Acta Polytechnica CTU Proceedings **6**:1–5, 2016
doi:10.14311/APP.2016.6.0001

available online at http://ojs.cvut.cz/ojs/index.php/app

ON CONSTRUCTION OF A RELIABLE GROUND TRUTH FOR EVALUATION OF VISUAL SLAM ALGORITHMS

Jan Bayer*, Petr Čížek, Jan Faigl

Czech Technical University, Faculty of Electrical Engineering, Technicka 2, Prague, Czech Republic

* corresponding author: bayerja1@fel.cvut.cz

Abstract. In this work we are concerning the problem of localization accuracy evaluation of visual-based Simultaneous Localization and Mapping (SLAM) techniques. Quantitative evaluation of the SLAM algorithm performance is usually done using the established metrics of Relative pose error and Absolute trajectory error which require a precise and reliable ground truth. Such a ground truth is usually hard to obtain, while it requires an expensive external localization system. In this work we are proposing to use the SLAM algorithm itself to construct a reliable ground truth by offline frame-by-frame processing. The generated ground truth is suitable for evaluation of different SLAM systems, as well as for tuning the parametrization of the on-line SLAM. The presented practical experimental results indicate the feasibility of the proposed approach.

Keywords: RGB-D SLAM, localization, ground truth construction, legged robot, rough terrain traversal, adaptive motion gait.

1. Introduction

Reliable localization is an essential prerequisite in many practical robotic scenarios, thus its accuracy is the most important parameter in the evaluation of newly developed localization techniques. There are numerous methods based on various principles that can localize the robot in its operational environment. However, when benchmarking such localization methods it is necessary to use a reliable ground truth together with some established metrics (e.g. [1, 2]).

In this work, we address the problem of localization accuracy assessment in a visual Simultaneous Localization and Mapping (SLAM) task [3]. In principle, it is possible to use publicly available datasets, e.g., the Freiburg Uni dataset [1], the RGB-D SLAM dataset [2], or the KITTY dataset [4], to evaluate the accuracy of SLAM method. However, our particular environment, setup, and sensory equipment might not exhibit similar properties as the ones used in these standardized datasets. Therefore we need to construct our own dataset and acquire a ground truth to assess the localization accuracy.

All the above-mentioned datasets are supplied with a centimeter-level precision ground truth provided by an external localization system like a differential GPS [4], which is based on a standard GPS enhanced with ground reference stations with known geospatial positions. Other types of external localization systems used in publicly available datasets are motion capture system [2] or a total station [5]. The former utilizes the data captured by camera sensors to localize the robot marked with special patterns well-distinguishable in the environment. The later measures distance between the observed object and the static ground station using time of flight principle. A 3-DOF position of the robot in a global reference frame is obtained from the distance and the angles measured by the total station.

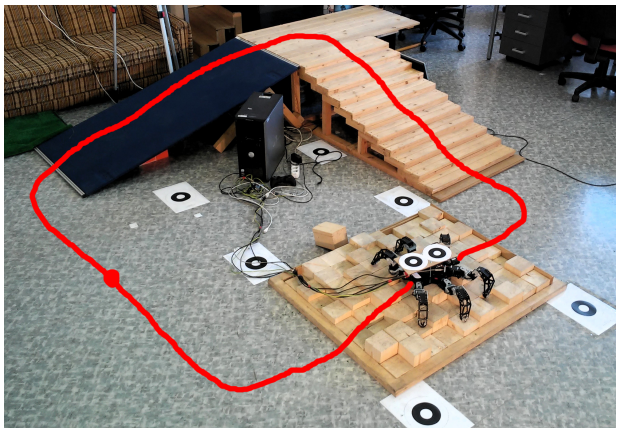


Figure 1. Experimental environment

However, acquisition of a reliable ground truth using any of the mentioned localization techniques might be a considerable problem while all of the aforementioned principles are very expensive. Thus we are proposing to use the dataset itself for the ground truth construction by offline frame-by-frame processing of the dataset using the SLAM algorithm. In such a case, we are not interested in an on-line performance of the algorithm rather than the reliability and accuracy of the constructed trajectory estimate to be used as a ground truth. Such a processing is assumed to provide the best trajectory estimate for the given dataset and localization technique. Moreover, by taking into account certain factors in the dataset construction., which are discussed further in this paper, it is possible to improve the reliability of the generated ground truth.

