SLAM for Indoor Mapping of Wide Area Construction Environments

Vincent Ress *, Wei Zhang, David Skuddis, Norbert Haala, Uwe Soergel

Institute for Photogrammetry and Geoinformatics, University of Stuttgart, Germany - forename.lastname@ifp.uni-stuttgart.de

KEY WORDS: SLAM, Dense Reconstruction, Mobile Robot, BIM, Construction Sites, Close Range Sensing

ABSTRACT:

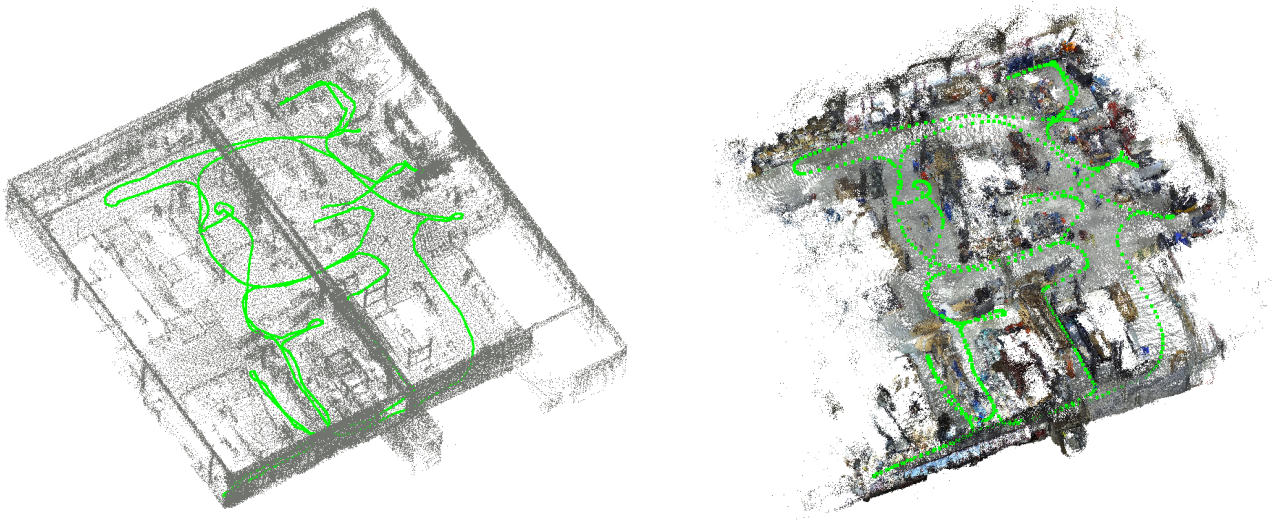
Simultaneous localization and mapping (SLAM), i.e., the reconstruction of the environment represented by a (3D) map and the concurrent pose estimation, has made astonishing progress. Meanwhile, large scale applications aiming at the data collection in complex environments like factory halls or construction sites are becoming feasible. However, in contrast to small scale scenarios with building interiors separated to single rooms, shop floors or construction areas require measures at larger distances in potentially texture less areas under difficult illumination. Pose estimation is further aggravated since no GNSS measures are available as it is usual for such indoor applications. In our work, we realize data collection in a large factory hall by a robot system equipped with four stereo cameras as well as a 3D laser scanner. We apply our state-of-the-art LiDAR and visual SLAM approaches and discuss the respective pros and cons of the different sensor types for trajectory estimation and dense map generation in such an environment. Additionally, dense and accurate depth maps are generated by 3D Gaussian splatting, which we plan to use in the context of our project aiming on the automatic construction and site monitoring.

