A handheld stereo vision and LiDAR system for outdoor dense RGB-D mapping using depth map completion based on learned priors

Michael Bleier, Yijun Yuan, Andreas Nüchter

Computer Science XVII: Robotics, Julius-Maximilians-Universität Würzburg, Germany
(michael.bleier, yijun.yuan, andreas.nuechter)@uni-wuerzburg.de

Keywords: 3D Mapping, LiDAR, Handheld Mapping System, Depth Map Completion, Dense RGB-D, Learned Depth Covariance Function.

Abstract

This paper proposes a handheld mapping system consisting of a stereo camera setup combined with low-cost automotive LiDAR. The prototype system is applicable to various types of mobile monocular, stereo vision and LiDAR data collection and processing. Capturing dense RGB-D data outdoors with low-cost sensors is challenging, especially when low latency is required. Readily available commercial RGB-D sensors are typically limited to a range of less than 10 m, which is too small to capture large outdoor structures. Currently available low-cost automotive LiDAR scanners feature a suitable range but provide only sparse data. To enable low-latency dense RGB-D scans we augment the sparse LiDAR data with the RGB data stream based on learned models. We apply monocular depth estimation based on a single image and apply scale correction based on learned priors and sparse automotive LiDAR scans. Using the laser scan data, accurate metric information is incorporated directly into the scale estimation stage. For validation, the learning-based depth map completion is compared to traditional LiDAR mapping using scan matching on an outdoor data set acquired with the proposed handheld. While the model-based regression of the sparse LiDAR data produces significantly less accurate results in our experiments, it is able to compute dense RGB-D data from a single sparse 3D scan and monocular RGB image with low latency.

1. Introduction

The appearance of RGB-D sensors, such as Microsoft Kinect or Intel Realsense sensors, enabled low-cost, dense 3D capture of indoor scenes at high framerates. Hence, these sensors are also frequently employed for mobile robots or manipulation tasks using robotic grippers. Off-the-shelf RGB-D sensors are typically limited for outdoor applications due to their limited range. In addition, sensors that rely on pattern projection are often degraded by bright sunlight. As a result, capturing low-latency RGB-D data of large-scale outdoor scenes typically requires more expensive sensor setups that combine high-resolution LiDAR scanners with cameras.

Capturing low-latency RGB-D outdoors on large structures is a challenge on a budget. Low-cost automotive laser scanners provide high update rates of typically 10-20 Hz but the point cloud of a single scan is rather sparse. Similarly, feature-based Visual Odometry approaches (Qin et al., 2018, Geneva et al., 2020) provide real-time 3D reconstruction only for the feature points. In contrast, Structure-from-Motion (SfM) techniques using dense matching provide detailed models of large-scale environments but require significant time for post-processing to solve the photogrammetric bundle adjustment over multiple views. Stereo vision with a baseline of 10-40 cm, which is suitable for a handheld system, does not provide sufficient depth accuracy at a typical distance in the range of 10-100 m regularly observed in outdoor scenes.

Moreover, learning-based approaches are used for depth prediction. Recent learning-based monocular depth methods (Yin et al., 2023, Hu et al., 2024, Koch et al., 2018, Li et al., 2017) provide high-quality depth estimation of structures. However, since the estimation is based on a single image it naturally loses scale and intrinsic information and the relative positions between different structures are distorted.

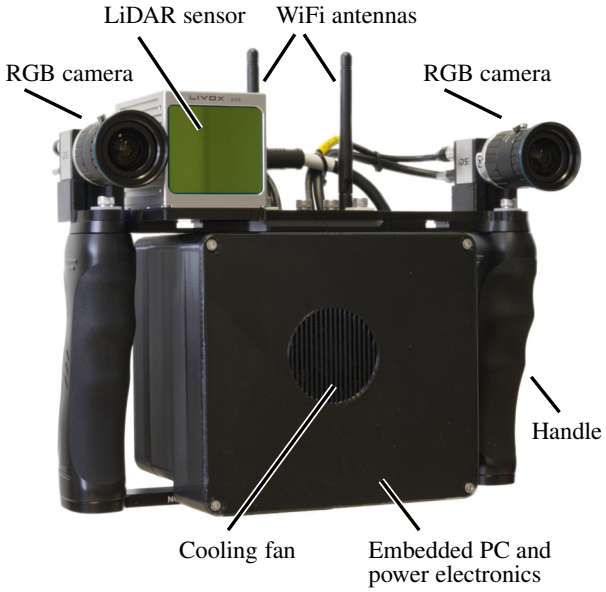


Figure 1. Proposed handheld mapping system with stereo RGB cameras and Livox Avia LiDAR sensor.

