



Benchmarking Simultaneous Localisation and Mapping Pipelines in Feature-Poor Warehouse Aisles

Beatriz Petrovic^{1*}, Sophie Sithole², Minh Chukwu³

¹Department of Mechanical Engineering, Transylvanian School of Mechanical Engineering, Barbados

²Department of Mechanical Engineering, Limpopo Institute of Mechanical Engineering, Hungary

³Department of Mechanical Engineering, Thames Academy of Mechanical Engineering, Norway

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijrra.2026.v1.i1.pp14-19](https://doi.org/10.62715/ijrra.2026.v1.i1.pp14-19)

Received: 07 May 2026

Accepted: 19 July 2026

Published: 13 September 2026

Abstract

Mobile robots in warehouses spend much of their working time in the place where simultaneous localisation and mapping performs worst: long, narrow aisles between racks that look the same from one bay to the next. Along the aisle a laser scan offers nothing to constrain position, cameras see repeated shelving and bare floor, and a wrong loop closure between two identical bays can fold the map. Public benchmarks, built around offices, streets and flying vehicles, contain little of this, so a pipeline that ranks well on them may still fail on the warehouse floor. This article proposes a protocol for benchmarking SLAM pipelines specifically in feature-poor aisles. It defines aisle classes by measurable descriptors, specifies a sensor rig and its calibration, describes how to obtain ground truth from surveyed fiducials where motion capture cannot reach, prescribes a set of trajectories that provoke the known failure modes, fixes rules for configuring the pipelines under test, and sets out metrics that separate along-aisle drift from cross-aisle error and count false loop closures. The protocol is a proposal built from published evaluation practice. It has not been carried out here, and the article states what a first application would have to demonstrate.

Keywords: SLAM benchmarking, Warehouse robotics, Degenerate geometry, Perceptual aliasing, Trajectory evaluation, Lidar odometry

© 2026 Beatriz Petrovic et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Simultaneous localisation and mapping is often described as a mature technology, and for many settings the description is fair. The survey by Cadena et al. (2016) nonetheless argues that maturity depends on the combination of robot, environment and performance requirement, and that robustness in hard environments remains open. Warehouse aisles are such an environment. They are geometrically simple, which sounds like an advantage and is the opposite: two parallel planes of racking and a flat floor leave the direction of travel unconstrained.

Operators have historically avoided the problem with infrastructure. Reflectors, floor codes and magnetic tape give automated vehicles a position without any need to interpret the scene. Autonomous

* Corresponding author: Beatriz Petrovic, Department of Mechanical Engineering, Transylvanian School of Mechanical Engineering, Barbados.
E-mail: beatriz.petrovic@transylvanian.edu

© 2026 Beatriz Petrovic et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

mobile robots that navigate from natural features promise cheaper installation and easier rerouting, and their dependence on SLAM brings the aisle problem back. An integrator choosing among open-source pipelines has little evidence to go on, because the common benchmarks were recorded elsewhere.

This article proposes a benchmarking protocol aimed at that gap. It is a method paper: it specifies what to record, how to obtain a reference, which pipelines to compare under which rules, and what to report. No dataset has been collected and no pipeline has been evaluated for this article. The protocol is put forward for others to apply, criticise and amend.

2. Background

The design rests on two bodies of work: the analysis of why estimation degrades in aisles, and the established practice of SLAM evaluation.

2.1. Why aisles defeat SLAM

Three mechanisms are at work. The first is geometric degeneracy. Scan matching aligns each new scan to a map or to the previous scan, and in a corridor with featureless walls the alignment cost is flat along the corridor axis. [Censi \(2007\)](#) showed how the covariance of an iterative closest point estimate can be computed in closed form, which makes the weak direction visible, and [Zhang et al. \(2016\)](#) proposed detecting degeneracy from the eigenvalues of the optimisation problem and updating the state only in well-conditioned directions. Pipelines that lack such a safeguard report confident and wrong motion.

The second mechanism is perceptual aliasing: different places that produce the same sensor reading. Racking is modular by design, and one bay matches the next in a laser scan and often in an image. Place recognition then proposes loop closures between distinct locations, and a single accepted false closure can corrupt a pose graph beyond repair ([Cadena et al., 2016](#)). The third is change. Pallets arrive and leave, so the occupied volume of a rack face differs between the mapping run and the working run, and people and forklifts occlude the view.

