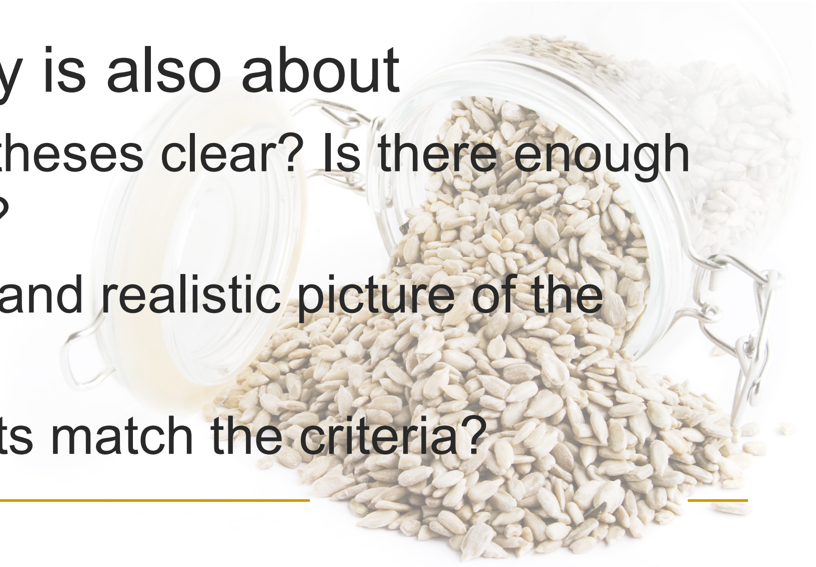
- *“Defining a standard benchmark for mobile service robots”* (The RoSta wiki – 2008)
  - Benchmark:
    - A standard by which something is evaluated or measured.
    - A surveyor's mark made on some stationary object and shown on a map; used as a reference point.
  - Benchmarking:
    - To measure the performance of an item relative to another similar item in an impartial scientific manner. (source: http://en.wiktionary.org/wiki/benchmark)
- A benchmark is a standard itself and second, benchmarking is a comparing measurement of performance.

GEM vs Benchmarking

- *"General Guidelines for Robotics Papers using Experiments"* (John Hallam – March 2008 DRAFT)
  - Is it an experimental paper?
  - Are the system assumptions/hypotheses clear?
  - Are the performance criteria spelled out explicitly?
  - What is being measured and how?
  - Do the methods and measurements match the criteria?
  - Is there enough information to reproduce the work?
  - Do the results obtained give a fair and realistic picture of the system being studied?
  - Are the drawn conclusions precise and valid?

GEM vs Benchmarking

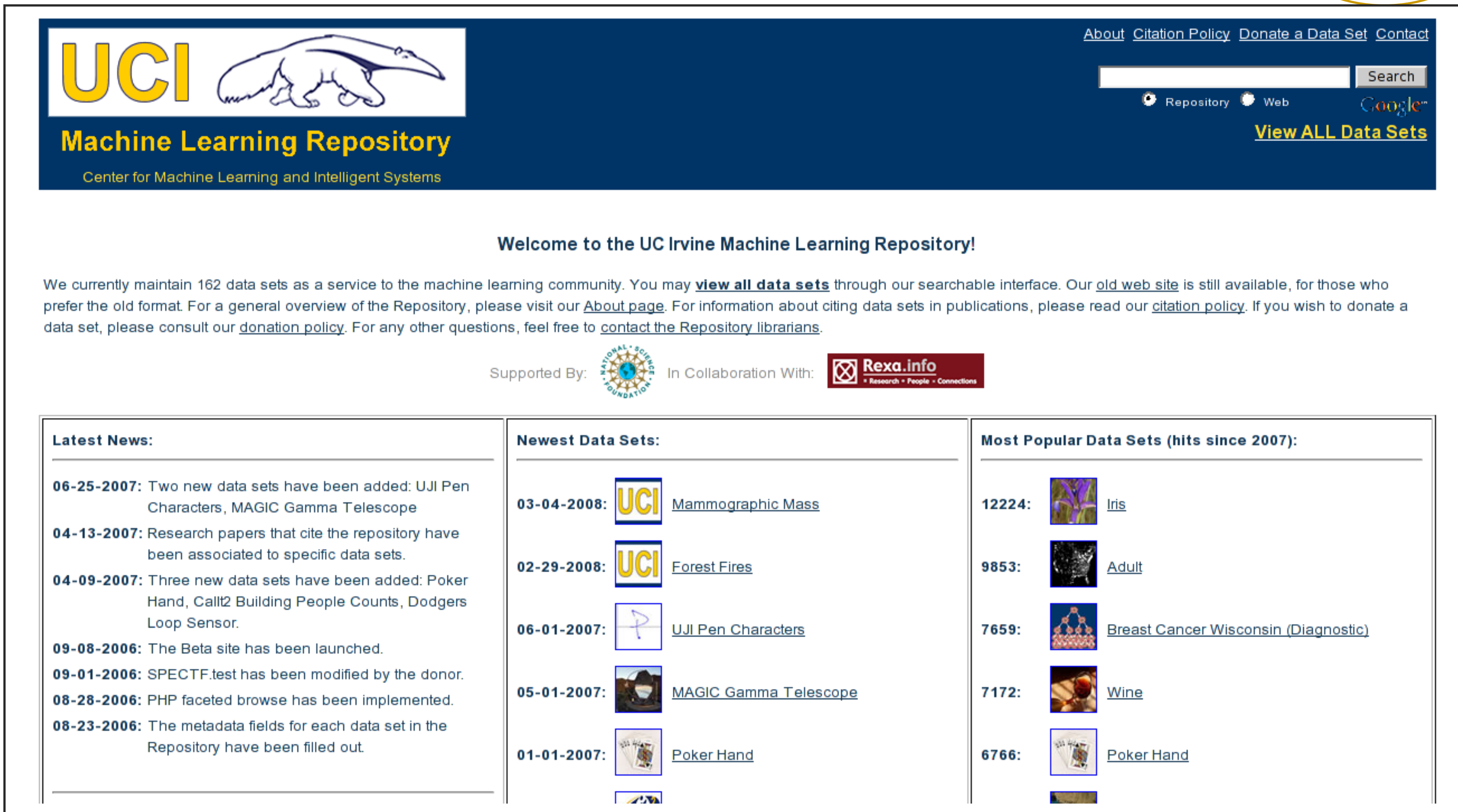
- Is a benchmark enough to state we are following GEM?
  - A benchmark forces us to use explicit (external) assumption/ hypothesis when performing system evaluation
  - Explicit performance criteria are part of a benchmark as well as the detailed definition of what is being measured and how
  - Benchmark aims at reproducing the results of system evaluation
- Good Experimental Methodology is also about
  - Are the system assumptions/hypotheses clear? Is there enough information to reproduce the work?
  - Do the results obtained give a fair and realistic picture of the system being studied?
  - Do the methods and measurements match the criteria?



