# Integrated Multi-Stereo Camera System for Robust Indoor Localization with Temporal Fusion

Faezeh Mortazavi[1*], Alexander Kuzminykh[2], Volker Ahlers[2], Claus Brenner[1], Monika Sester[1]

[1] Institute of Cartography and GeoInformation, Leibniz Universität Hannover, Germany
- first name.last name@ikg.uni-hannover.de
[2] Data|H – Institute for Applied Data Science Hannover, University of Applied Sciences and Arts Hannover, Germany
- first name.last name@hs-hannover.de

**Technical Commission II**

**Keywords:** Indoor Localization, Point Cloud, Stereo Camera, LiDAR Sensor, Voxelization

**Abstract**

This paper presents a novel multi-stereo camera system for robust indoor localization, leveraging point cloud data and temporal fusion techniques. The system integrates three synchronized stereo cameras to capture point clouds from multiple angles, enhancing coverage and improving point cloud density in complex indoor environments. By combining data from different perspectives and accumulating point clouds over time, the method mitigates the limitations in the short range of point clouds derived from stereo cameras, ensuring broader coverage for effective localization. To manage the computational complexity of large-scale point clouds and reduce noise in accumulated data, voxelization is applied to downsample the point clouds while preserving key geometric features. The localization process is driven by a predictive point cloud odometry method, refined through the Iterative Closest Point (ICP) algorithm. Experimental results demonstrate the system's ability to achieve accurate localization within a pre-built LiDAR map. This study highlights the feasibility of using low-cost stereo camera systems as an alternative to LiDAR-based solutions for indoor localization.

## 1. Introduction

Indoor localization is a critical component in a variety of applications, ranging from robotics and autonomous navigation to augmented reality and indoor mapping. Traditional methods often rely on high-cost sensors like LiDAR to generate dense and accurate point clouds for localization. However, the use of low-cost sensors, such as stereo cameras, offers a more accessible alternative, though it comes with significant challenges, including limited range, coverage, and accuracy.

Despite the affordability and accessibility of stereo cameras, their use in large indoor environments, particularly for accurate point cloud localization, presents considerable difficulties. In large indoor environments, a single stereo camera may not provide the necessary detail, particularly when objects are distant, often resulting in the capture of primarily ground points, which are insufficient for robust point cloud localization. These challenges call for creative solutions that can fully leverage the capabilities of stereo cameras while overcoming their inherent limitations.

The field of indoor localization has seen significant advancements with the integration of stereo vision and point cloud technologies, particularly in applications like autonomous navigation in robotics. One of the most straightforward approaches for point cloud localization is to use the same type of data for both the pre-built map and subsequent measurements for localization tasks (Suzuki et al., 2010; Ruchti et al., 2015). Techniques like Iterative Closest Point (ICP) (Segal et al., 2009) are widely employed for point cloud registration because of their effectiveness in aligning point cloud data.

In recent years, there has been a growing interest in using low-cost sensors such as stereo cameras for localization, offering an accessible alternative to high-cost LiDAR systems. Given the high costs associated with LiDAR systems, recent research has increasingly focused on leveraging more cost-effective alternatives, such as stereo cameras, to achieve comparable localization results while reducing expenses (Caselitz et al., 2016; Xu et al., 2017). For instance, Han et al. (2019) developed a method that enhances stereo camera localization by incorporating constraints from pre-existing LiDAR maps, effectively improving accuracy in complex environments. In line with this trend, the work by Kim et al. (2018) explores the integration of stereo cameras for localization within 3D LiDAR maps. Our previous work (Mortazavi et al., 2023) demonstrated that accumulating LiDAR scans allows for the creation of a denser and more detailed representation of the environment, which significantly enhances global localization accuracy. Accumulation, in this context, refers to combining consecutive or overlapping LiDAR scans to increase the level of detail in the resulting point cloud. Building on this idea, our current research aims to investigate whether similar aggregation strategies can be applied to point clouds generated from stereo cameras, thereby achieving cost-effective and accurate localization.

This paper proposes a novel multi-stereo camera system designed to enhance point cloud localization within indoor maps. We build upon earlier ideas of mounting stereo cameras to a forklift (Kuzminykh et al. 2023), but simplify the setup by a considerable amount. The system employs three stereo cameras facing different directions for having a wide field of view, collecting data from the front, left, and right perspectives. By combining the point clouds generated from these cameras and integrating data across multiple timestamps, the system achieves a wider coverage area and improved point cloud density (Figure 1). Another key reason for merging these point clouds is to address the limitations posed by the short range of point clouds

derived from stereo cameras. In some instances, there may be no objects in the immediate vicinity of one camera, but by integrating data from cameras facing different directions, we can ensure that sufficient environmental features are captured from other angles. This increases the likelihood of having adequate data for accurate localization within the map. In addition, the reliability of point clouds generated by stereo cameras can be significantly improved, enabling low-cost options to become effective for high-accuracy indoor localization. This enhanced point cloud is then utilized within our localization approach to accurately position within the pre-built map.