This work proposes a handheld mapping system composed of a stereo camera setup combined with an automotive LiDAR. To enable low-latency dense RGB-D scans, we complete the sparse LiDAR data using the RGB data stream based on learned models. Using the LiDAR data accurate metric depth information is incorporated directly into the estimation stage. We employ a state-of-the-art approach for monocular depth image estimation, Metric3Dv2 (Hu et al., 2024), and scale the result based on

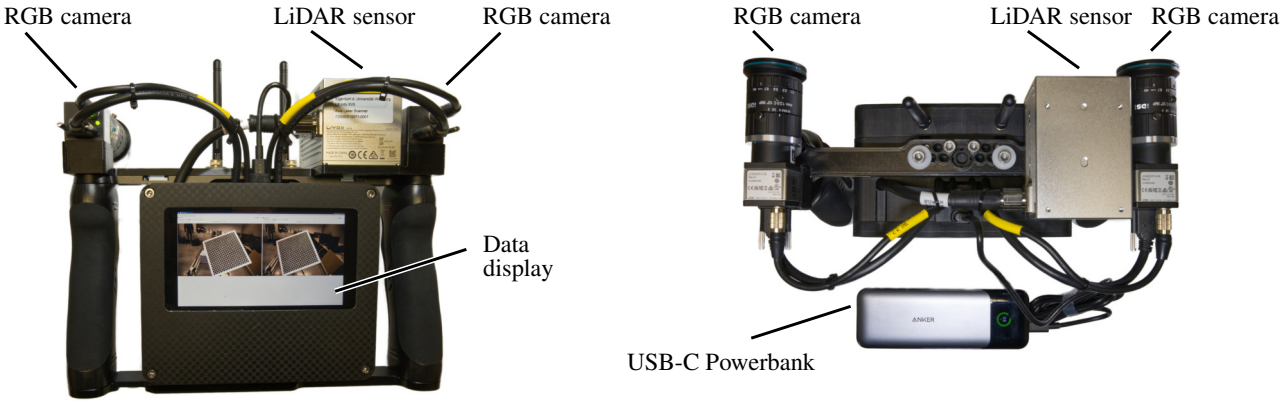


Figure 2. Backside and top of the proposed handheld mapping system for data collection.

learned priors (Yuan et al., 2024). To compare the quality, we capture a dataset with the proposed hardware and compare the learning-based depth map completion to traditional 3D mapping results using scan matching.

The main contributions of this paper are the description of the proposed prototype hardware, a method for completing sparse LiDAR data based on monocular depth estimation, and experiments on real world data for comparing the resulting depth maps with a model of the scene created from registered LiDAR scans.

2. Hardware Setup

Different handheld personal mapping systems have been proposed in the literature. Some setups rely on multi-camera systems, such as GuPho (Torresani et al., 2021, Menna et al., 2022), goScout3D (Bräuer-Burchardt et al., 2023) or Ant3D (Perfetti et al., 2024). Visual Odometry is used to estimate an initial trajectory on the fly, which is post-processed into a dense 3D model using a Structure-from-Motion (SfM) pipeline. This can also be informed by monocular depth estimation (Padkan et al., 2023). Fisheye or wide angle cameras are used to achieve a wide field of view (Holdener et al., 2017), which is particular helpful in constrained spaces.

Other handheld systems use a combination of a monocular camera and LiDAR, such as the R^{3}LIVE handheld (Lin and Zhang, 2022), or multi-sensor setups including automotive LiDAR and RGB-D cameras (Proudman et al., 2022). These handheld devices use Simultaneous Localization and Mapping (SLAM) techniques to register the observations from individual sensors into a consistent global model of the outdoor environment.

2.1 Handheld Mapping System

Our custom-built handheld mapping device, inspired by the previous work, is shown in Figure 1 and Figure 2. It employs low-cost automotive LiDAR and industrial vision cameras. The mechanical structure is based on a commercial camera rig (NEEWER CA016 Video Camera Cage Rig) with the sensors mounted to the top bar. The sensors are attached to the rig using a custom 3D-printed adapter. Below the sensors, a 3D printed enclosure is mounted with an embedded PC for data recording and power supply electronics.