Experiences to Imitate

- Research in Robotics is facing the themes of Good Experimental Methodology and Benchmarking rather late. Other fields in Computer Science have paved the way:
  - Machine Leaning @ UCI
  - Stereo vision @ Middlebury
  - Performance Evaluation of Tracking and Surveillance
  - PASCAL (object recognition database collection)
  - …
- What can/cannot be copied from those?
  - Machine Learning
  - Stereo Matching

Machine Learning @ UCI



Benchmarking Machine Learning

- *"PROBEN1 – A Set of Neural Network Benchmark Problems and Benchmarking Rules"* (Lutz Prechelt,1994)
  - A collection of problems for neural network learning in the realm of pattern classification and function approximation
  - Along with the datasets, Proben1 defines a set of rules for how to conduct and how to document neural network benchmarking.
  - The purpose of the problem and rule collection is to give researchers easy access to data for the evaluation of their algorithms and networks and to make direct comparison of the published results feasible.
- Delve datasets and utilities …

Stereo @ Middlebury (I)

vision.middlebury.edu

stereo • mview • MRF • flow

Stereo Evaluation • Datasets • Code • Submit

Daniel Scharstein • Richard Szeliski

Welcome to the Middlebury Stereo Vision Page, formerly located at www.middlebury.edu/stereo. This website accompanies our taxonomy and comparison of two-frame stereo correspondence algorithms [1]. It contains:

- An on-line evaluation of current algorithms
- Many stereo datasets with ground-truth disparities
- Our stereo correspondence software
- An on-line submission script that allows you to evaluate your stereo algorithm in our framework

How to cite the materials on this website:
We grant permission to use and publish all images and numerical results on this website. If you report performance results, we request that you cite our paper [1]. Instructions on how to cite our datasets are listed on the datasets page. If you want to cite this website, please use the URL "**vision.middlebury.edu/stereo/**".

References:

[1] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.
International Journal of Computer Vision, 47(1/2/3):7-42, April-June 2002.
Microsoft Research Technical Report MSR-TR-2001-81, November 2001.

Support for this work was provided in part by NSF CAREER grant 9984485 and NSF grant IIS-0413169. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Stereo @ Middlebury (II)

vision.middlebury.edu

stereo • mview • MRF • flow

Stereo Evaluation • **Datasets** • Code • Submit

Middlebury Stereo Datasets

2001 datasets - 6 datasets of piecewise planar scenes [1]
(Sawtooth, Venus, Bull, Poster, Barn1, Barn2)

2003 datasets - 2 datasets with ground truth obtained using structured light [2]
(Cones, Teddy)

2005 datasets - 9 datasets obtained using the technique of [2], published in [3, 4]
(Art, Books, Dolls, Laundry, Moebius, Reindeer, Computer, Drumsticks, Dwarves)

2006 datasets - 21 datasets obtained using the technique of [2], published in [3, 4]
(Aloe, Baby1-3, Bowling1-2, Cloth1-4, Flowerpots, Lampshade1-2, Midd1-2, Monopoly, Plastic, Rocks1-2, Wood1-2)

How to cite our datasets:
We grant permission to use and publish all images and disparity maps on this website. However, if you use our datasets, we request that you cite the appropriate paper(s): [1] for the 2001 datasets, [2] for the 2003 datasets, and [3] or [4] for the 2005 and 2006 datasets.

References:

[1] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.
International Journal of Computer Vision, 47(1/2/3):7-42, April-June 2002.

[2] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light.
In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, volume 1, pages 195-202, Madison, WI, June 2003.

[3] D. Scharstein and C. Pal. Learning conditional random fields for stereo.
In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, MN, June 2007.

[4] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching.
In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, MN, June 2007.

Stereo @ Middlebury (III)

vision.middlebury.edu

stereo • mview • MRF • flow

Stereo | Evaluation • Datasets • Code • Submit

Middlebury Stereo Evaluation - Version 2

New features and main differences to version 1.
Submit and evaluate your own results.