1. INTRODUCTION

The problem of estimating the exterior orientation of optical sensors and the simultaneous reconstruction of the three-dimensional (3D) environment is commonly known as SfM (Structure from Motion) in Computer Vision and SLAM (Simultaneous Localisation and Mapping) in robotics (Cadena et al., 2016). In the context of SLAM, the estimation of the exterior orientation is performed sequentially and in real-time, frequently aiming at data collection in various indoor scenarios from images and/or laser scanning data. The work presented in our paper is motivated by a project aiming at the monitoring of construction sites in the context of the cluster of Excellence Integrative Computational Design and Construction for Architecture (IntCDC) at the University of Stuttgart (IntCDC, 2024a). Overarching goal of this work is to harness the potential of digital technologies to manufacturing and construction in the building sector. Construction industry has traditionally been a labor-intensive branch, yet it stands to benefit from autonomous robots that promise to deliver construction work that is more accurate and efficient compared to manual or conventional methods. One key task in this context is the direct digital data capture and monitoring of construction sites e.g. for the generation of BIM and digital twins. This documentation is an important application scenario for the geodetic capture of 3D point clouds. While in the past, Terrestrial Laser Scanning (TLS) from fixed stations defined the standard approach, meanwhile mobile systems applying SLAM-based methods are increasingly used. The state-of-the-art for data collection in such scenarios, is e.g. demonstrated by the results of the Hilti SLAM Challenge (Hilti, 2023). For this benchmark, data acquisition was carried out by a hand-held system equipped with an IMU, a multi-camera head and a laser scanner device (Zhang et al., 2023). According to the applied sensor type, SLAM algorithms are categorised into LiDAR and visual SLAM. Current visual SLAM methods can provide dense representations reasonably well, but are typically limited to well-textured environments and rather small spaces, such as single rooms, which are typically captured at short measurement distances. In contrast, real-world applications of large indoor scenes like construction sites and factory halls remain challenging and potentially benefit from the larger measurement range of LiDAR scans. Figure 1 exemplarily visualizes a reconstructed point cloud and trajectory for the LiDAR and Visual SLAM approaches as discussed in our paper.

While approaches limited to a planar 2D space are quite mature and already enter consumer market, methods using 3D point clouds are still topic of research; however, such efforts are strongly pushed by applications connected with autonomous driving and robotics. Visual SLAM algorithms apply monocular, stereo or even RGB-D imagery. Compared to LiDAR sensors, cameras are significantly cheaper and therefore enable a much wider range of applications. Furthermore, the analysis of the captured images is not limited to geometric information extraction during localization and mapping. Due to rich information embedded in RGB imagery, visual SLAM is advantageous for visualization purposes and for semantic segmentation of the environment. On the other hand, the sensor principle of typical RGB-D devices limit their application to close range scenarios, while monocular SLAM frequently suffers from scale drift (Mur-Artal and Tardos, 2017). Even though the angular resolution of LiDAR sensors declines proportional to distance caused by diffraction, in practice, the direct range measurement principle allows for larger scene extent compared to visual SLAM schemes. As a result, current benchmarks (cf. (Hilti, 2023)) show that LiDAR-based methods achieve significantly higher accuracy in trajectory reconstruction. However, this depends strongly on the scanning pattern and the line density as defined by the spatial resolution of the sensor. Further improvements in terms of localization reliability and accuracy for both groups of methods can be achieved by fusion with additional and complementing observations such as IMU or odometer data. In order to support data collection in larger scale environments like construction sites and factory halls, we use measures both from LiDAR and stereo cameras. As an example, we chose a factory hall, in which we monitor indoor construction activities (IntCDC, 2024b). An exemplary result

* Corresponding author



a) LiDAR-SLAM b) Visual SLAM

Figure 1. Visualisation of the resulting point clouds and trajectories for the LiDAR and Visual SLAM approaches presented.

of our results from LiDAR and visual SLAM is already depicted in Figure 1.

The remainder of the paper is organized as follows: the following section on related work first presents current benchmarks aiming at scenarios similar to our task: data collection at large scale complex and dynamic environments. Furthermore, LiDAR and visual SLAM approaches feasible for such applications are introduced. Section 3 then presents data collection by our robot system (cf. Section 3.1) and our processing pipelines for LiDAR and visual SLAM (cf. Sections 3.2 and 3.3). Our results for data collected in a wide-area factory hall are then presented in section 4. Since our future work aims on the combination of LiDAR and visual SLAM, we especially discuss the pros and cons of the respective sensors for data collection in such an environment.

2. RELATED WORK

The following chapter is divided into three sections. The first introduces typical data sets used for SLAM applications and highlights the most important properties. In this context, only public available sources are considered. Subsequently, both established and the latest approaches in the fields of LiDAR and Visual SLAM will be presented.

2.1 Large-Scale SLAM Benchmarks

Within the last years multiple data sets targeting various applications in the area of computer vision, simultaneous localization and mapping and robotics have been created. The majority of resources predominantly focus on capturing compact regions such as small rooms and address either visual or LiDAR SLAM by providing either camera or LiDAR data. One of the most common large-scale multi-modal data collections used to evaluate corresponding SLAM approaches is the KITTI data set (Geiger et al., 2013). It contains stereo images, point clouds acquired by a 360 degree LiDAR and readings from an Inertial Measurement Unit (IMU) attached to a car. Various benchmarks for the calculation of depth maps, odemetry estimations or object detection and tracking tasks have been derived based on this resource. A comparable and even larger collection of driving data including radar information and additional cameras for a 360 degree view is the nuScenes (Caesar et al., 2020) set. Both resources are providing high quality data from various sensors, but are limited to outdoor scenes captured solely on streets. The data collection, which is in parts most closely aligned with our use case, is the HILTI benchmark (Zhang et al., 2023). For three scenes of this collection the data were captured with an mobile robot equipped with a 3D front LiDAR, four RGB-D cameras for a 360 degree surround view and an IMU. Since all scenes were taken on construction sites, even the environmental conditions are comparable. Nevertheless, the focus on the HILTI benchmark is laid on the accurate localization of the robots. Therefore, the trajectories driven are not optimized for the purpose of a detailed reconstruction of the building or the objects around. Furthermore, the robot only moved in an underground car park with controlled lighting conditions, so that there were hardly any glare or reflection effects. In addition, no work was carried out in the robot’s field of vision - apart from the rover and an instructor following it the scene was static.