The sensors used are a Livox AVIA Lidar and two IDS U3-30C0CP global-shutter cameras with Sony IMX392 2.35 MPixel RGB CMOS sensors. The LiDAR sensor has a Field of View (FoV) of $70.4° \times 77.2°$. It is a triple-echo sensor with a range of 450 m and a specified range precision of 2 cm. We use the non-repetitive circular scanning mode with 24.000 points per scan. This enables creating more dense scans by accumulating laser scans of an area since the scan pattern changes over time. The cameras are equipped with 4 mm lenses, which results in a similar FoV of $77.3° \times 61.9°$. Hardware signals are used to trigger the two cameras simultaneously and precise timestamps are assigned to both the images and LiDAR scans. We record stereo images at 100 Hz and laser scans at 10 Hz. For the work at hand we extract a synchronized stereo image and LiDAR scan data stream with 10 Hz by selecting the pair of RGB images closest to the receive timestamp of the LiDAR data.

The system is equipped with a LattePanda Sigma embedded PC for data processing. It features a 12-core Intel Core i5-1340P CPU and fast SSD storage to handle the high-frequency image data. The entire setup is powered via USB-C. This allows powering the system either with a USB-C power bank or a USB-C power supply. The system includes a power delivery controller set to 20 V, which allows a power consumption of up to 140 W. When the embedded computer is running at maximum load and all sensors are operating, the combined power requirement is approximately 100 W. In typical operation, a 86 Wh power bank provides about 1.5 to 2 hours of continuous operation.

2.2 Software Architecture

The software architecture is built on top of the Robot Operating System 2 (ROS)[1]. We use the LiDAR drivers provided by the manufacturer and the GenICam interface of the vision cameras. ROS tools are used for data visualization and storage of sensor messages in ROS file format. Network Time Protocol (NTP) is used for synchronizing the clocks over the network. A touchscreen is integrated on the back of the 3D-printed housing for data display. This allows the data to be checked during recording for quality control. In addition, a push button is mounted in the 3D printed case near the rig's handles, which allows the user to start and stop the recording session while holding the system with both hands. Remote control via a WiFi network is also possible.

A standard stereo calibration of the cameras is performed using a calibration board with ChArUco markers. Additionally, we find the co-calibration between the LiDAR coordinate frame

[1] https://www.ros.org/

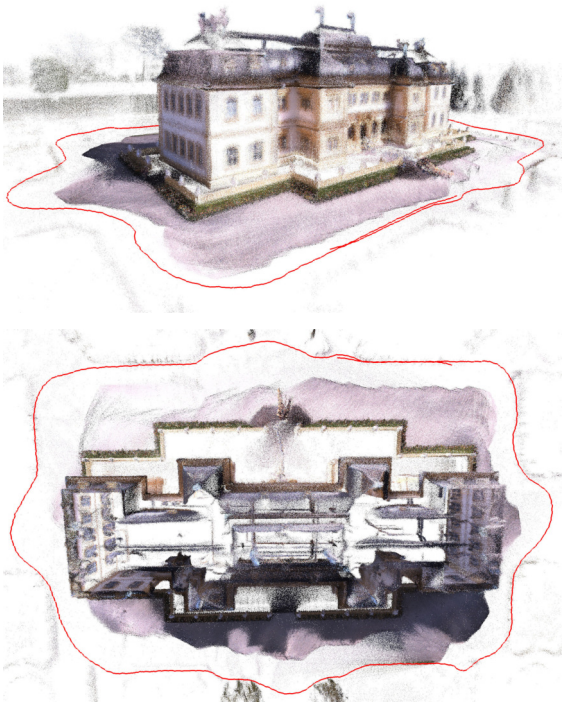


Figure 3. Example of a colored point cloud of Veitshöchheim Palace created using the proposed system. The red line visualizes the trajectory of the handheld mapping system.

and the right camera. As a pre-processing step a stereo rectification of the images is applied and the data processing is computed on the rectified images. For experiments with monocular depth estimation, we only use the right camera of the handheld system since it is closest to the LiDAR sensor. This way the offset between the sensors is small and the viewing direction as well as the field of view are similar.