To verify our hypothesis, we are reporting on the results of such a ground truth construction evaluation based on the RGB-D SLAM method [6] compared to the ground truth provided by the WhyCon motion capture system [7]. We evaluated the proposed approach on a challenging dataset obtained with the RGB-D camera mounted on a hexapod walking robot crawling in rough terrain (see Fig. 1). Such a scenario requires full 6-DOF position estimation, while the camera is subjected to unpredictable motion and the resulting dataset is very different in comparison to all the standardized datasets obtained mainly by wheel platforms.

The presented approach to ground truth construction is somehow similar to the task of structure-from-motion studied in computer vision community. In [8], the authors use color images from the RGB-D Microsoft Kinect camera to recover its pose, but for the scene and object recognition purposes. However, in our work we are focusing solely on the localization accuracy rather than the quality of the model (map) of the environment.

The paper is structured as follows. Section 2 explains our method of getting the ground truth for the evaluation and the used evaluation metrics. The set-up of a practical experiment together with the evaluation results are described in Section 3. Conclusion and remarks for future work are dedicated to Section 4.

2. Proposed solution

The main motivation of our work is to provide a reliable ground truth estimate whenever a precise external localization is not available to be used as the ground truth. For this reason, we are proposing to use the captured dataset and process it with a SLAM system offline frame-by-frame with special settings of parameters to minimize the trajectory error. Afterward, the resulting trajectory estimate can be used as a ground truth to compare to the online generated trajectories, while in deployment scenarios we are interested in the online performance of SLAM methods. Such an estimated ground truth can be used to compare different SLAM methods or to find the best parametrization for the online SLAM trajectory estimation.

2.1. Evaluation metrics

The SLAM algorithms can be compared either by the quality of the generated map or the accuracy of the estimated trajectory. As the evaluation metrics, we use the established metrics of ATE (absolute trajectory error) and RPE (relative pose error) presented in [2].

ATE is given as:

$$\mathbf{F}_i = \mathbf{Q}_i^{-1}\mathbf{S}\mathbf{P}_i, \tag{1}$$

where $\mathbf{Q}_i^{-1}$ is transformation matrix, which maps $(i+1)$th point of ground truth to ith point of ground truth. $\mathbf{P}_i$ is a similar matrix, which maps ith point of the estimated trajectory to the $(i+1)$th point of the trajectory. $\mathbf{S}$ is a transformation matrix, which aligns the estimated trajectory to the ground truth.

ATE measures the absolute error of all estimated 6-DOF positions, which means that ATE can be represented by ATE_t and ATE_ϕ, where ATE_t are euclidean distances between estimated poses and corresponding ground truth positions and ATE_ϕ is the absolute error of the estimated orientations.

RPE measures the drift of the estimated trajectory (error in shape of the trajectory estimate) and it is given by the equation:

$$\mathbf{E}_i = (\mathbf{Q}_i^{-1}\mathbf{Q}_{i+\Delta})^{-1}(\mathbf{P}_i^{-1}\mathbf{P}_{i+\Delta}), \tag{2}$$

where Δ defines a fixed frame interval. We used $\Delta = 1$ for the evaluation, which means that RPE measures a drift of visual odometry.

RPE can be also represented by RPE_t (the relative translational error) and RPE_ϕ (the relative rotational error).

3. Evaluation

To test our hypothesis and prove its feasibility we designed a practical experiment where we evaluated the localization accuracy provided by an offline frame-by-frame processing of a visual SLAM dataset captured by a hexapod robot traversing rough terrain with a ground truth provided by an external localization system.

In particular, we focused on evaluation with the structured light camera and RGB-D SLAM algorithm based on [6]. We choose to use the structured light sensor because it is inexpensive and provide the SLAM method with a metric information unlike the stereo and monocular SLAM methods which can estimate the trajectory up to scale. For the establishment of a reliable ground truth, we used motion capture system WhyCon [7] which tracks the circular markers attached to the robot using two Logitech HD Pro Webcam C920 cameras. Used external localization provided centimeter-level precision ground truth.