The key contributions of this work include:

1. The design and implementation of a multi-stereo camera setup for improved point cloud generation.
2. The development of a temporal fusion technique that combines point clouds across time, using methods such as ICP.
3. The integration of the enhanced point cloud into a localization approach, with an experimental evaluation demonstrating the effectiveness of the proposed system in achieving robust indoor localization within a pre-built map.

## 2. Methodology

### 2.1 Dataset and Experimental Setup

Two datasets were created for development and demonstration of our experimental system: a reference point cloud by LiDAR and video streams, consisting of color and depth information, by our multi-stereo camera setup. The point clouds and video streams were captured indoors within an area of around $160\,\mathrm{m} \times 130\,\mathrm{m}$ in between the construction phase and the official begin of the fair "Hannover Messe 2024" in Hanover, Germany. The data shows the interior of a fair hall, including the facility, presentation booths, roll-up banners and more. Occasionally, dynamic objects such as working staff were recorded; however, the data is primarily composed of static geometry.

Given the well-known characteristics of the two sensor types, LiDAR data provides higher precision, accuracy, and range per frame, while stereo depth data, though limited in depth range and accuracy, offers the advantage of capturing color information and achieving a higher data density within the field of view, as well as a higher frame rate.

In the following, we refer to the point cloud collected by LiDAR after a full rotation of the sensor as a frame. Similarly, the recorded color and depth image data captured by our camera setup at a single time instance, as well as the corresponding extracted point cloud, are also referred to as frames.

**2.1.1 LiDAR-Based Reference Map :** The LiDAR data was collected using a Velodyne Puck LITE, a mechanical, spinning LiDAR sensor equipped with 16 vertically aligned laser emitters. The sensor was mounted on a photography tripod, which was secured to a push cart to maintain a consistent sensor height throughout the data collection. Data was collected for 9 minutes, starting from an entrance to the hall and moving at a steady pace through the environment, eventually returning to the starting point to facilitate loop closure optimization during later processing. Due to the absence of suspension on the push cart and the rigid tripod setup, the mounted sensor experienced significant vibrations, which could potentially affect the quality of the collected data. The captured point cloud stream was processed using standard SLAM techniques to generate a comprehensive point cloud representation of the area of interest. While ubiquitous coverage of the entire hall was not intended, the region of interest for our experiments was adequately covered. Data collection and processing were performed using LidarView and CloudCompare, resulting in a point cloud that appears reasonably accurate, with no obvious deformations or visible artifacts. The point cloud primarily consists of spatial positions with corresponding laser return intensities (Figure 2).

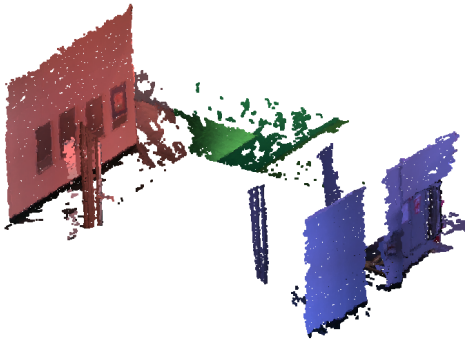


Figure 1. Points collected by the front camera (green) and by the left and right cameras (red, blue). The front camera has only ground points in its field of view and range.

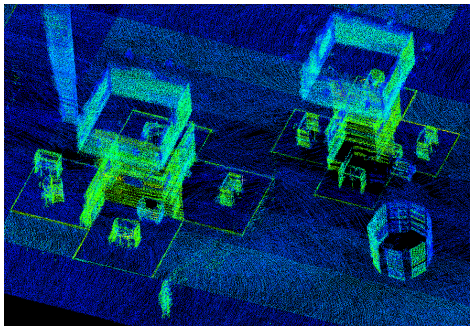


Figure 2. Captured scene via LiDAR, showing a steel pole (top-left), two presentation booths (center), roll-up banners (bottom-right) and a moving person (bottom).