2.2. Existing benchmarks and metrics

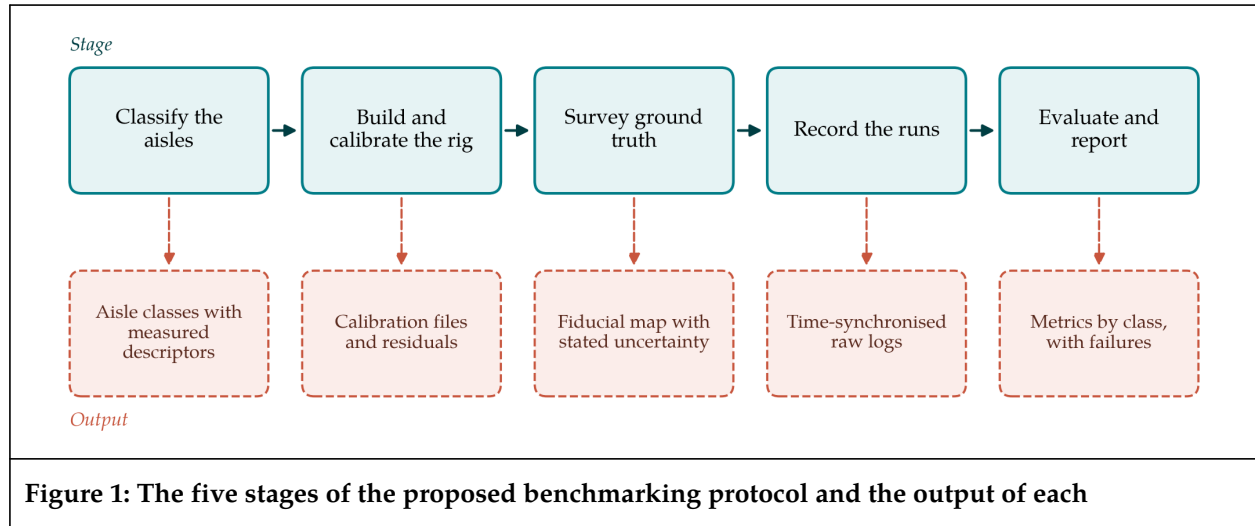
The reference datasets of the field were recorded for other purposes. The TUM RGB-D benchmark covers handheld and robot-mounted cameras in offices and a hall ([Sturm et al., 2012](#)), KITTI covers road driving ([Geiger et al., 2012](#)) and EuRoC a micro aerial vehicle in a machine hall and a small room ([Burri et al., 2016](#)). Their lasting contribution is methodological. [Sturm et al. \(2012\)](#) defined the absolute trajectory error, computed after rigid alignment of estimate and reference by the closed-form method of [Umeyama \(1991\)](#), and the relative pose error over a fixed interval. [Kümmerle et al. \(2009\)](#) argued for measuring errors in relative displacements between selected poses, which avoids dependence on a global frame. [Zhang and Scaramuzza \(2018\)](#) showed that the choice of alignment changes the reported error and must be stated. Open tools implement these measures ([Grupp, 2017](#)).

3. The Proposed Protocol

The protocol has five stages, each with a defined output, as set out in Figure 1. The stages are described below in the order in which they would be carried out.

3.1. Stage 1: Classify the aisles

"Feature-poor" must be made measurable before it can be benchmarked. Each test aisle is described by its length, its width, the ratio of the two, the type of rack face (open beams, mesh, solid panels or wrapped pallets), the fill level at recording time, the floor finish and the lighting. From these the protocol defines three classes: a reference class of short aisles with open, varied rack faces; a degenerate class of long aisles with uniform or solid faces; and an aliased class in which several parallel aisles are identical to one another. Reporting results by class prevents a good average from hiding failure in the class that matters.



3.2. Stage 2: Build and calibrate the sensor rig

All pipelines must be fed from one recording, so the rig carries the union of the sensors they need: a planar safety-type laser scanner at the usual mounting height, a three-dimensional lidar, a stereo or depth camera, an inertial measurement unit and wheel encoders. Sensors are hardware-synchronised where possible and time offsets are estimated otherwise. Intrinsic and extrinsic calibration follows published practice for joint spatial and temporal calibration of cameras and inertial sensors (Furgale et al., 2013), and the calibration files and their residuals are released with the data, since a poor extrinsic estimate is easily mistaken for a weakness of the algorithm.