In the following sections, algorithms for visual and LiDAR SLAM applications connected with our work are presented and described in more detail.

2.2 LiDAR SLAM

Although the field of LiDAR SLAM originates from 2D LiDAR sensors, we will focus on more recent 3D LiDAR SLAM approaches in this overview. An early popular real-time approach of such kind is LOAM (Zhang and Singh, 2014). Here, from the LiDAR point set planar and edge features are derived. The poses of the sensor are determined from a high-frequency odometry function and a low-frequency mapping function by minimizing the errors of corresponding features. Newer approaches also include mechanisms such as Scan Context (Kim and Kim, 2018) or LoGG3D-Net (Vidanapathirana et al., 2022) to detect the revisiting of already mapped locations (loop closures) and to improve consistency through pose graph optimization (Behley and Stachniss, 2018), (Ramezani et al., 2022). In some modern methods, the points are still reduced to edge and planar features

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-209-2024 | © Author(s) 2024. CC BY 4.0 License.

during preprocessing (Shan et al., 2020), while in others the points are combined into disk-like surface elements (Behley and Stachniss, 2018), (Ramezani et al., 2022) or in so-called dense approaches simply (downsampled) points are used (Dellenbach et al., 2022), (Xu et al., 2022), (Vizzo et al., 2023), (Skuddis and Haala, 2024). In modern LiDAR SLAM systems IMU data are used to increase robustness in situations with fast movements and to reduce the orientational drift through gravity estimations (Ramezani et al., 2022), (Shan et al., 2020), (Xu et al., 2022), (Skuddis and Haala, 2024). While in IMU-free LIDAR SLAM approaches it is common to model the trajectory within one scan as linear motion (Dellenbach et al., 2022), (Neuhaus et al., 2019), (Vizzo et al., 2023), IMU supported LiDAR inertial SLAM approaches enable higher resolution trajectory representations, which can result in improved accuracy (Ramezani et al., 2022), (Shan et al., 2020), (Xu et al., 2022), (Skuddis and Haala, 2024). In recent works, for LiDAR SLAM Neural Network representations of the environment have been proposed to obtain more consistent maps (Wiesmann et al., 2023), (Zhong et al., 2023), (Deng et al., 2023). In benchmark datasets, however, they cannot yet achieve the accuracy of conventional methods (Geiger et al., 2013), (Zhang et al., 2023).

2.3 Visual SLAM

With ORB-SLAM (Mur-Artal et al., 2015), an efficient, robust and real-time capable SLAM approach that includes both place recognition capabilities required for loop-closing and global optimizations combined with the ability for lifelong mapping, which is especially required for the exploration of large environments, was introduced. ORB-SLAM and its successors, ORB-SLAM2 (Mur-Artal and Tardos, 2017) and ORB-SLAM3 (Campos et al., 2021) are using hand-crafted feature extraction algorithms and optimization methods focusing on improved tracking accuracy. This results in low computing requirements during operation, but also causes relatively sparse point clouds and a low level of detail in the created maps. More recent approaches such as DROID-SLAM (Teed and Deng, 2021) are integrating Neural Networks and train them on various scenes to further increase the robustness by reducing accumulating drift and the number of failures due to loss of feature track. A fully differentiable method design allows to combine and tune Neural Network Layers, for example, for dense pixel matching or to update the camera poses, with standard algorithms, for instance, to perform global optimizations. In addition, a clear separation in a frontend, which performs time critical tasks on the input stream of the images, and a backend, in which the computing-intensive processes are outsourced, still allows real-time capability. The study by (Zhang et al., 2022) demonstrates the high effectiveness of DROID-SLAM in robotic applications with planar motion.