**2.1.2 Stereo Camera Setup for Localization :** The multi-stereo data was collected using a setup of three Intel RealSense D455 cameras, which estimate depth through stereo vision with the support of an infrared (IR) dot pattern projector. The cameras were mounted on a custom-designed, 3D-printed platform to facilitate seamless data integration (Figure 3). When mounted on the platform, all three cameras were aligned in a common plane. The camera setup consists of three cameras: one fac-

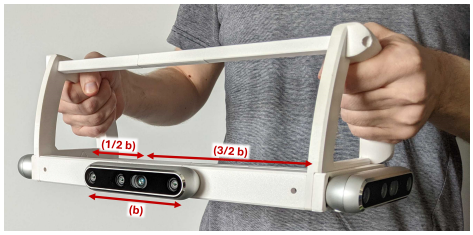



Figure 3. Three stereo depth cameras with baseline $b$ attached to a portable, hand-holdable platform, excluding cables that would be at the bottom.

ing forward and two positioned on the left and right sides. The side cameras are rotated by approximately $\pm 87^{\circ}$ relative to the front camera, a configuration determined by the specifications to achieve a seamless field of view. The front camera is positioned with a stereo baseline $b$ of $95\,\mathrm{mm}$. The side cameras are shifted by $(3/2)b$ to the sides and $(1/2)b$ towards the back relative to the front camera. This arrangement was chosen to facilitate easy and ergonomic usage of the sensor platform rather than for data collection considerations. If the side cameras were positioned further apart, the gaps in their overlapping fields of view with the front camera would increase. This would lead to a higher likelihood of capturing unconnected geometries, complicating the registration of frames between cameras. Maintaining overlapping fields of view is crucial for our localization process, as it enables effective integration and accurate alignment of point cloud data from multiple perspectives.

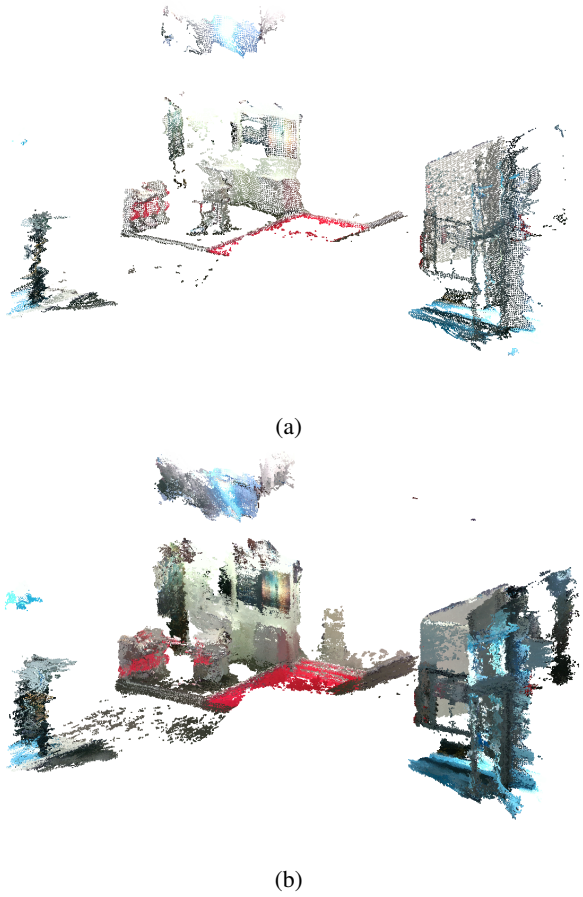
Starting and ending at the same position as the LiDAR dataset, we recorded data for two minutes while moving through the scene with the hand-held camera platform. The recording duration was limited due to storage constraints. Since the platform lacks stabilization and is operated manually, the camera footage exhibits some shakiness; however, we maintained a smooth and steady movement as much as possible. Although the area covered in this run is smaller than that of the LiDAR dataset, it includes enough geometric variation to serve the purpose of our study.

The cameras recorded data using the Intel RealSense Viewer at HD resolution and 30 frames per second (FPS), with the highest quality depth settings applied. This recording session took place half an hour after the LiDAR run, leading to slight differences between the datasets, particularly in non-static objects.

### 2.2 Data Fusion and Processing

Our approach towards localization of the camera data is driven by the aspiration of collecting real-time, spatio-temporal data with high coverage and for low implementation cost. Combining low-cost sensors with low-cost edge computers, such as Raspberry Pi or NVIDIA Jetson Nano, provides the hardware basis for applications with the said goal. Nevertheless, the limitations of low-cost stereo depth cameras in terms of range and precision, as well as the computational constraints of these edge devices, necessitate a more advanced and integrated system architecture. A viable strategy involves distributing computational tasks between local and external resources. In this paper, we focus specifically on the logical separation of the localization algorithm while excluding considerations of the physical distribution of local and external computations.