3.3. Stage 3: Survey the ground truth

Motion capture, the reference used in small indoor datasets, cannot cover an aisle of industrial length, and satellite positioning is unavailable under a roof. The protocol therefore uses surveyed fiducials. Visual tags of the AprilTag family (Olson, 2011) are fixed to rack uprights at regular spacing on both sides, and their positions are measured with a total station from stations tied into a common control network. A dedicated upward- or side-facing camera, not available to the pipelines under test, observes the tags and yields reference poses wherever a tag is in view. The uncertainty of the reference is estimated from the survey and stated; errors below it are not interpreted. Tags are masked from the images supplied to visual pipelines so that the reference does not become a feature.

3.4. Stage 4: Record the runs

Each trajectory is chosen to provoke one mechanism. A straight traverse of a single long aisle at constant speed isolates along-aisle drift. A serpentine route through adjacent identical aisles invites false loop closures. A route that leaves the racking, crosses an open dock area and returns tests recovery and genuine loop closure. A repeat of the first route after the rack contents have changed tests sensitivity to change, and a run with pedestrians and trucks present tests dynamic clutter. Runs are recorded at the working speed of the vehicle, including the stops and turns in place that real missions contain. Raw data are stored unfiltered.

3.5. Stage 5: Evaluate and report

The pipelines compared should span the main designs in use: planar lidar SLAM by particle filter (Grisetti et al., 2007) and by graph optimisation with submaps (Hess et al., 2016), three-dimensional lidar odometry with and without tight inertial coupling (Zhang and Singh, 2014; Shan et al., 2020), and visual and visual-inertial systems (Qin et al., 2018; Campos et al., 2021). Two configuration rules apply. Each pipeline is run once with its published default parameters and once with parameters tuned on a held-out aisle that is never used for scoring, and both results are reported. Each sequence is processed several times, because pipelines with multithreading or random sampling are not deterministic.

Table 1 lists the metrics. The departure from common practice is the decomposition of error into an along-aisle and a cross-aisle component in the frame of the racking, since a single root-mean-square figure conceals the degeneracy the benchmark exists to expose.

Table 1: Metrics specified by the proposed protocol and the failure each is meant to reveal		
Metric	How it is obtained	What it reveals
Absolute trajectory error	Position error after rigid alignment (Sturm et al., 2012)	Global consistency of the trajectory
Along-aisle error	Error component parallel to the rack line, per aisle	Drift from geometric degeneracy
Cross-aisle error	Error component perpendicular to the rack line	Clearance margin for safe driving
Relative pose error	Error in displacement over fixed path lengths	Local odometry quality
False loop closures	Accepted closures between poses in different surveyed places	Susceptibility to perceptual aliasing
Completion rate	Share of repeated runs finishing without tracking loss	Reliability, not accuracy
Compute load	Processor and memory use on a stated computer	Feasibility on vehicle hardware

Reports give the distribution of each metric over repeated runs, by aisle class, together with every run that failed. A pipeline that loses tracking is recorded as a failure for that run and is not dropped from the averages without comment.

4. Fitting the Protocol to an Operating Warehouse

A working site imposes limits that a laboratory does not. Aisles can seldom be closed for long, so the survey should be planned so that the tags, once fixed, can stay in place and be re-observed on later visits. Rack contents cannot be controlled, and the protocol treats this as information: fill level is photographed and logged for each run. Safety rules govern where a test vehicle may travel alongside staff, which may restrict the dynamic-clutter run to supervised periods. Commercial confidentiality may prevent the release of camera images, in which case lidar and inertial data with the reference trajectory still permit most of the comparison. Where no total station is available, a calibrated measuring wheel and laser distance meter give a coarser reference whose larger uncertainty must be carried through to the interpretation.

5. Limitations and What Remains to Be Shown

The protocol is proposed and untested. Its weakest element is the ground truth. Fiducial-based reference poses are discontinuous, available only where a tag is seen, and their accuracy depends on survey quality and on the calibration of the reference camera. A first application would need to show, by an independent check such as repeated total-station shots to a prism on the vehicle at rest, that the reference is good enough to rank the pipelines.