Most recent approaches are addressing the visualization and representation of the resulting maps. Traditional SLAM methods employ voxel grids, point clouds, or mesh representations as scene representations to construct dense mapping. However, these schemes face serious challenges in obtaining fine-grained dense maps. Methods such as HI-SLAM (Zhang et al., 2024) extend existing concepts to include Neural Radiance Fields (NeRF) (Mildenhall et al., 2020) as 3D representation of the environment. For this purpose, the estimates such as pose and depth values of the keyframes are used to incrementally optimize the corresponding weights of the integrated Neural Network. In addition, by Multiresolution Hash Encoding (Müller et al., 2022) the required training times can be reduced significantly allowing a fast update of the resulting neural map. Recently, 3D Gaussian splats have been proposed as efficient rendering technique of Radiance Fields for high-quality and dense mapping with low memory consumption (Kerbl et al., 2023). Beyond its efficiency for high resolution image rendering, Gaussian splatting holds an explicit geometry scene structure and appearance, benefiting from the exact modeling of scenes representation. This technology has been rapidly applied in several fields, and it seems to be also very promising for subsequent 3D modelling as one long-term goal of our project.

3. SLAM-BASED MAPPING OF A LARGE FACTORY HALL

Within the following sub-chapters the sensor system used to capture the environment is introduced and a exploration of the characteristics of the recorded building is provided (cf. Chapter 3.1). The LiDAR and visual SLAM approaches (cf. Chapter 3.2/3.3) adapted and evaluated within this work are subsequently outlined .

3.1 Sensor Platforms and Data Acquisition

As test environment a part of the Large-Scale Construction Robotics Laboratory (LCRL) of the Cluster of Excellence IntCDC was chosen (IntCDC, 2024b). The building consists of a large construction hall including multiple robotic building part prefabrication plants, instructor work spaces, material depots and traditional fabrication tools. The setup includes wide areas with open space as well as small corridors. Multiple structures built from various materials such as concrete, steel or wood are located within the recording area. To support the evaluation of the resulting point clouds, markers have been placed at different locations within the hall. The data acquisition took place during normal operation so on-site staff was passing the sensors and robots and objects moved during our recording.

a) Overview

b) Detail - Cameras

c) Detail - LiDAR

Figure 2. Images of the robotik platform used for data acquisition.

The sensor platform is a 6-wheeled robotic system providing all the basic functionalities such as power supply, computational resources and driving capabilities (cf. Figure 2). A 3D LiDAR sensor (RoboSense BPearl) with 32 lines and a maximum detection range of 100m (30m@10% NIST) is mounted at the front of the robotic system. In addition, four ZED 2 stereo cameras providing a 360°-surround view of the environment are attached to an extractable tower in the center of the robot. All relative sensor orientations are known from a prior calibration.

Due to logistical reasons, only one third of the hall with the size of approx. $1600m^2$ was explored by the robotic platform. For a most detailed reconstruction of the construction site, all areas that could be reached with the outer dimensions of the robot were entered. To provide a very accurate reference point cloud a Trimble X7 survey grade Terrestrial LiDAR Station was used. During data collection 360 degree scans were captured at eight scanning positions in the hall of the LCRL and merged to a final map by using available reference points.

3.2 DMSA LiDAR SLAM

Within the scope of this work, we investigated three LiDAR SLAM methods to processes our recorded LiDAR / IMU data. In detail, we selected KISS-ICP (Vizzo et al., 2023), a popular LiDAR SLAM method that claims to be easy to integrate and robust, CT-ICP (Dellenbach et al., 2022), a highly accurate LiDAR-only open source algorithm, and our DMSA SLAM (Skuddis and Haala, 2024). While using the standard parameter settings proposed by the corresponding authors, the SLAM methods KISS-ICP and CT-ICP diverged after only a few seconds during processing. We suspect that the methods have difficulties with processing the very sparse LiDAR data of the RoboSense BPearl. Thus, the results presented in more detail in chapter 4 are solely generated by our DMSA SLAM approach.

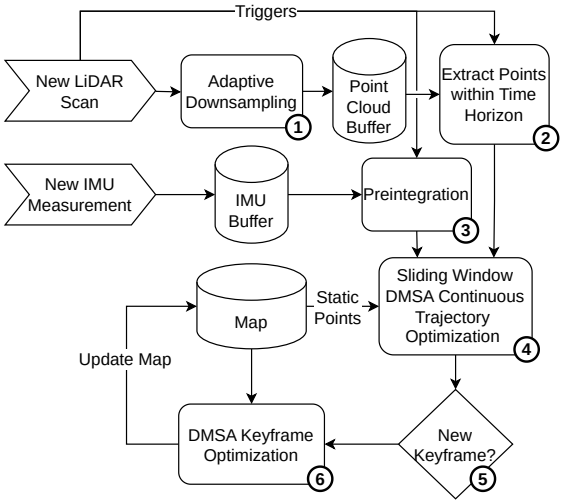


Figure 3. DMSA SLAM overview. Numbers indicate the processing order. Adopted from (Skuddis and Haala, 2024).