☐ Open a new window for each link

Error Threshold = 1		Sort by nonocc			Sort by all			Sort by disc					
Algorithm	Avg. Rank	Tsukuba ground truth nonocc	all	disc	Venus ground truth nonocc	all	disc	Teddy ground truth nonocc	all	disc	Cones ground truth nonocc	all	disc
AdaptingBP [17]	2.8	1.11 6	1.37 3	5.79 7	**0.10** 1	0.21 2	**1.44** 1	4.22 4	7.06 2	11.8 4	**2.48** 1	7.92 2	**7.32**
DoubleBP2 [35]	2.8	**0.88** 1	**1.29** 1	**4.76** 1	0.13 3	0.45 5	1.87 5	3.53 2	8.30 3	**9.63** 1	2.90 3	8.78 7	7.79
DoubleBP [15]	4.8	0.88 2	1.29 2	4.76 2	0.14 5	0.60 12	2.00 7	3.55 3	8.71 5	9.70 2	2.90 4	9.24 10	7.80
SubPixDoubleBP [30]	5.5	1.24 10	1.76 13	5.98 8	0.12 2	0.46 6	1.74 4	**3.45** 1	8.38 4	10.0 3	2.93 5	8.73 6	7.91
AdaptOvrSegBP [33]	9.5	1.69 21	2.04 20	5.64 6	0.14 4	**0.20** 1	1.47 2	7.04 13	11.1 7	16.4 11	3.60 10	8.96 9	8.84
PlaneFitBP [32]	10.4	0.97 5	1.83 14	5.26 5	0.17 7	0.51 7	1.71 3	6.65 9	12.1 12	14.7 7	4.17 19	10.7 19	10.6
SymBP+occ [7]	10.6	0.97 4	1.75 12	5.09 4	0.16 6	0.33 3	2.19 8	6.47 8	10.7 6	17.0 14	4.79 23	10.7 20	10.9
AdaptDispCalib [36]	11.2	1.19 8	1.42 4	6.15 9	0.23 9	0.34 4	2.50 11	7.80 18	13.6 20	17.3 16	3.62 11	9.33 11	9.72
Segm+visib [4]	11.5	1.30 15	1.57 5	6.92 18	0.79 19	1.06 17	6.76 20	5.00 5	**6.54** 1	12.3 5	3.72 12	8.62 5	10.2
C-SemiGlob [19]	11.8	2.61 28	3.29 23	9.89 25	0.25 11	0.57 9	3.24 14	5.14 6	11.8 8	13.0 6	2.77 2	8.35 4	8.20
SO+borders [29]	12.2	1.29 14	1.71 9	6.83 15	0.25 12	0.53 8	2.26 9	7.02 12	12.2 13	16.3 9	3.90 14	9.85 15	10.2
DistinctSM [27]	13.5	1.21 9	1.75 11	6.39 11	0.35 13	0.69 15	2.63 13	7.45 17	13.0 16	18.1 18	3.91 15	9.91 17	8.32
OverSegmBP [26]	13.7	1.69 22	1.97 17	8.47 22	0.51 16	0.68 14	4.69 17	6.74 10	11.9 11	15.8 8	3.19 8	8.81 8	8.89
CostRelax [11]	26.9	4.76 34	6.08 33	20.3 36	1.41 26	2.48 27	18.5 32	8.18 22	15.9 26	23.8 30	3.91 16	10.2 18	11.8 23
ReliabilityDP [13]	27.9	1.36 16	3.39 24	7.25 19	2.35 31	3.48 32	12.2 29	9.82 28	16.9 27	19.5 26	12.9 36	19.9 35	19.7 32
TreeDP [8]	28.6	1.99 24	2.84 22	9.96 26	1.41 25	2.10 25	7.74 23	15.9 34	23.9 34	27.1 35	10.0 32	18.3 32	18.9 31
GC [1d]	29.3	1.94 23	4.12 28	9.39 24	1.79 29	3.44 31	8.75 24	16.5 35	25.0 36	24.9 32	7.70 30	18.2 31	15.3 29
DP [1b]	32.9	4.12 32	5.04 32	12.0 30	10.1 39	11.0 39	21.0 34	14.0 31	21.6 31	20.6 28	10.5 33	19.1 33	21.1 33
PhaseBased [31]	34.2	4.26 33	6.53 34	15.4 34	6.71 35	8.16 35	26.4 37	14.5 32	23.1 32	25.5 33	10.8 35	20.5 36	21.2 34
SSD+MF [1a]	34.6	5.23 37	7.07 35	24.1 37	3.74 33	5.16 33	11.9 28	16.5 36	24.8 35	32.9 37	10.6 34	19.8 34	26.3 36
STICA [16]	35.8	7.70 38	9.63 39	27.8 38	8.19 36	9.58 36	40.3 39	15.8 33	23.2 33	37.7 38	9.80 31	17.8 30	28.7 38
SO [1c]	36.3	5.08 36	7.22 37	12.2 31	9.44 38	10.9 38	21.9 35	19.9 38	28.2 39	26.3 34	13.0 37	22.8 38	22.3 35
PhaseDiff [23]	37.0	4.89 35	7.11 36	16.3 35	8.34 37	9.76 37	26.0 36	20.0 39	28.0 38	29.0 36	19.8 39	28.5 39	27.5 37
Infection [10]	37.4	7.95 39	9.54 38	28.9 39	4.41 34	5.53 34	31.7 38	17.7 37	25.1 37	44.4 39	14.3 38	21.3 37	38.0 39

References

[1] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 2002.
a - SSD + min-filter (i.e., shiftable windows), window size = 21
b - Dynamic programming, similar to Bobick and Intille (IJCV 1999)
c - Scanline optimization (1D optimization using horizontal smoothness terms)
d - Graph cuts using alpha-beta swaps (Boykov, Veksler, and Zabih, PAMI 2001)

[2] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. ICCV 2001.

[3] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. ECCV 2002.