Several design choices also await evidence. Whether the three aisle classes separate pipeline behaviour as intended, or whether a continuous descriptor of degeneracy would serve better, can only be learned from data. The tuned-parameter rule reduces the advantage of the experimenter who knows one pipeline well, without removing it. A benchmark recorded at one site reflects that site's racking, floor and lighting, and generality requires several. Finally, trajectory accuracy is a proxy: the property an operator cares about is whether the vehicle completes missions, and the link between the two is assumed here and not demonstrated.

6. Conclusion

Warehouse aisles combine geometric degeneracy, perceptual aliasing and continual change, and none of the standard datasets represents them. The protocol proposed in this article specifies aisle classes, a common sensor rig, a surveyed fiducial reference, trajectories matched to failure mechanisms, configuration rules and a set of metrics that splits error along and across the aisle and counts false loop closures. It is offered as a starting point. Its usefulness will be settled by the first groups who record data with it and report where it needed to change.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W. and Siegwart, R. (2016). [The EuRoC micro aerial vehicle datasets](#). *International Journal of Robotics Research*, 35(10): 1157-1163. <https://doi.org/10.1177/0278364915620033>
- Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I. and Leonard, J.J. (2016). [Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age](#). *IEEE Transactions on Robotics*, 32(6): 1309-1332. <https://doi.org/10.1109/TRO.2016.2624754>
- Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M. and Tardós, J.D. (2021). [ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM](#). *IEEE Transactions on Robotics*, 37(6): 1874-1890. <https://doi.org/10.1109/TRO.2021.3075644>
- Censi, A. (2007). [An accurate closed-form estimate of ICP's covariance](#). *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA)*, 3167-3172.
- Furgale, P., Rehder, J. and Siegwart, R. (2013). [Unified temporal and spatial calibration for multi-sensor systems](#). *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1280-1286.
- Geiger, A., Lenz, P. and Urtasun, R. (2012). [Are we ready for autonomous driving? The KITTI vision benchmark suite](#). *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354-3361. <https://doi.org/10.1109/CVPR.2012.6248074>
- Grisetti, G., Stachniss, C. and Burgard, W. (2007). [Improved techniques for grid mapping with Rao-Blackwellized particle filters](#). *IEEE Transactions on Robotics*, 23(1): 34-46. <https://doi.org/10.1109/TRO.2006.889486>
- Grupp, M. (2017). [evo: Python package for the evaluation of odometry and SLAM](#). Open-source software, github.com/MichaelGrupp/evo.
- Hess, W., Kohler, D., Rapp, H. and Andor, D. (2016). [Real-time loop closure in 2D LIDAR SLAM](#). *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, 1271-1278. <https://doi.org/10.1109/ICRA.2016.7487258>
- Kümmerle, R., Steder, B., Dornhege, C., Ruhnke, M., Grisetti, G., Stachniss, C. and Kleiner, A. (2009). [On measuring the accuracy of SLAM algorithms](#). *Autonomous Robots*, 27(4): 387-407. <https://doi.org/10.1007/s10514-009-9155-6>
- Olson, E. (2011). [AprilTag: A robust and flexible visual fiducial system](#). *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, 3400-3407. <https://doi.org/10.1109/ICRA.2011.5979561>
- Qin, T., Li, P. and Shen, S. (2018). [VINS-Mono: A robust and versatile monocular visual-inertial state estimator](#). *IEEE Transactions on Robotics*, 34(4): 1004-1020. <https://doi.org/10.1109/TRO.2018.2853729>

- Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C. and Rus, D. (2020). LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5135-5142.
- Sturm, J., Engelhard, N., Endres, F., Burgard, W. and Cremers, D. (2012). A benchmark for the evaluation of RGB-D SLAM systems. *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 573-580. <https://doi.org/10.1109/IROS.2012.6385773>
- Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4): 376-380. <https://doi.org/10.1109/34.88573>
- Zhang, J. and Singh, S. (2014). LOAM: Lidar odometry and mapping in real-time. *Robotics: Science and Systems X*. <https://doi.org/10.15607/RSS.2014.X.007>
- Zhang, J., Kaess, M. and Singh, S. (2016). On degeneracy of optimization-based state estimation problems. *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, 809-816.
- Zhang, Z. and Scaramuzza, D. (2018). A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7244-7251. <https://doi.org/10.1109/IROS.2018.8593941>