In DMSA SLAM, LiDAR points within a sliding time window are optimized together with so called static points and IMU data in a tightly coupled manner. For processing the captured data within this work a sliding window time horizon of 0.8 s is selected. Adaptive downsampling within the preprocessing enables handling of LiDAR data from narrow spaces as well as from spacious places. New keyframes are selected based on overlap and distance thresholds. In addition to LiDAR points belonging to a keyframe, the gravity direction is estimated and added to the keyframe data. When a new keyframe is added to the map, all keyframes with significant overlap to the current keyframe are optimized. Figure 3 gives an overview of the processing steps. For details we refer to (Skuddis and Haala, 2024).

3.3 Dense Multi-Camera RGB-D SLAM Pipeline

In the processing chain of the four camera imagery, we developed a dense visual SLAM pipeline visualized in Figure 4. To fully leverage the 360-degree surround view, the data of each camera is first processed via the DROID-SLAM method (Teed and Deng, 2021). In this stage, keyframes for each camera view are adaptively selected based on the optical flow distance between neighboring images. For these keyframes, dense per-pixel depths are estimated by predicting dense flow and performing bundle adjustment per camera stream to minimize reprojection errors with the predicted flow as references. Subsequently, using the known extrinsics between the cameras, we transform the keyframe poses of each camera into a common coordinate system, with the optical center of the front camera as reference coordinate.

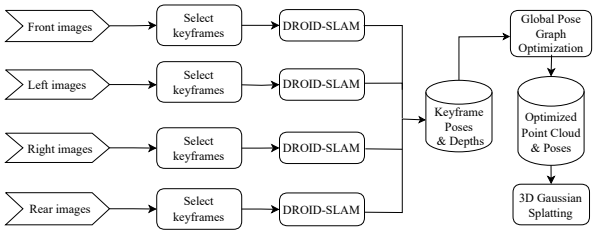


Figure 4. Diagram of dense visual SLAM pipeline.

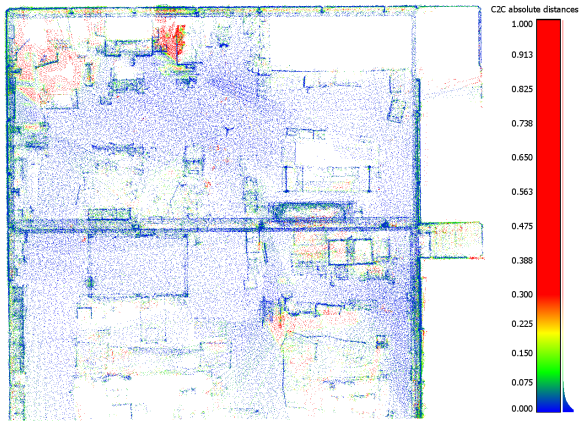
Although this process yields a global map integrating all camera views and observations, inconsistencies may still arise due to remaining errors, such as drifted trajectories, in the parameters estimated individually for each camera - particularly in scenarios where no loop closure is available. To construct a globally consistent map, we combine the keyframes from all four cameras in a joint bundle block adjustment. For this purpose we adopted the global bundle adjustment of the DROID-SLAM backend. This optimization technique identifies loop closures across different camera views and observations at different times. The typical visual SLAM pipelines (Campos et al., 2021, Teed and Deng, 2021) often result in a point cloud map. However, this format may not be sufficient for robots to localize themselves within the map while navigating collision-free, due to the inherent sparseness of point clouds. Additionally, applications such as first-person navigation through a scene along a newly rendered path require realistic images from novel viewpoints. To overcome these challenges, we have adopted the 3D Gaussian Splatting method (Kerbl et al., 2023) for training a radiance field of the scene. This approach utilizes both the estimated poses and the globally consistent point cloud produced in the last stage as initial Gaussian positions. While following the default configuration in (Kerbl et al., 2023), we have made a key enhancement by adding stereo depth for additional supervision, thereby improving geometric reconstruction. Figure 6 presents an example of input stereo depth, which is often incomplete due to issues like low texture or limited stereo baseline. The depth images rendered using our trained 3D Gaussian model address these difficulties effectively, producing fully complete depths with clearer object borders.