[4] M. Bleyer and M. Gelautz. A layered stereo algorithm using image segmentation and global visibility constraints. ICIP 2004.

[5] L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. SIGGRAPH 2004.

[6] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. CVPR 2005, PAMI 30(2):328-341, 2008.

[7] J. Sun, Y. Li, S.B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. CVPR 2005.

[8] O. Veksler. Stereo correspondence by dynamic programming on a tree. CVPR 2005.

[9] P. Mordohai and G. Medioni. Stereo using monocular cues within the tensor voting framework. PAMI 28(6):968-982, 2006.

[10] G. Olague, F. Fernández, C. Pérez, and E. Lutton. The infection algorithm: an artificial epidemic approach for dense stereo correspondence. Artificial Life, 2006.

[11] R. Brockers, M. Hund, and B. Mertsching. Stereo vision using cost-relaxation with 3D support regions. Image and Vision Computing New Zealand (IVCNZ), 2005.

[12] K.-J. Yoon and I.-S. Kweon. Adaptive support-weight approach for correspondence search. PAMI 28(4):650-656, 2006.

[13] M. Gong and Y.-H. Yang. Near real-time reliable stereo matching using programmable graphics hardware. CVPR 2005.

[14] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nistér. High-quality real-time stereo using adaptive cost aggregation and dynamic programming. 3DPVT 2006.

[15] Q. Yáng, L. Wang, R. Yang, H. Stewénius, and D. Nistér. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. CVPR 2006.

[16] H. Audirac, A. Beloiarov, F. Núñez, and J. Villegas. Dense disparity map based on STICA algorithm. Expo Forestal, Mexico, 2005.

[17] A. Klaus, M. Sormann and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. ICPR 2006.

[18] C. Lei, J. Selzer, and Y. Yang. Region-tree based stereo using dynamic programming optimization. CVPR 2006.

Benchmarking SLAM

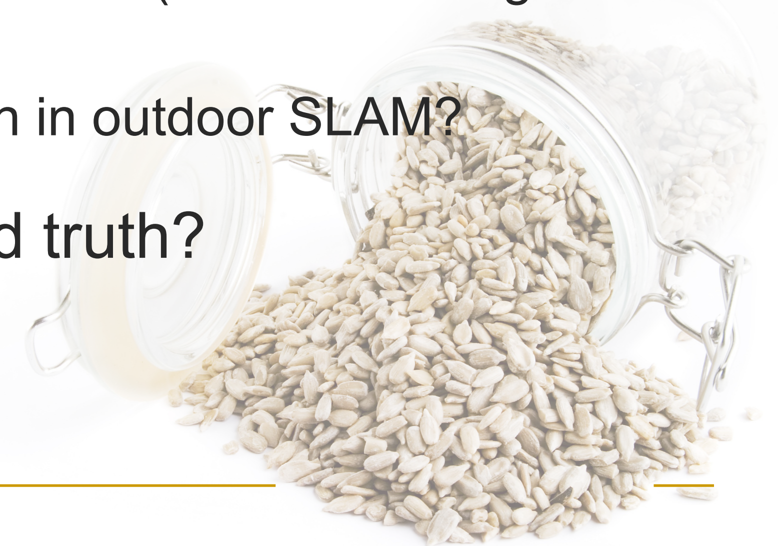
- Benchmarking of a fully fledged robotic application might be complex and hard to tackle as a whole …
- (Simultan ght be one of the eas s … provided:
  - We can
  - The co
  - The co
- A note: a xperimental!

How do we evaluate SLAM?

- To set up a benchmark for SLAM we need to define a way to asses the performance of a SLAM algorithm
  - Quantitative measures of map/path quality, w.r.t. ground truth
  - Performance variation as map size grows
  - How realistic/pessimistic/optimistic is the estimation error
  - Large loop recognition and closure
  - …
- It seems clear there is no single measure to evaluate SLAM, but we need to collect a set of measures plus we need ***ground truth!***

The Ground Truth Issue

- Quantitative measurements w.r.t. ground truth are subject to the precision of ground truth collecting device:
  - What is the reasonable precision we need in ground truth?
  - When facing indoor mapping, executive drawings might be a reasonable ground truth, but what about the robot path?
  - What is the accuracy required for the task (of course navigation is different from turning an handle).
  - Do we need RTK-GPS Ground Truth in outdoor SLAM?
- Can’t we get along without ground truth?
  - Large loop recognition and closure
  - Indirect ground truth computation …



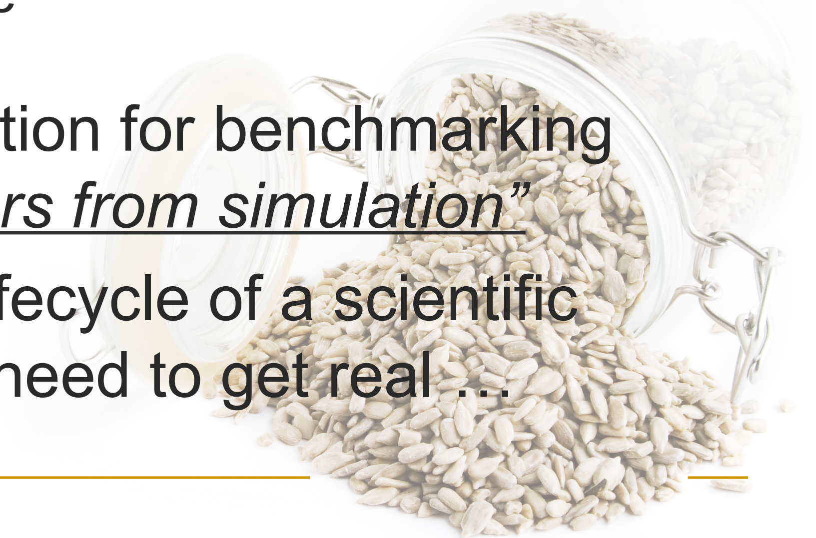
A Tricky Trick for Ground Truth

- *"Benchmarking Urban 6D SLAM"* (Wulf et al. – Benchmarking Workshop @ IROS 2007)
  - Highly accurate RTK-GPS receivers can not be used in outdoor urban areas
  - Surveyed maps can be obtained from the national land registry offices
  - Monte Carlo Localization can be used with such accurate maps to estimate ground truth positioning from the data and a manual supervision step to validate the MCL results.