3. Data Processing

The prototype hardware system allows for different types of data collection and processing, such as monocular, stereo vision, LiDAR scanning, or a combination. Since the focus of the presented work is on the acquisition of large outdoor scenes, the following sections consider monocular RGB data in combination with LiDAR. The stereo setup of the handheld provides scale constraints for dense bundle adjustment in SfM. However, direct acquisition of 3D scans from stereo vision at larger distances is limited given the baseline of approximately 30 cm. Therefore, in the following sections, we look at depth estimation using LiDAR and learning-based monocular depth methods.

3.1 Trajectory Estimation

The Livox LiDAR has a large FoV of more than 70°. Therefore, in urban areas, enough structure is visible in each scan, and scan matching is applicable. We apply straight-forward registration based on the Iterative Closest Point (ICP) algorithm. Octree reduction with a voxel size of 25 cm is applied. To deal with the sparse automotive data we apply scan-to-map registration instead of scan-to-scan registration, which is more stable. We keep a meta scan of the last 3-5 seconds, which corresponds to 30-50 scans. Every new incoming scan is registered against this scan group. Since the automotive scanner has a changing scan pattern, a fixed number of scans for creating the meta scan is suitable. Moreover, we apply distance or movement-based subsampling, such that we do not aggregate a high number of scans for the same area when standing still. For large trajectories, loop-closing based on (Borrmann et al., 2008) is applied.

Figure 3 shows an example trajectory created using ICP-based scan matching and the resulting point cloud. The point cloud is colored using the RGB images. The trajectory of the handheld system is visualized as a red line.

3.2 Depth Map Completion of Sparse LiDAR Data

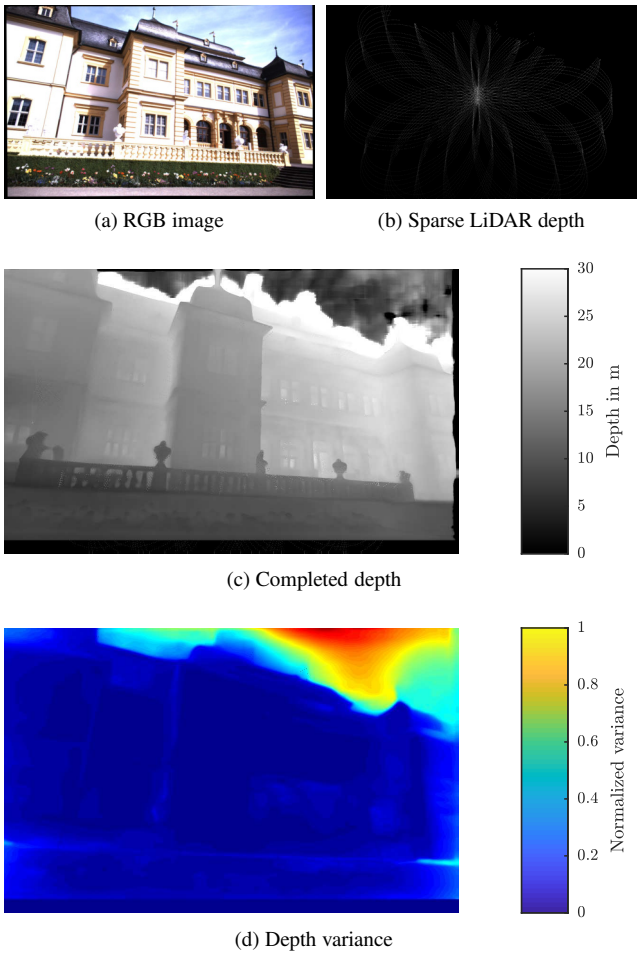
Most commonly, stereo reconstruction is performed from multi-view setups, for example, by triangulating 3D points from corresponding 2D image correspondences from two different cameras, or by SfM using a single camera. Recent learning-based monocular depth methods (Yin et al., 2023, Hu et al., 2024, Koch et al., 2018, Li et al., 2017) provide high-quality depth estimation of structures from single RGB images. Current models, are able to preserve edges and planar regions and produce s consistent depth that accurately captures the structure of the scene. We use the Metric3Dv2 (Hu et al., 2024) model for monocular depth estimation.