4. RESULTS

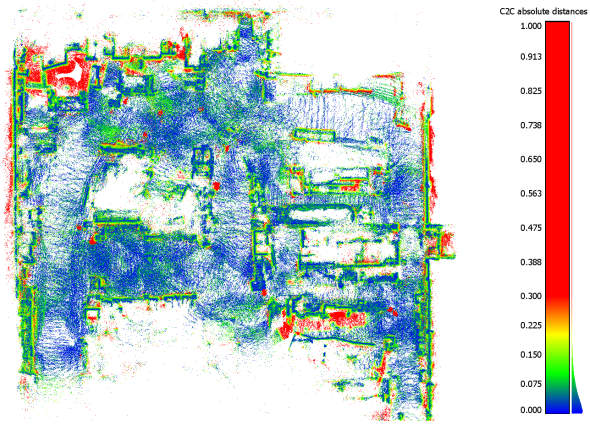
In the following, we discuss the performance of the aforementioned methods step-by-step. While Figure 1a) and b) already presented the resulting point clouds including the determined trajectories from both our LiDAR and visual SLAM methods. Figure 5 further evaluates and visualizes these results based on a comparison to a TLS reference (cf. Figure 5a). To compare the created maps with the TLS reference, the corresponding

points of each cloud were transformed to the reference coordinate system and then finely adjusted using Iterative Closest Point (ICP). The colorization of Figure 5b and 5c is based on the Euclidean distance to the nearest neighbour of the reference point cloud.

a) Reference from TLS



b) Lidar SLAM [m]



c) Visual SLAM [m]

Figure 5. Visualization of the Cloud-to-Cloud (C2C) distance [m] of the presented DMSA LiDAR and RGB-D SLAM pipelines. Colorization based on the minimum distance of the points to the nearest neighbour of the reference point cloud (RGB).

As the graphics illustrate, the point cloud resulting from the LiDAR SLAM approach (cf. Figure 5b) provides a relatively sparse but precise representation of the environment. With DMSA SLAM (Skuddis and Haala, 2024) we were able to successfully process the data from the LiDAR sensor together with the IMU data after a few adaptations to the initially published pipeline/parameters. The alignment of the environment representation created on the basis of Visual SLAM showed that the resulting map was down-scaled by approximately 5% compared to the reference. Scaling issues commonly arise in monocular SLAM, but can be addressed across various applications through the utilization RGB-D cameras (Campos et al., 2021). However, since such cameras frequently apply stereo or structured light, they typically suffer from a restricted measurement range (max. $15 - 20m$) and, compared to LiDAR sensors, less precise depth information. As an example, the applied ZED2 camera features a $120mm$ stereo baseline. According to the manufacturer, the maximum range of the ZED 2 is $20m$ before the depth’s accuracy decreases significantly. Our assumption is that particularly in wide area environments these limitations still contribute to (significant) scale errors. However, to simplify comparisons, for the following evaluations the scale of our visual SLAM result was corrected by a corresponding upscaling of the affected point cloud.

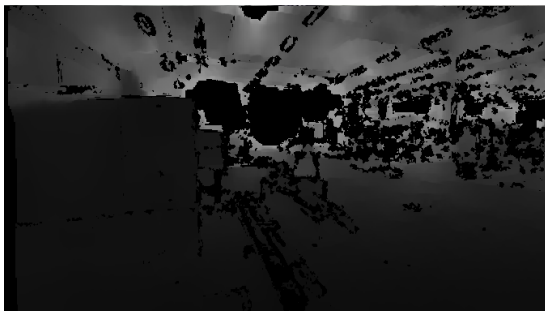
While LiDAR SLAM typically does not suffer from scaling issues and provides reliable range measures at considerable distances, our visual SLAM using all four RGB-D cameras produces a denser point cloud after global optimization. However, it frequently suffers from a higher noise impact. This becomes particularly evident at the right and left walls of the hall, which look quite bumpy compared to the result achieved from LiDAR data. On the one hand, this is caused by the lower reliability of the depth estimates generated from the stereo camera, especially for distant objects, and on the other hand to challenging lighting conditions within the captured environment. Since both methods are not as much effected from shadowing effects as the reference point cloud measured from only a few dedicated scanning positions inside the hall, areas with no information in the reference cloud lead to high distances in the evaluation. These areas effect on the one hand the fine adjustment via ICP and on the other hand restrict possible quality assessments. For this reason we excluded those areas without a sufficient number of observations (cf. red boxes - Figure 5a). Based on the remaining data we received a mean point distance of $\mu = 4,1cm$ ($\sigma = 6,8cm$) for the point cloud created by the presented LiDAR SLAM approach and $\mu = 7,4cm$ ($\sigma = 9,5cm$) for the resulting map of our visual SLAM pipeline.