- Isn't there a simpler solution?

A Simulated Solution

- *“Towards Quantitative Comparisons of Robot Algorithms: Experiences with SLAM in Simulation and Real World Systems”* (Balaguer et al. - Benchmarking @ IROS 2007)
  - Simulators can be available for free (almost)
  - Ground Truth is perfect and easy to collect ;-)
  - Experiments are "easy" to replicate
- Simulation seems to be the solution for benchmarking problems *“however real life differs from simulation”*
- Simulation is useful during the lifecycle of a scientific idea, but, at some point, robots need to get real …



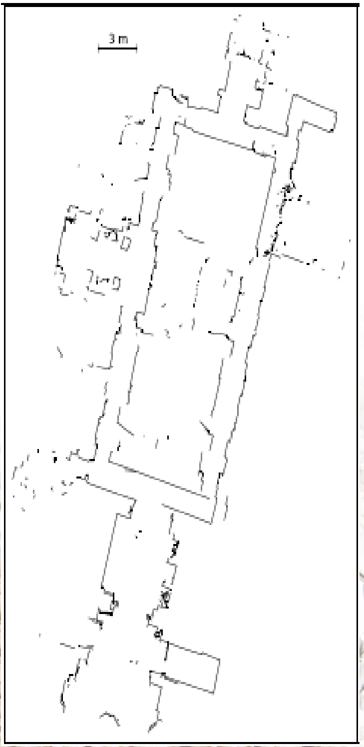
Robots Get Real!

- When robots become real, things get more cumbersome for development and benchmarking as well
  - Algorithms should be compared on the same real situations
  - Data should be provided for comparison (also the results!)
  - Ground truth should be collected and provided as well
- Publicly available Datasets become the solution
  - Freshly grained real data for all ;-)
  - Results are easy to replicate provided a Good Experimental Methodology is used
  - However most of them have no ground truth :-(

Segment Based Mapping (I)

- *"Good Experimental Methodologies for Robotic Mapping: A Proposal"* (Amigoni et al. – ICRA 2007)
  - The mapping system has to be applied to publicly available data.
  - The values of the parameters should be indicated.
  - Some experiments in which the mapping system does not perform well should be shown.

	method	public data	comparisons	env. size	# line segments	displacement	ground-truth	evaluation	parameter values	loop	proc. time
[4]	yes [5]	no	inferable	yes	yes	no	visual	yes	yes	yes
[6]	no	no	inferable	n.a.	n.a.	no	visual	yes	yes	n.a.
[7]	no	no	no	n.a.	yes	no	visual	n.a.	yes	n.a.
[8]	no	no	yes	n.a.	n.a.	yes	pose estimate	n.a.	no	n.a.
[9]	no	no	inferable	yes	inferable	no	visual	n.a.	no	yes
[10]	no	no	inferable	yes	inferable	simulated	visual and pose estimate	n.a.	yes	n.a.
[11]	no	no	yes	n.a.	yes	simulated	visual and pose estimate	n.a.	yes	n.a.
[12]	no	no	yes	n.a.	n.a.	no	visual and pose estimate	n.a.	no	yes
[13]	no	yes	inferable	n.a.	inferable	simulated	visual	n.a.	yes	n.a.
[14]	yes [5]	no	no	n.a.	inferable	no	visual	n.a.	yes	yes
[15]	no	yes	yes	yes	n.a.	yes	numerically w.r.t. ground-truth map	yes	yes	yes
[16]	no	no	inferable	yes	n.a.	yes	visual	yes	no	n.a.
[17]	no	no	inferable	n.a.	n.a.	no	visual	n.a.	no	n.a.
[18]	no	no	no	n.a.	n.a.	no	visual	n.a.	no	n.a.
[19]	no	no	no	n.a.	yes	no	pose estimate	n.a.	no	n.a.
[20]	no	no	inferable	n.a.	n.a.	no	visual	n.a.	yes	n.a.
[21]	no	no	no	n.a.	n.a.	yes	visual	n.a.	yes	yes
[22]	no	no	yes	n.a.	inferable	no	visual	n.a.	yes	yes
[23]	no	no	no	yes	n.a.	no	visual and pose estimate	n.a.	no	yes
[24]	no	no	yes	yes	n.a.	no	visual	n.a.	no	yes