However, monocular methods do not accurately recover the metric scale. In addition, similar to human vision the methods suffer from ambiguous illusions where alternative interpretations of depth cues exist. Metric3Dv2 addresses some of these issues by using a canonical camera transformation to make the learned model more robust to changes in, for example, focal length. Still, the structure of the scene is not always accurately captured. We constrain the solution and scale the monocular depth accurately using the true measurements of the LiDAR data. However, we only have sparse depth measurements available from laser scans. Therefore, we need an approach to regress the depth image adjustment from LiDAR to apply it to the entire depth image extracted from monocular depth methods. And we need to do this in a way that takes into account the structure of the scene.

Here, we apply our method ScaleCov (Yuan et al., 2024), which supports sparse LiDAR depth completion as shown in Figure 4. The top row shows the input monocular RGB image and the sparse LiDAR depth image (single scan of the Livox Avia). The middle image shows the resulting completed depth map. Note that the applied color map is the same for the sparse LiDAR depth and the completed depth. Additionally, the bottom image visualizes the depth variance. Note that parts of the image that cannot be assigned a valid depth, such as the sky, also have a high variance. Therefore, we do not trust the final completed depth if the variance is above a certain threshold.

ScaleCov uses a learned covariance function as a prior similar to DepthCov (Dexheimer and Davison, 2023). However, unlike DepthCov, we do not directly regress the depth image. With sparse LiDAR input, this leads to poor estimates for certain structures without any true depth observations. Regressing the true measurements for these structures without any observations leads to arbitrary results and does not accurately capture the structure of the scene.

Therefore, we use the monocular depth estimation from Metric3Dv2 as a starting point. We then compute a sparse scale adjustment by computing scale values for all image points where



(a) RGB image (b) Sparse LiDAR depth

(c) Completed depth

(d) Depth variance

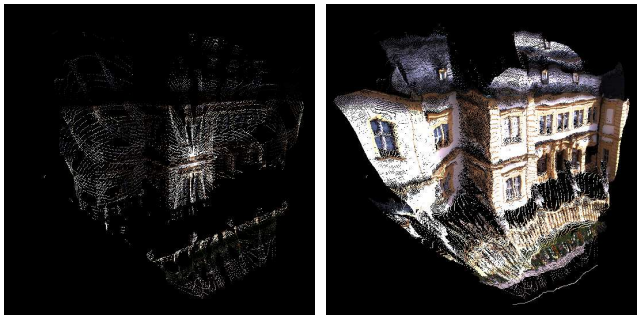
Figure 4. Completed LiDAR depth map using ScaleCov. Top: Input RGB image and sparse LiDAR depth map. Middle: Completed depth map. Bottom: Normalized variance.

the depth computed from LiDAR points overlaps the estimated depth. For these depth image points, we have a true measurement and compute an accurate scale adjustment. This results in a sparse scale image to adjust the monocular depth image. We regress this scale image on the true LiDAR measurements using a deep-learned covariance function in a Gaussian process regression. Regressing the scale to a structure without any true observations from the laser scan is less problematic because the structure estimated by Metric3Dv2 is still preserved.

Figure 5 shows the completed point cloud created from the depth image in Figure 4 side by side with the sparse LiDAR data. In this example, the overall scale is correctly adjusted. However, the parts of the image with little or no true LiDAR observations show large errors. For example, note that the window near the left edge of the completed point cloud is not straight. Therefore, we use the completed depth only for the areas covered by the true LiDAR measurements and do not trust the completed depth outside the FoV of the LiDAR scanner. This improves the consistency of the generated RGB-D frames.

4. Experiments

We apply the proposed handheld system for mobile mapping. Figure 6 shows example point clouds of Veitshöchheim Palace. The data is collected with the handheld system and an average



(a) Sparse LiDAR point cloud (b) Completed point cloud

Figure 5. Comparision of sparse input and completed point cloud. Left: Sparse LiDAR point cloud with 17.627 points. Right: Completed point cloud with 401.687 points.

walking speed of 0.84 $\frac{m}{s}$. During data collection the sensors always point towards the captured object. The trajectory around Veitshöchheim Palace is 251 m long and was captured in 5 min. It consists of 3000 LiDAR scans and RGB images.

The images in the top row of Figure 6 show the results of a point cloud created by scan matching from the full dataset, including all 3000 scans. There are about 48.8 million points in the point cloud. The points are colored using the RGB images.

To compare the result of the depth map completion we select a subset of 15 scans evenly distributed along the trajectory. Using only 15 laser scans the sparse result is visualized in the middle is obtained. This point cloud has approximately 0.3 million points. If we apply the proposed depth map completion to the same 15 frames, we get the result in the bottom row. The completed point cloud has about 5.7 million points. After applying ScaleCov the overall scale of the resulting point clouds is metrically correct. However, we observe some distortions and outliers in the final data. In particular, the smaller structures, such as the windows of the palace, show errors in the regressed depth.