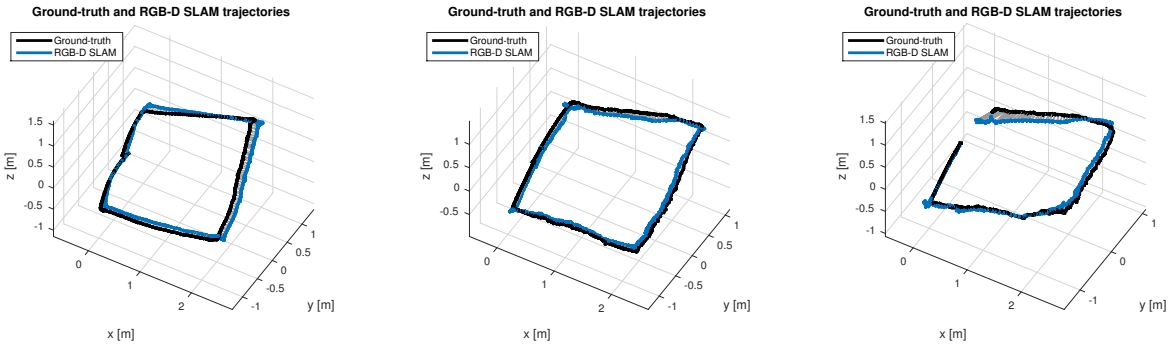
We used the RGB-D camera Asus Xtion Pro for the visual SLAM, which acquires a color image and also depth measurements. The experimental dataset captured using the RGB-D camera mounted on a hexapod walking platform contains sequences of RGB and depth images in full Asus Xtion Pro resolution 640×480 px at 10 frames per second. The whole dataset for the evaluation was captured using the ROS framework [9].

3.1. RGB-D SLAM

The considered RGB-D SLAM system is based on [6] and it operates as follows. First positions of the salient image features are extracted from the RGB image supplemented by their depth provided by the RGB-D sensor. Then, a rigid transformation is found

Trial	Frame-by-frame processing					Online processing				
	ATE_t [cm]	ATE_ϕ [rad]	RPE_t [cm]	RPE_ϕ [rad]	End dist. [cm]	ATE_t [cm]	ATE_ϕ [rad]	RPE_t [cm]	RPE_ϕ [rad]	End dist. [cm]
No. 1	4.00	–	0.92	–	7.11	9.60	–	1.13	–	3.5
No. 2	4.32	–	0.54	–	12.54	8.13	–	0.56	–	8.93
No. 3	3.68	–	0.40	–	5.21	22.02	–	0.87	–	33.15
No. 4	9.97	0.08	1.21	0.01	9.13	12.50	0.09	1.31	0.01	7.12
No. 5	11.44	0.06	0.80	0.01	15.96	15.64	0.06	0.94	0.02	27.55

TABLE 1. Trajectory estimation results



(A) . No.3 (B) . No.4 (C) . No.5

FIGURE 2. Trajectory estimation results.

using the RANSAC [10] algorithm between the currently processed frame and a subset of already mapped frames. This subset consists of n_p directly preceding frames, n_g graph neighboring frames and n_r random frames from the whole map which serves to detect large loop closures. The estimated pose is then added to the map (pose graph) and refined using the **g2o** graph optimizing algorithm [11]. The optimization is especially beneficial in case of large loop closures when the whole trajectory estimate is improved.

3.2. EXPERIMENTAL SETUP

Experimental environment (see Fig. 1) that has been used for capturing of the dataset consists of a plane, low stairs, ramp, and square filled with wooden blocks of different heights. The dataset has been captured by a hexapod walking robot equipped with adaptive motion gait [12] which enables the robot to traverse difficult obstacles. The robot was guided by an operator along the squared path of approx. 9 m length.

Our experimental environment has been designed to represent some basic problems, which can SLAM method face. These include high speed of rotation at the corners of the trajectory, which affects the precision of rotation estimation. The stairs and wooden blocks parts of the trajectory force angular rotations along the geometrical axis and the descending ramp force the SLAM method to deal with the limited number of image features. Moreover, the locomotion of the hexapod robot induces unpredictable and abrupt camera motions during the whole trajectory.