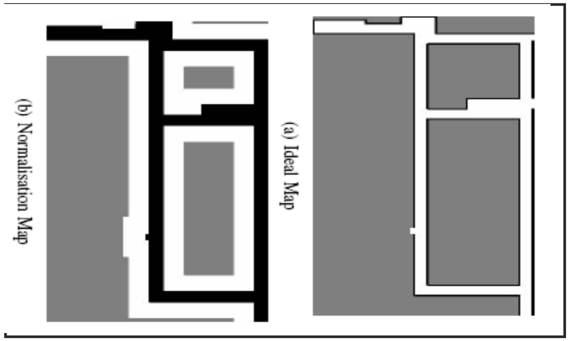


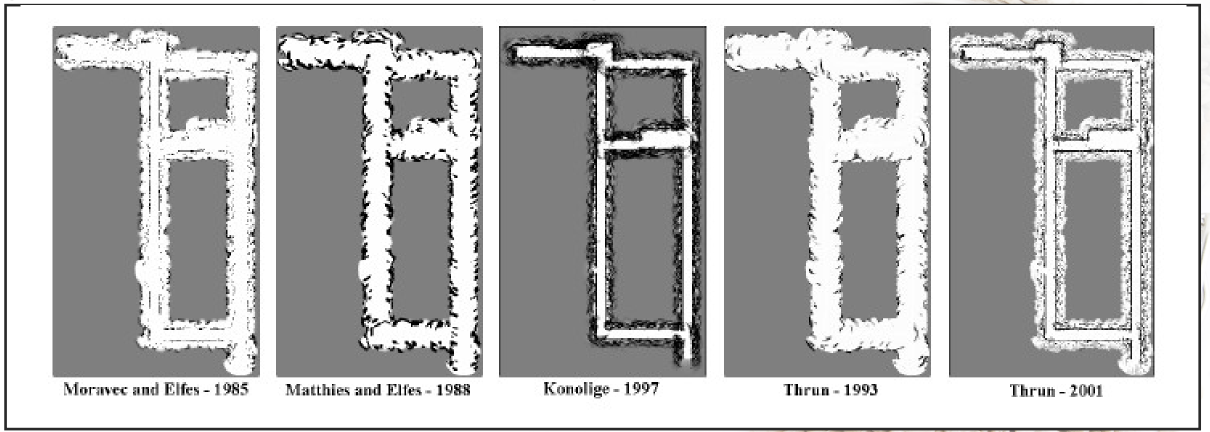
Segment Based Mapping (II)

- In order to *evaluate* and to *compare* different methods:
  - When a ground-truth map is available (this is not always the case), it should be used to assess the quality of the produced map, by evaluating its distance from the ground-truth map (e.g., according to the Hausdorff metric).
  - All the data about the produced maps should be clearly indicated (e.g., dimensions of mapped environment, resulting number of line segments, time required to build the map, …)
  - The behavior of the mapping system for different values of the parameters should be shown.
  - The map produced following a closed loop path in the environment should be shown, to evaluate the ability of the method not to “diverge”.

Grid Based Mapping

- *"Occupancy Grid Mapping: An Empirical Evaluation"* (Collins et al. – 2007)
  - An image comparison algorithm based on correlation
  - A direct comparison method called Map Scoring designed for probabilistic maps and a modified one ignoring free-space
  - A path analysis technique which tests the usefulness of a map as a means of navigation rather than treating it as a picture.





What if I’m different?

- Many SLAM algorithm exist and they differ in too many ways to be easily compared:
  - What if I’m using Occupancy Grid Representation instead of Segment Based Representation?
  - What if I’m working in a 3D world using 6DoF instead of moving in the classical 3DoF flatland?
  - What if I don’t have a laser scan or if my research is in SLAM with vision?
  - …
- Can’t we figure out a benchmarking procedure/metric that could take into account all these situations?

Fixing the Representation

- Recall the possible measures to assess the performance of a SLAM algorithm could be:
  - Quantitative measures of map/path quality, w.r.t. ground truth
  - Performance variation as map size grows
  - Large loop recognition and closure
  - …
- The tricky one is: map quality w.r.t. ground truth
  - Identify set of landmarks in the executive drawings (e.g., corners)
  - Find those landmarks by hand in you representation and compute the error
  - If they are enough, you have a lower bound on the “luck” you had in finding them ;-)

An alternative solution

- Quantitative measure of effectiveness in performing a certain (set of) mobile robotics task(s) based on that map!
  - We are not really interested in any accuracy w.r.t. ground truth provided we can plan, navigate, and localize in our map
  - Moreover any representation is OK for us if it allows these task, and who cares about the sensor if we can plan, navigate and localize :-)
- Here it comes the trick! The definitive SLAM benchmarking solution is benchmarking of Planning, Navigation and Localization :-P

SLAM Localization Benchmark

- Suppose we have the *data* and the *ground truth* from some paths in the environment:
  - Use a round or two in the environment to perform SLAM
  - Use the resulting map to localize with new data collected in the very same environment possibly on a different path, different light condition or even with people around.
  - We can measure how much it will take to localize, what's the localization error and the robustness to changing conditions.
- Pros and Cons:
  - We'll need to implement and provide a localization algorithm :-(
  - We'll be able to compare our algorithm with other representation or sensing suites ;-)

Any issue with this?

- Some issues could arise from this benchmark:
  - How much the localization algorithm influences the SLAM benchmarking?
  - Should we force all the people to use the same localization algorithm? How much it depends on the representation?
  - What's going on? Are we scoring the SLAM algorithm the Localization algorithm or the representation they use?
  - What if we have two different SLAM algorithm using different sensors, representation and different localization procedure?
- Who cares after all?
  - Being able to face the very last situation would be a success!
  - At some point we need to compare complete systems one against the other …

RAWSEEDS: Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets

Politecnico di Milano – Matteo Matteucci
University of Freiburg – Wolfram Burgard
Università di Milano-Bicocca – Domenico G. Sorrenti
Universidad de Zaragoza – Juan Domingo Tardos

What is RAWSEEDS ?