Figure 7 visualizes a comparison between the reference cloud created using scan matching from 3000 scans and the result obtained from learning-based depth map completion with only 15 input frames. The point cloud on the left is colored by the distance between the reference and completed point cloud. Note that the overall scale of the palace is correctly captured. Large errors are observed especially in areas with few true measurements, such as the roof of the palace or the ground in front of the palace. Looking at the error histogram of the cloud-to-cloud distance on the left in Figure 7, we observe that most of the errors are below 20 cm. However, coarse errors in the depth structure and outliers are present in the completed point cloud.

Considering accuracy, scan matching of laser scans or computing Dense Bundle Adjustment on multiple views leads to better results. However, if only sparse depth data is available or a low latency is required, the proposed depth completion produces correctly scaled dense RBG-D frames with only a single RGB image and a sparse laser scan as input. The model inference for monocular depth estimation and completion of the scale map is computed in 200-500 ms using a single desktop GPU. This latency is similar to stereo vision processing pipelines, such as publicly available GPU implementations of Semi Global Matching (Hirschmüller, 2008, fixstars Development Team, 2024).

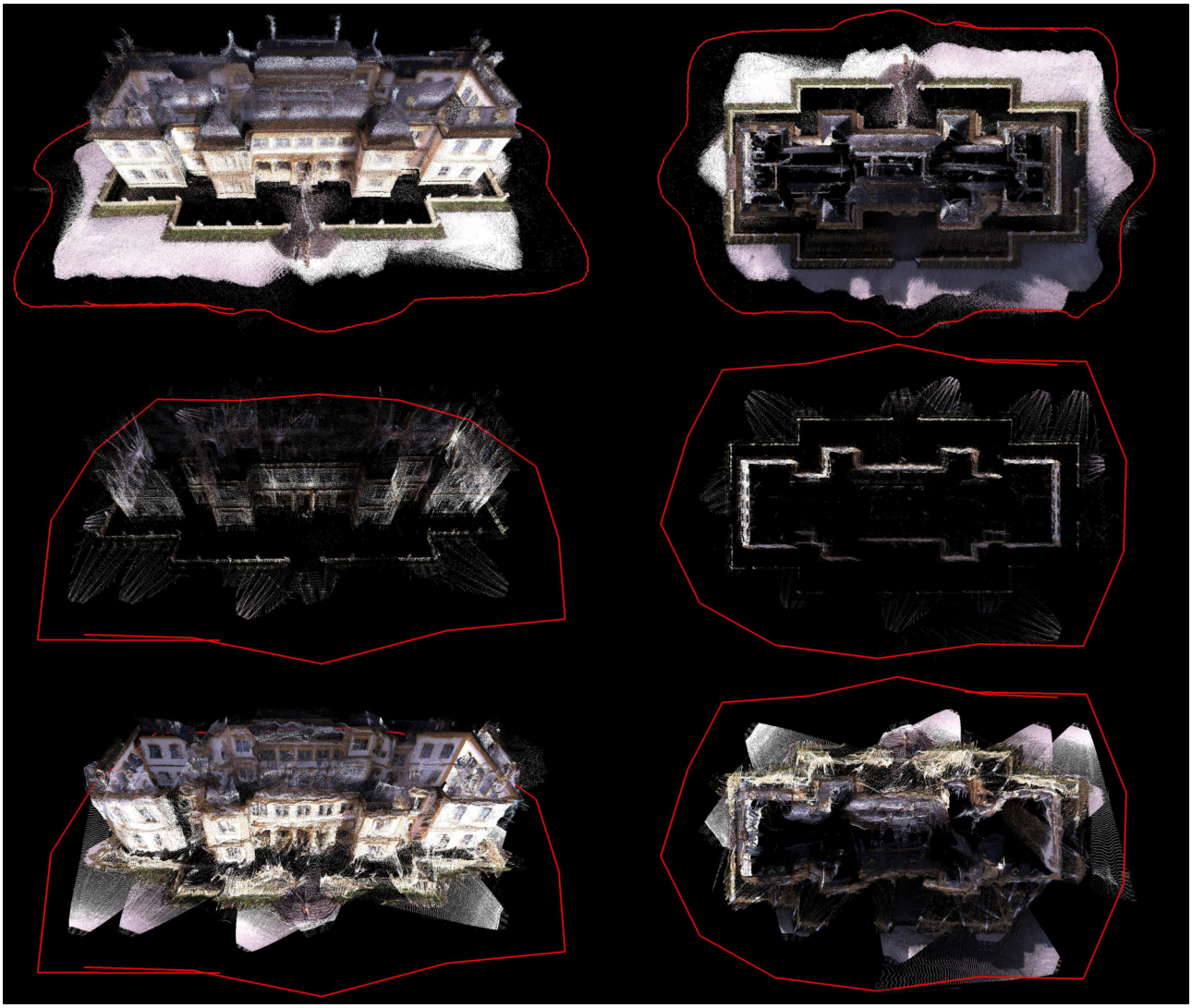


Figure 6. Resulting point clouds of Veitshöchheim Palace with the trajectory of the handheld system visualized as a red line. Top: Aggregated LiDAR point cloud from 3000 scans as a reference. Middle: Sparse LiDAR point cloud from 15 scans. Bottom: Result obtained using depth map completion based on the same 15 scan positions.

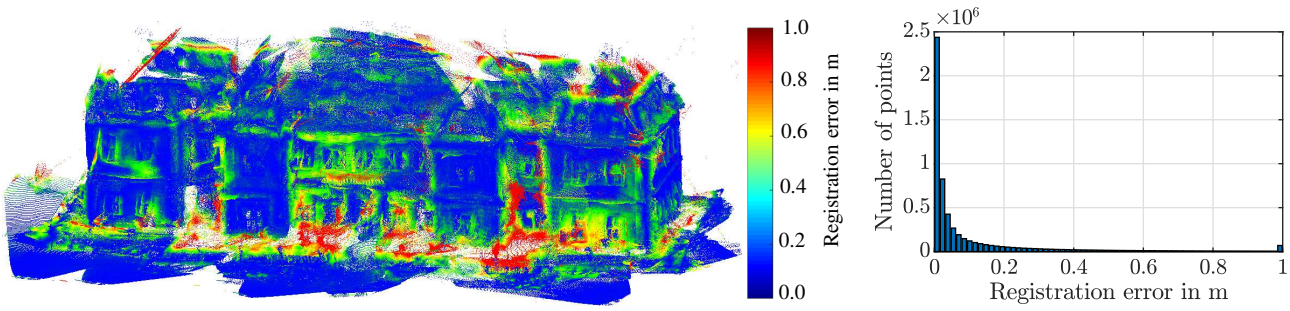