3.3. RESULTS

Altogether 5 trajectories were captured and analyzed in order to test our hypothesis. Ground truth for the first three trajectories; however, does not contain information about the heading of the robot, and thus only translational component of the ATE and RPE can be verified.

The considered RGB-D SLAM parametrization for the frame-by-frame dataset processing has been the SURF [13] feature extractor detecting 600 features and keeping 400 best matches with a very long comparison horizon of $n_p = 16$, $n_g = 4$ and $n_r = 1$. RANSAC stopping criteria have been set to 85% of inliers and 1000 iterations considering point an inlier if it is not outside the range of 3 pixels. Optimization precision has been set to the best values optimizing each individual frame. Such a processing took more than 4 hours per one trajectory on Intel Core i5 machine.

To show the feasibility of our approach we processed the dataset online as well. The parametrization for the online processing has been the SURF feature extractor detecting 400 features and keeping 250 of them, $n_p = 3$, $n_g = 3$ and $n_r = 1$ with 500 RANSAC iterations and optimization after every 10 frames. Such a processing achieves an online performance of average 5.2 frames per second on the same evaluation hardware.

The results are summarized in Table 1. Resulting trajectories for trials No. 3, No. 4 and No. 5 are depicted in Figure 2.

3.4. Discussion

Table 1 indicates that an average ATE_t, when using frame-by-frame processing, is about 1% of the trajectory length and the average RPE_t is under 1 cm. Notice the ATE_t and RPE_t values are always higher for the online processing results. On the other hand, differences in the ATE_ϕ and RPE_ϕ are neglectable although the slight imperfections in the orientation estimation are the most influencing factor in the accuracy of the localization [14]. The results also indicate that the frame-by-frame processing provides the best result obtainable by a given SLAM method. Thus, it is possible to use the estimated ground truth for the evaluation of the online versions of the same algorithm to find the best parametrization which is deployable online onboard of the mobile robot.

Note, the resulting localization accuracy is also greatly influenced by the loop closure in case the robot visits previously mapped location. If there is no loop closure, then the trajectory estimate is given as the visual odometry only, which error is asymptotically unbound. In case the loop closure is detected, the optimization process corrects the pose estimates along the whole trajectory. The phenomenon is illustrated in Figure 2c, where the loop did not close; thus, the trajectory estimate exhibits higher error in the end part of the trajectory, and Figure 3 which presents the results of the frame-by-frame processing of the same trajectory with open-loop enabled and disabled. The results for the closed-loop trajectory are ATE_t = 4.19 cm, RPE_t = 0.94 rad·10^{-2} and End dist = 7.33 cm, while for the open-loop trajectory are ATE_t = 8.72 cm, RPE_t = 0.96 rad·10^{-2} and End dist = 13.73 cm.

Although the drift of the visual odometry represented by the RPE_t is similar in both cases, the resulting end distance and absolute trajectory error are higher when the open-loop scenario is considered. Thus, we recommend designing the experimental trajectory to exploit the loop closures to obtain a better ground truth estimate.

Note the prerequisite of our method is that there are no systematic errors, which can significantly decrease the precision of ground truth estimate. To avoid systematic errors we recommend to use a special shape of the robot's path, e.g., square.

4. Conclusion

In this paper, we have studied the problem of the localization accuracy assessment in a SLAM task whenever a reliable ground truth provided by an external localization system is not available. We have proposed to use frame-by-frame processing of the captured dataset to obtain a sufficiently reliable ground truth for the quantitative evaluation of the online SLAM systems. The estimated ground truth is then suitable for the comparison of different SLAM methods or determination of the best parametrization of the online SLAM. Note, the important thing for the dataset creation is