- EU Funded Project in the VI Frame Program from the 1st of November 2006 to April 2009
- A Specific Support Action to collect and publish a benchmarking toolkit for (S)LAM research
- Involved Institutions:
  - Politecnico di Milano (Italy – Coordinator)
  - Università di Milano-Bicocca (Italy – Partner)
  - University of Freiburg (Germany – Partner)
  - Universidad de Zaragoza (Spain – Partner)

Benchmarking Beyond Radish

- Nowadays we feel the lack of tools and methods to compare and evaluate market strength products. To aim at this we foster publishing of:

  - Extended multi-sensor data sets for the testing of systems on real-world scenarios
  - Benchmarks and methodologies for quantitative evaluation and comparison of algorithms/sensors
  - Off-the-shelf algorithms, with demonstrated performances, to be used for research bootstrap and comparison.

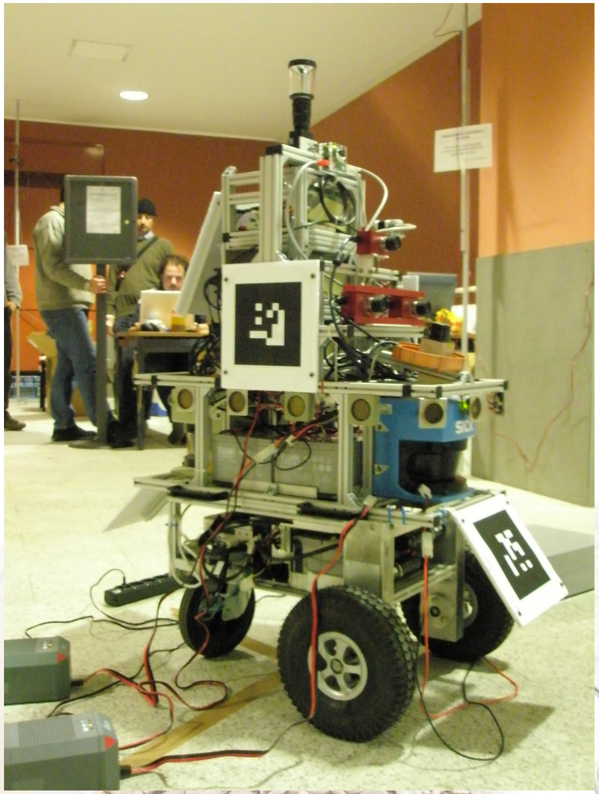
The RAWSEEDS Activities

- Definition and collection of benchmarks and methodologies for the assessment/comparison of algorithms for (S)LAM
- Creation of a website from which researchers and companies will be able to download these benchmarks, contribute new material and communicate with each other.
- Dissemination of knowledge about the RAWSEEDS benchmarks and the website

www.rawseeds.org

RAWSEEDS Sensor Suite

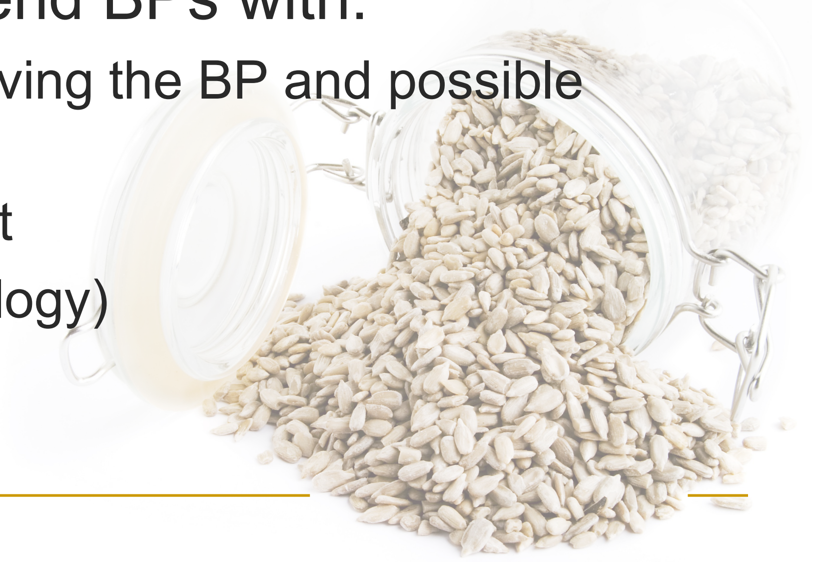
- Use of an extensive sensing suite
  - B/W + Color cameras (mono/stereo)
  - 3D cameras
  - LRFs (2D)
  - Omnidirectional camera
  - Sonars
  - GPS and D-GPS
  - Other proprioceptives (e.g., odometry, IMU)
- Sensors are synchronized and data acquired at maximum frequency



Benchmarks Problems & Solutions

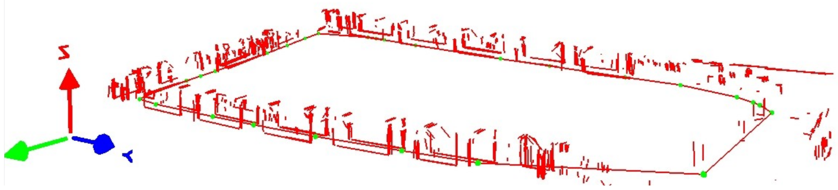
- Benchmark Problems (BPs) aim at testing algorithms:
  - Include detailed description of the task
  - Multi-sensor Data Set related to the task
  - Evaluation Methodology and Tools