Figure 7. Point cloud of the result from 15 scans using depth map completion compared to the reference scan created using scan matching from 3000 scans. Left: Completed point cloud colored by distance to the reference cloud with corresponding colorbar. Right: Corresponding error histogram of the cloud-to-cloud distance.

https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-31-2024

5. Conclusions

In summary, this work provides details on the hardware setup and processing pipeline of a handheld mapping system, which leverages low-cost automotive LiDAR in combination with a stereo camera setup. We apply monocular depth estimation combined with scale correction based on learned priors. In contrast to aggregating LiDAR scans or applying Dense Bundle Adjustment in a window of multiple views, the proposed approach is able to generate dense RGB-D scans from a single sparse laser scan and a single RGB image.

In addition, we provide an evaluation of depth estimation using learning-based approaches in the context of depth map completion of sparse LiDAR data. We provide real-world results on datasets acquired using the proposed mapping system. The evaluation shows that the model-based regression of the sparse LiDAR data produces significantly less accurate results in our experiments. However, when latency is important or only sparse data is available, the method is able to compute dense RGB-D data with correct scale from a single sparse 3D scan and a monocular RGB image with latencies similar to stereo vision processing pipelines.

In this work, we focused on the regressing of single sparse depth maps. Given a moving handheld system the result can be further improved by multi-view fusion, such as scan aggregation and applying Dense Bundle Adjustment in a local window of input frames, using the last collected LiDAR and RGB frames.

Acknowledgments

The authors acknowledge the support of this work from the Elite Network of Bavaria for the academic program Satellite Technology - Advanced Space Systems.

