# Towards a benchmark for RGB-D SLAM evaluation

Jürgen Sturm[1] Stéphane Magnenat[2] Nikolas Engelhard[3] François Pomerleau[2]
Francis Colas[2] Daniel Cremers[1] Roland Siegwart[2] Wolfram Burgard[3]

***Abstract*— We provide a large dataset containing RGB-D image sequences and the ground-truth camera trajectories with the goal to establish a benchmark for the evaluation of visual SLAM systems. Our dataset contains the color and depth images of a Microsoft Kinect sensor and the ground-truth trajectory of camera poses. The data was recorded at full frame rate (30 Hz) and sensor resolution (640x480). The ground-truth trajectory was obtained from a high-accuracy motion-capture system with eight high-speed tracking cameras (100 Hz). Further, we provide the accelerometer data from the Kinect. Finally, we propose an evaluation criterion for measuring the quality of the estimated camera trajectory of visual SLAM systems.**

## I. Introduction

Simultaneous localization and mapping (SLAM) has a long history in robotics and computer-vision research [11], [6], [1], [15], [7], [4]. Different sensor modalities have been explored in the past, including 2D laser scanners [12], [3], 3D scanners [14], [16], monocular cameras [13], [7], [9], [19], [20] and stereo systems [8]. Recently, low-cost RGB-D sensors became available, of which the most prominent one is the Microsoft Kinect. Such sensors provide both color images and dense depth maps at video frame rates. Henry et al. [5] were the first to use the Kinect sensor in a 3D SLAM system. Others have followed [2], and we expect to see more approaches using RGB-D data for visual SLAM in the near future.

Various datasets and benchmarks have been proposed for laser- and camera-based SLAM, such as the Freiburg, Intel and Newcollege datasets [18], [17]. However until now, no suitable dataset or benchmark existed that can be used to evaluate, measure, and compare the performance of RGB-D SLAM systems. As we consider objective evaluation methods to be highly important for measuring progress in the field (and demonstrating this in a verifiable way), we decided to provide such a dataset. To the best of our knowledge, this is the first RGB-D dataset for visual SLAM benchmarking.

[1] Jürgen Sturm and Daniel Cremers are with the Computer Vision and Pattern Recognition Group, Computer Science Department, Technical University of Munich, Germany. {sturmju,cremers}@in.tum.de

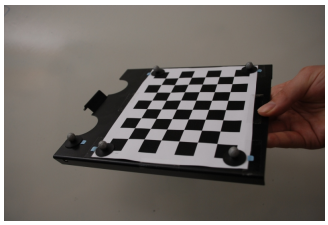
[2] S. Magnenat, F. Pomerlau, F. Colas and R. Seigwart are with the Autonomous Systems Lab, ETH Zurich, Switzerland. {stephane.magnenat,francis.colas}@mavt.ethz.ch and f.pomerleau@gmail.com

[3] Nikolas Engelhard and Wolfram Burgard are with the Autonomous Intelligent Systems Lab, Computer Science Department, University of Freiburg, Germany. {engelhar,burgard}@informatik.uni-freiburg.de

(a) Typical office scene

(b) Motion capture system

(c) Microsoft Kinect sensor with reflective markers



(d) Checkerboard with reflective markers used for calibration

Fig. 1: The office environment and the experimental setup in which the RGB-D dataset with ground truth camera poses was recorded.

## II. Experimental Setup and Data Acquisition

We acquired a large set of data recordings containing both the RGB-D data from the Kinect and the ground truth estimates from the mocap system. We moved the Kinect along different trajectories in typical office environments (see Fig. 1a). The recordings differ in their translational and angular velocities (fast/slow movements) and the size of the environment (one desk, several desks, whole room). We also acquired data for three specific trajectories for debugging purposes, i.e., we moved the Kinect (more or less) individually along the x/y/z-axes and rotated it individually around the x/y/z-axes.

We captured both the color and depth images from an off-the-shelf Microsoft Kinect sensor using PrimeSense's OpenNI-driver. All data was logged at full resolution (640×480) and full frame rate (30 Hz) of the sensor on a Linux laptop running Ubuntu 10.10 and ROS Diamondback. Further, we recorded IMU data from the accelerometer in the Kinect at 500 Hz and also read out the internal sensor parameters from the Kinect factory calibration.

Further, we obtained the camera trajectory by using an external motion capturing system from MotionAnalysis at 100 Hz (see Fig. 1b). We attached reflective targets to the Kinect (see Fig. 1c) and used a modified checkerboard for

calibration (Fig. 1d) to obtain the transformation between the optical frame of the Kinect sensor and the coordinate system of the motion capture system. Finally, we also video-taped all recordings with an external video camera to capture the camera motion and the environment from a different view point.

The original data has been recorded as a ROS bag file. In total, we collected 50 GB of Kinect data, divided into separate nine sequences. The dataset is available online under the Creative Commons Attribution license at

`https://cvpr.in.tum.de/research/datasets/rgbd-dataset`

The website contains—next to additional information about the data formats—videos for simple visual inspection of the dataset.