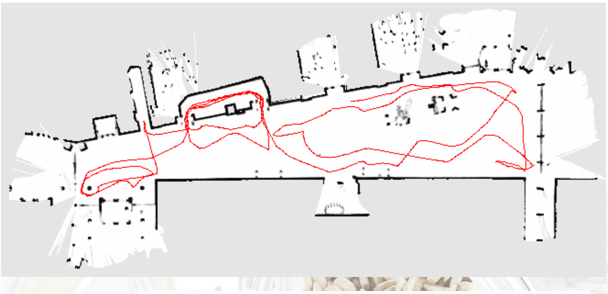
- Benchmark Solutions (BSs) extend BPs with:
  - Description of the algorithm for solving the BP and possible implementation (src or binary)
  - Algorithm output on the BP dataset
  - Evaluation (using the BP methodology)

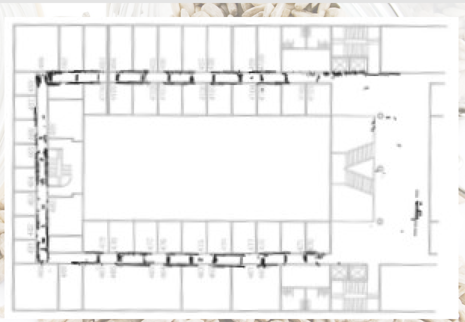


Benchmarks Problems & Solutions

- State of the art solutions for the tasks will be provided as examples such as:
  - Occupancy grids and 2D maps
  - Full 3D maps with segments
  - Map of features from MONOSLAM



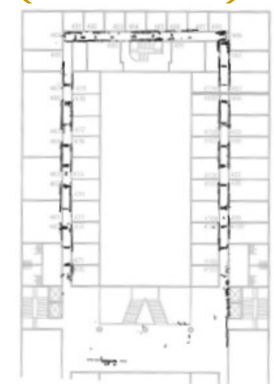


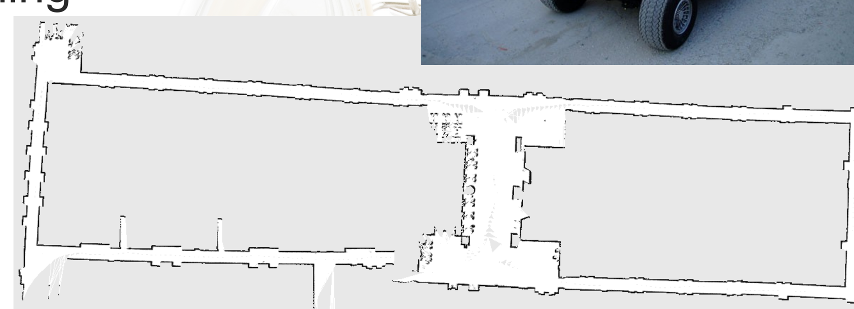


- You can contribute with:
  - Discussion on the RAWSEEDS forum
  - The definition of evaluation methodology
  - A solution (BS) for a Benchmark Problem

RAWSEEDS Today

- Done with the platform setup
  - Indoor
  - Outdoor
- Location Selected
  - Indoor
  - Campus
  - Outdoor
- Definition of Ground truth
  - Camera Network for Indoor positioning
  - RTK-GPS for outdoor position
  - Executive design of environments
- First data under validation
- First solutions developed





RAWSEEDS Measures

- Localization performance
  - Positioning with respect to executive plant & ground truth
- Mapping performance
  - Accuracy measured with respect to predefined landmarks
- SLAM performance
  - Error in path reconstruction
  - Error in positioning before loop closure
  - Map accuracy after loop closure
  - Localization error in your map for new trajectory

- Suggestions are welcome!

Scattered Conclusions (I)

- Ok Simulation, but at some point we need to get real and datasets are the easiest way to replicate results
- Benchmarking is nothing without Good Experimental Methodologies
  - Use publicly available data or provide the data & the solution
  - Give all the details about the system and the benchmarking
- Most SLAM numerical results are based on ground truth:
  - errors in path reconstruction,
  - errors in environment reconstruction (with the limit of unfair comparison of SLAM algorithms using different representations)
  - capability of large loop recognition and closure
  - large map management

Scattered Conclusions (II)

- Do we care about time? What about online operation?
  - "I got this real time algorithm that gives you a random map in zero time. Its quality to time ratio is infinite!" J.D. Tardos
  - If we are interested in the set up of a world model to be used by the robot why should we care about online? Just drive the robot around and after off-line SLAM you are set!

- We should try to compare SLAM algorithms using different representations and sensors …
  - which is the best representation?
  - do we have a real benefit from expensive sensor suites?
  - how much a SLAM fancy algorithm improves robot performance?

References & Time for Questions

- L. Prechelt. “Proben1 -- A Set of Neural Network Benchmark Problems and Benchmarking Rules”. Technical Report 21/94 Universitat Karlsruhe, 1994.
- J. Hallam. “General Guidelines for Robotics Papers using Experiments”,2008
- F. Amigoni, S. Gasparini, M. Gini. “Good Experimental Methodologies for Robotic Mapping: A Proposal” In Proceedings of ICRA 2007, 2007
- T. Collins, J.J. Collins, C. Ryan. “Occupancy Grid Mapping: An Empirical Evaluation” In Proceedings of Mediterranean Conference on Control and Automation, 2007
- O. Wulf, A. Nuchter, J. Hertzberg, B. Wagner. “Benchmarking Urban 6D SLAM”, in Proceedings of Workshop on Performance Evaluation and Benchmarking for Intelligent Robots and Systems, 2007.
- B. Balaguer, S. Carpin, S. Balakirsky. “Towards Quantitative Comparisons of Robot Algorithms: Experiences with SLAM in Simulation and Real World Systems” in Proceedings of Workshop on Performance Evaluation and Benchmarking for Intelligent Robots and Systems, 2007.