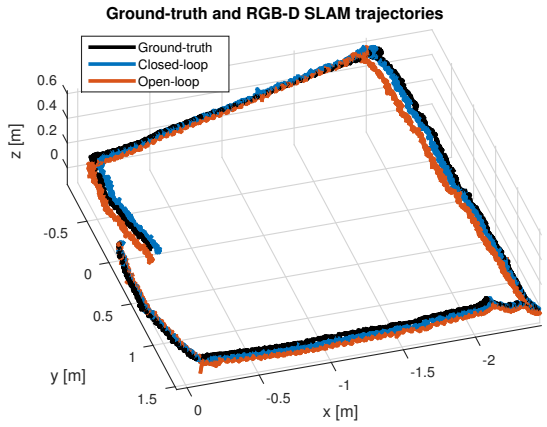


Figure 3. Comparison of closed-loop and open-loop frame-by-frame processed trajectories

the loop closing, which can significantly improve the robot localization and it is also important to design the shape of the path in order to avoid systematic errors. We tested the presented method with a hexapod crawling robot and RGB-D SLAM system, but we suppose that the proposed method is applicable to different SLAM methods and different platforms too. Presented results support the feasibility of the proposed approach and can be utilized in cases when expensive and hard to setup external localization systems are not available or suitable.

Acknowledgements

The presented work was supported by the Czech Science Foundation (GAČR) under research project No. 15-09600Y. The support of grant No. SGS16/235/OHK3/3T/13 to Petr Čížek is also gratefully acknowledged.

References

[1] R. Kümmerle, B. Steder, C. Dornhege, et al. On measuring the accuracy of SLAM algorithms. *Autonomous Robots* **27**(4):387–407, 2009. doi:10.1007/s10514-009-9155-6.

[2] J. Sturm, N. Engelhard, F. Endres, et al. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE International Conference on Intelligent Robots and Systems*, pp. 573–580. 2012. doi:10.1109/IROS.2012.6385773.

[3] S. Thrun, W. Burgard, D. Fox. *Probabilistic robotics*. MIT press, 2005. doi:10.1162/artl.2008.14.2.227.

[4] M. Menze, A. Geiger. Object Scene Flow for Autonomous Vehicles. In *Conference on Computer Vision and Pattern Recognition*, pp. 3061–3070. 2015. doi:10.1109/CVPR.2015.7298925.

[5] N. Sünderhauf, K. Konolige, S. Lacroix, P. Protzel. *Visual Odometry Using Sparse Bundle Adjustment on an Autonomous Outdoor Vehicle*, pp. 157–163. Springer Berlin Heidelberg, 2006. doi:10.1007/3-540-30292-1_20.

[6] F. Endres, J. Hess, N. Engelhard, et al. An evaluation of the RGB-D SLAM system. In *IEEE International Conference on Robotics and Automation*, pp. 1691–1696. 2012. doi:10.1109/ICRA.2012.6225199.

[7] T. Krajník, M. Nitsche, J. Faigl, et al. A practical multirobot localization system. *Journal of Intelligent & Robotic Systems* **76**(3-4), 2014. DOI:10.1007/s10846-014-0041-x.

[8] S. Bao. Semantic structure from motion. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2025–2032. 2011. DOI:10.1109/CVPR.2011.5995462.

[9] M. Quigley, K. Conley, B. P. Gerkey, et al. ROS: an open-source Robot Operating System. In *IEEE International Conference on Robotics and Automation (ICRA) – Workshop on Open Source Robotics*. 2009.

[10] M. A. Fischler, R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **24**(6):381–395, 1981.

[11] R. Kümmerle, G. Grisetti, H. Strasdat, et al. g2o: A general framework for graph optimization. In *IEEE International Conference on Robotics and Automation*, pp. 3607–3613. 2011. DOI:10.1109/ICRA.2011.5979949.

[12] J. Mrva, J. Faigl. Tactile sensing with servo drives feedback only for blind hexapod walking robot. In *10th International Workshop on Robot Motion and Control, RoMoCo*, pp. 240–245. 2015. DOI:10.1109/RoMoCo.2015.7219742.

[13] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding* **110**(3):346–359, 2008. DOI:10.1016/j.cviu.2007.09.014.

[14] P. Čížek, J. Faigl. On Localization and Mapping with RGB-D Sensor and Hexapod Walking Robot in Rough Terrains. In *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2273–2278. 2016.