Objects within the hall that are temporarily stationary (semi-static) and are thus located in the same place over several frames, result particularly in the point clouds of the camera-based method as local maxima (green/red dots) in the computed distances and thus appear as red or green clusters in the point cloud (c.f. Figure 5b/c). Will these objects move from their first observed position, the corresponding points will remain as fragments in the generated point cloud. Examples are the operator following the robot or staff temporarily working at a certain location. Since the cameras of the robot are providing a $360°$-view they are more sensitive to semi-static objects than the LiDAR sensor which is only scanning the front of the robot. In addition, the resolution of the LiDAR is significant lower than the data gathered by the camera. For these reasons, semi-static objects in the point cloud generated by the BPearl sensor appear less frequent in comparison to the camera based approach.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-209-2024 | © Author(s) 2024. CC BY 4.0 License.


a) Input - RGB


b) Rendered - RGB


c) Input - Depth


d) Rendered - Depth


d) Input - semi-static objects


e) Rendered - semi-static objects


f) Input - depth semi-static objects


g) Rendered - depth semi-static objects


h) Input - direct sunlight


i) Rendered - direct sunlight

Figure 6. Comparison of input RGB and depth images with those rendered by our 3D Gaussian splatting model. Our reconstruction notably enhances depth quality and completeness.

The advantages of using 3D Gaussian Splats as representation of the environment is presented in Figure 6. By using Gaussian Splats a nearly photo-realistic representation of the environment can be created and even new perspectives, which have not been visited by the robot during data collection, can be rendered and visualized. Additionally, the combination of multiple observations lead to a significant improvement of the depth information. Thus, even small details such as the tripod of a tachymeter or the outlet of the air-condition system, which are not clearly visible on the depth map generated by the original sensor, become recognisable (cf. Figure 6c/d). The optimization process in the Gaussian splats approach may render certain semi-static objects invisible due to the occasionally redundant observation of the same building parts (cf. Figure 6d/e/f/g), which is also helpful for image rendering at poses suffering from challenging lighting conditions during image acquisition (cf. Figure 6h/i). The efficient rendering methods presented by (Kerbl et al., 2023) also enable the dynamic loading and real-time visualization of large environment models on typical consumer GPUs. We were able to attain update rates exceeding 30 fps using a NVIDIA GeForce RTX 3090 Ti. Furthermore, the improved depth information for each (virtual) station is very helpful for subsequent applications aiming at Augmented Reality or high precision navigation.

5. CONCLUSIONS AND FUTURE WORK

As our main contribution we demonstrated the feasibility of SLAM-based methods for the acquisition and mapping of large factory halls or construction sites. Based on the data from our multi-sensor platform, we confirmed our original assumption that LiDAR sensors provide high geometric accuracy during the reconstruction of large scale environments, while camera-based methods can generate useful details which can be excellently visualised using 3D Gaussian splats. Compared to classic geodetic methods such as TLS, a number of advantages of SLAM-based methods for the mapping of large building interiors can be identified. In general, the flexible data collection of mobile platforms helps to avoid occluded areas especially in complex environments, which is typically labor-intensive for stationary measurement. Even though our SLAM pipeline does not (yet) allow for real-time processing, it already is very time efficient. The analyzed sequence of our factory hall depicted in Figure 1 was captured in 0.3h, while TLS data acquisition took 1h for preparation and to 2h for acquisition. In simple terms, the more complex and finer the geometry of the environment, the greater the time savings for a mobile approach. In combination with a map-based trajectory planning and an automatic guidance even a fully autonomous acquisition process becomes imaginable.

While the achieved (mean) accuracies of our approach $(4 - 8cm)$ are sufficient for navigating through the captured environment, plenty of tasks on construction sites such as the installation of doors, windows or similar built-in parts require more precise measurements. Our future work thus aims on fusing LiDAR and image data while striving to get the best of both worlds - reliable position estimates and geometries from LiDAR SLAM in combination with rich representations of the environment from the visual SLAM. In order to obtain more reliable accuracy estimates for the resulting trajectories and maps, we also plan to integrate reference points located inside the hall into our evaluation process.

Since 3D Gaussian splats are characterized by efficient memory usage and, in combination with the already developed rendering methods, require low computing power, they provide an ideal opportunity of representing large environment models. Due to the explicit form of representation, they are also a good basis for subsequent 3D segmentation tasks (Kim et al., 2024). In our ongoing efforts to enhance the precision and quality of visualizations and 3D point clouds generated by our visual SLAM pipeline, we are also working on refining the joint calibration and photogrammetric assessment of the four cameras. If semantic information is also determined, digital construction plans such as Building Information Model (BIM) or Building and Habitats object Model (BoHM) can also be created or updated on the basis of the collected data. In this way, automatically generated building status reports that compare the actual status with the target status are conceivable.