References

Borrmann, D., Elseberg, J., Lingemann, K., Nüchter, A., Hertzberg, J., 2008. Globally consistent 3D mapping with scan matching. *Robotics and Autonomous Systems*, 56(2), 130–142. doi.org/10.1016/j.robot.2007.07.002.

Bräuer-Burchardt, C., Preißler, M., Ramm, R., Breitbarth, A., Dittmann, J. T., Munkelt, C., Verhoek, M., Kühmstedt, P., Notni, G., 2023. Mobile 3D sensor for documenting maintenance processes of large complex structures. *Engineering for a changing world: Proceedings : 60th ISC, Ilmenau Scientific Colloquium, Technische Universität Ilmenau, September 04-08, 2023*. doi.org/10.22032/dbt.58857.

Dexheimer, E., Davison, A. J., 2023. Learning a depth covariance function. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi.org/10.1109/CVPR52729.2023.01261.

fixstars Development Team, 2024. libSGM: a CUDA implementation performing Semi-Global Matching. GitHub. `https://github.com/fixstars/libSGM`. Accessed 18 October 2024.

Geneva, P., Eckenhoff, K., Lee, W., Yang, Y., Huang, G., 2020. OpenVINS: A research platform for visual-inertial estimation. *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France. doi.org/10.1109/ICRA40945.2020.9196524.

Hirschmüller, H., 2008. Stereo Processing by Semi-Global Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328–341. doi.org/10.1109/tpami.2007.1166.

Holdener, D., Nebiker, S., Blaser, S., 2017. Design and implementation of a novel portable 360 stereo camera system with low-cost action cameras. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 105–110. doi.org/10.5194/isprs-archives-XLII-2-W8-105-2017.

Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S., 2024. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10579–10596. doi.org/10.1109/TPAMI.2024.3444912.

Koch, T., Liebel, L., Fraundorfer, F., Korner, M., 2018. Evaluation of CNN-based single-image depth estimation methods. *Proceedings of the European Conference on Computer Vision (ECCV) 2018 Workshops*, Springer International Publishing, Cham, 331–348. doi.org/10.1007/978-3-030-11015-4_25.

Li, J., Klein, R., Yao, A., 2017. A two-streamed network for estimating fine-scaled depth maps from single RGB images. *Proceedings of the IEEE International Conference on Computer Vision*, 3372–3380. doi.org/10.1109/ICCV.2017.365.

Lin, J., Zhang, F., 2022. R^3LIVE: A robust, real-time, RGB-colored, LiDAR-Inertial-Visual tightly-coupled state estimation and mapping package. *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 10672–10678. doi.org/10.1109/ICRA46639.2022.9811935.

Menna, F., Torresani, A., Battisti, R., Nocerino, E., Remondino, F., 2022. A modular and low-cost portable VSLAM system for real-time 3D mapping: From indoor and outdoor spaces to underwater environments. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W1-2022, 153-162. doi.org/10.5194/isprs-archives-XLVIII-2-W1-2022-153-2022.

Padkan, N., Battisti, R., Menna, F., Remondino, F. et al., 2023. Deep learning to support 3D mapping capabilities of a portable VSLAM-based system. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1), 363–370. doi.org/10.5194/isprs-archives-XLVIII-1-W1-2023-363-2023.

Perfetti, L., Fassi, F., Vassena, G., 2024. Ant3D—A Fisheye Multi-Camera System to Survey Narrow Spaces. *Sensors*, 24(13). doi.org/10.3390/s24134177.

Proudman, A., Ramezani, M., Digumarti, S. T., Chebrolu, N., Fallon, M., 2022. Towards real-time forest inventory using handheld LiDAR. *Robotics and Autonomous Systems*, 157, 104240. doi.org/10.1016/j.robot.2022.104240.

Qin, T., Li, P., Shen, S., 2018. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4), 1004-1020. doi.org/10.1109/TRO.2018.2853729.

Torresani, A., Menna, F., Battisti, R., Remondino, F., 2021. A V-SLAM Guided and Portable System for Photogrammetric Applications. *Remote Sensing*, 13(12). doi.org/10.3390/rs13122351.

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C., 2023. Metric3D: Towards zero-shot metric 3D prediction from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9043–9053. doi.org/10.1109/ICCV51070.2023.00830.

Yuan, Y., Bleier, M., Nüchter, A., 2024. SceneFactory: A Workflow-centric and Unified Framework for Incremental Scene Modeling. *arXiv preprint*. doi.org/10.48550/arXiv.2405.07847.