## III. Evaluation

For evaluating visual SLAM algorithms on our dataset, we propose a metric similar to the one introduced by [10]. The general idea is to compute the relative error between the true and estimated motion w.r.t. the optical frame of the RGB camera. As we have ground-truth pose information for all time indices, we propose to compute the error as the sum of distances between the relative pose at time $i$ and time $i + \Delta$, i.e.,

$$error = \sum_{i=1}^{n} \left[ (\hat{\mathbf{x}}_{i+\Delta} \ominus \hat{\mathbf{x}}_i) \ominus (\mathbf{x}_{i+\Delta} \ominus \mathbf{x}_i) \right]^2 \qquad (1)$$

where $i = 1, \ldots, n$ are the time indices where ground truth information is available, $\Delta$ is a free parameter that corresponds to the time scale, $\mathbf{x}_i$ is the ground truth pose at time index $i$, $\hat{\mathbf{x}}_i$ the estimated pose at time index $i$, $\ominus$ stands for the inverse motion composition operator. If the estimated trajectory has missing values, i.e., there are timesteps $i_{j_1}, \ldots, i_{j_m}$ for which no pose $\hat{\mathbf{x}}_i$ could be estimated, the ratio of missing poses $m/n$ should be stated as well.

All data necessary to evaluate our measure are present in the dataset. We plan to release a Python script that computes these measures automatically given the estimated trajectory and the respective dataset. To prevent that (future) approaches are over-fitted on the dataset, we recorded all scenes twice, and held back the ground-truth trajectory in these secondary recordings. With this, we plan to provide a comparative offline evaluation benchmark for visual SLAM systems.

## IV. Conclusions

In this paper, we have presented a novel RGB-D dataset for benchmarking visual SLAM algorithms. The dataset contains color images, depth maps, and associated ground-truth camera pose information. Further, we proposed an evaluation metric that can be used to assess the performance of a visual SLAM system. We thus propose a benchmark that allows researchers to objectively evaluate visual SLAM systems. Our next step is to evaluate our own system [2] on this dataset in order to provide a baseline for future implementations and evaluations. In this way, we hope to detect (and resolve) potential problems present in our current dataset, such as calibration and synchronization issues between the Kinect and our mocap system as well as the effects of motion blur and the rolling shutter of the Kinect. Furthermore, we want to investigate ways to measure the performance of a SLAM system not only in terms of the accuracy of the estimated camera trajectory, but also in terms of the quality of the resulting map of the environment.

## References


[1] F. Dellaert. Square root SAM. In *Proc. of Robotics: Science and Systems (RSS)*, Cambridge, MA, USA, 2005.

[2] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard. Real-time 3D visual SLAM with a hand-held RGB-D camera. In *Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum*, Vasteras, Sweden, 2011.

[3] G. Grisetti, C. Stachniss, and W Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Transactions on Robotics (T-RO)*, 23:34–46, 2007.

[4] G. Grisetti, C. Stachniss, and W. Burgard. Non-linear constraint network optimization for efficient map learning. *IEEE Transactions on Intelligent Transportation systems*, 10(3):428–439, 2009.

[5] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In *Proc. of the Intl. Symp. on Experimental Robotics (ISER)*, Delhi, India, 2010.

[6] H. Jin, P. Favaro, and S. Soatto. Real-time 3-D motion and structure of point features: Front-end system for vision-based control and interaction. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[7] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. of the IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, Nara, Japan, 2007.

[8] K. Konolige, M. Agrawal, R.C. Bolles, C. Cowan, M. Fischler, and B.P. Gerkey. Outdoor mapping and navigation using stereo vision. In *Intl. Symp. on Experimental Robotics (ISER)*, 2007.

[9] K. Konolige and J. Bowman. Towards lifelong visual maps. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1156–1163, 2009.

[10] R. Kümmerle, B. Steder, C. Dornhege, M. Ruhnke, G. Grisetti, C. Stachniss, and A. Kleiner. On measuring the accuracy of SLAM algorithms. *Autonomous Robots*, 27:387–407, 2009.

[11] F. Lu and E. Milios. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, 4(4):333–349, 1997.

[12] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002. AAAI.

[13] D. Nistér. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, 16:321–329, 2005.

[14] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann. 6D SLAM – 3D mapping outdoor environments: Research articles. *J. Field Robot.*, 24:699–722, August 2007.

[15] E. Olson, J. Leonard, and S. Teller. Fast iterative optimization of pose graphs with poor initial estimates. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2006.

[16] B. Pitzer, S. Kammel, C. DuHadway, and J. Becker. Automatic reconstruction of textured 3D models. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2010.

[17] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *Intl. Journal of Robotics Research (IJRR)*, 28(5):595–599, 2009.

[18] C. Stachniss, P. Beeson, D. Hähnel, M. Bosse, J. Leonard, B. Steder, R. Kümmerle, C. Dornhege, M. Ruhnke, G. Grisetti, and A. Kleiner. Laser-based slam datasets and benchmarks at http://openslam.org.

[19] H. Strasdat, J. M. M. Montiel, and A. Davison. Scale drift-aware large scale monocular slam. In *Proc. of Robotics: Science and Systems (RSS)*, Zaragoza, Spain, 2010.

[20] J. Stühmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Proc. of the DAGM Symposium on Pattern Recognition (DAGM)*, Darmstadt, Germany, 2010.