6. ACKNOWLEDGEMENTS

Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2120/1 – 390831618.

REFERENCES

Behley, J., Stachniss, C., 2018. Efficient surfel-based slam using 3d laser range data in urban environments. *Robotics: Science and Systems*, 2018, 59.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., 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.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving.

Campos, C., Elvira, R., Rodriguez, J. J. G., M. Montiel, J. M., D. Tardos, J., 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 37(6), 1874–1890. http://dx.doi.org/10.1109/TRO.2021.3075644.

Dellenbach, P., Deschaud, J.-E., Jacquet, B., Goulette, F., 2022. Ct-icp: Real-time elastic lidar odometry with loop closure. *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 5580–5586.

Deng, J., Wu, Q., Chen, X., Xia, S., Sun, Z., Liu, G., Yu, W., Pei, L., 2023. Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8218–8227.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*.

Hilti, 2023. Hilti slam challenge. https://hilti-challenge.com/.

IntCDC, 2024a. Cluster of excellence integrative computational design and construction for architecture (intcdc). https://www.intcdc.uni-stuttgart.de/.

IntCDC, 2024b. Intcdc large-scale construction robotics laboratory (lcrl). https://www.intcdc.uni-stuttgart.de/research/research-infrastructure/.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 1–14. https://inria.hal.science/hal-04088161.

Kim, C. M., Wu, M., Kerr, J., Tancik, M., Goldberg, K., Kanazawa, A., 2024. Garfield: Group anything with radiance fields. *arXiv*.

Kim, G., Kim, A., 2018. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 4802–4809.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 405–421.

Mur-Artal, R., Montiel, J. M. M., Tardos, J. D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147–1163. http://dx.doi.org/10.1109/TRO.2015.2463671.

Mur-Artal, R., Tardos, J. D., 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262. http://dx.doi.org/10.1109/TRO.2017.2705103.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4), 1–15. http://dx.doi.org/10.1145/3528223.3530127.

Neuhaus, F., Koß, T., Kohnen, R., Paulus, D., 2019. Mc2slam: Real-time inertial lidar odometry using two-scan motion compensation. *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*, Springer, 60–72.

Ramezani, M., Khosoussi, K., Catt, G., Moghadam, P., Williams, J., Borges, P., Pauling, F., Kottege, N., 2022. Wildcat: Online continuous-time 3d lidar-inertial slam. *arXiv preprint arXiv:2205.12595*.

Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D., 2020. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 5135–5142.

Skuddis, D., Haala, N., 2024. Dmsa - dense multi scan adjustment for lidar inertial odometry and global optimization. To be pubsished in ICRA Proceedings.

Teed, Z., Deng, J., 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (eds), *Advances in Neural Information Processing Systems*, 34, Curran Associates, Inc., 16558–16569.

Vidanapathirana, K., Ramezani, M., Moghadam, P., Sridharan, S., Fookes, C., 2022. Logg3d-net: Locally guided global descriptor learning for 3d place recognition. *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2215–2221.

Vizzo, I., Guadagnino, T., Mersch, B., Wiesmann, L., Behley, J., Stachniss, C., 2023. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters*, 8(2), 1029–1036.

Wiesmann, L., Guadagnino, T., Vizzo, I., Zimmerman, N., Pan, Y., Kuang, H., Behley, J., Stachniss, C., 2023. Locndf: Neural distance field mapping for robot localization. *IEEE Robotics and Automation Letters*.

Xu, W., Cai, Y., He, D., Lin, J., Zhang, F., 2022. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4), 2053–2073.

Zhang, J., Singh, S., 2014. Loam: Lidar odometry and mapping in real-time. *Robotics: Science and systems*, 2number 9, Berkeley, CA, 1–9.

Zhang, L., Helmberger, M., Fu, L. F. T., Wisth, D., Camurri, M., Scaramuzza, D., Fallon, M., 2023. Hilti-Oxford Dataset: A Millimeter-Accurate Benchmark for Simultaneous Localization and Mapping. *IEEE Robotics and Automation Letters*, 8(1), 408-415.

Zhang, W., Sun, T., Wang, S., Cheng, Q., Haala, N., 2024. HI-SLAM: Monocular Real-Time Dense Mapping With Hybrid Implicit Fields. *IEEE Robotics and Automation Letters*, 9(2), 1548-1555.

Zhang, W., Wang, S., Haala, N., 2022. Towards robust indoor visual SLAM and dense reconstruction for mobile robots. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1, 211–219.

Zhong, X., Pan, Y., Behley, J., Stachniss, C., 2023. Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 8371–8377.