**2.2.1 Data Fusion :** The recorded video streams are processed using the Intel RealSense SDK, including the provided post-processig filters. These recordings are synchronized at the software level and converted into streams of point clouds, derived from the corresponding color and depth information. At each time instance, each stream generates its most recent point cloud, which is transformed based on the respective sensor location to form a stream of fused point clouds. Only fused frames containing new data from all three cameras are utilized. Consequently, to maintain software-level synchronization, frames from individual cameras may be dropped when necessary. This resulting stream of fused frames, along with the generated reference map, constitutes the primary input data for our work. To expand the coverage area and improve localization performance, fused frames are accumulated as the platform's movement is estimated. For this purpose, each subsequent frame is registered using ICP and its transformed points are appended to the initial frame. Each following frame and its corresponding data are then registered and appended to the progressively accumulated point cloud, while continuously tracking the estimated movement of the platform (Figure 4). After a certain number of accumulation steps, the current accumulated point cloud is passed to the subsequent processing stages of the system, while the accumulation process is restarted with the next set of frames. This overall fusion and accumulation procedure is implemented using the Open3D library.



(a)

(b)

Figure 4. Single frame (a) and 30 accumulated frames (b). While accumulation increases coverage and provides more structural detail for localization in a reference map, it can also amplify alignment errors over time due to sensor noise. .

An interesting aspect of this process is the determination of the number of frames to be accumulated. In initial implementations, a fixed number of frames is used; however, a more sophisticated approach would involve dynamically adjusting this number based on the velocity of the sensor platform. Additionally, various forms of supplementary information and methods, such as IMU data and visual odometry, could be incorporated to further refine the process.

The main purpose of this data fusion and accumulation is to provide a denser point cloud with higher coverage for global localization in a reference map. The quality of the accumulation process, in particular, is influenced by several factors: First, the registration of point clouds is designed with performance considerations in mind, as this part of the system is intended to run on an edge device. Second, the stereo data, which serves as input for the process, often includes significant sensor

noise and occasional distortions due to factors such as observed geometry and lighting conditions. These issues can cause the accumulating point cloud to decrease in precision over time as more data is appended. Given a selected moment in a stereo camera recording with minimal movement and observing a flat surface, the registration process contributes less to the overall error. In this scenario, the measured Root Mean Square Error (RMSE) at the flat surface increases from 4.63 cm to 5.76 cm after accumulating 30 frames. Repeating this evaluation on another flat surface, but this time with movement involving both translation and rapid rotation of the sensor platform, the measured RMSE increases from 3.59 cm to 8.82 cm. This result indicates how the sensor platform's movement, combined with more complex geometries observed by the cameras, affects the registration process. This trade-off between higher density and coverage leads to a decrease in precision. However, it can also mitigate the impact of distortions in a single frame by averaging them out with frames recorded immediately before or after, although this effect can occur in the opposite direction as well. As demonstrated later, voxelization of the data is effective for achieving precise localization. Manual measurements of the maximum deviation between the recorded points and the actual geometry indicate an upper limit of approximately 30 cm, which aligns with our preferred voxel size.

**2.2.2 Voxel Representation :** In addition to frame accumulation, voxelization plays a crucial role in optimizing the point cloud data for more efficient localization. Voxels can be considered as the 3D equivalent of pixels in a two dimensional (2D) space. Similar to the 2D case, voxels are placed on a 3D grid with uniform spacing in all dimensions. Analogous to a 2D square pixel representation, a voxel corresponds to a 3D cube (Chajdas, 2015). Voxel-based representation finds extensive use across various applications, such as finite-element simulations, object detection, classification, 3D reconstruction, localization, trajectory planning, and computer graphics rendering (Koketsu et al., 2004; Pantaleoni, 2011; Agus et al., 2010; Ma et al., 2021; Mao et al., 2021; Xie et al., 2018). The application of voxel representation can be attributed to its advantageous properties, including uniform resolution and a regular grid structure of independent cells, which simplifies complex computations and data handling (Chajdas, 2015).

In the context of stereo camera data, voxelization allows us to downsample the point cloud while maintaining key geometric features, thereby improving processing speed and reducing sensor noise. By applying a voxel grid filter, the system ensures that only essential points are retained, contributing to more robust and accurate localization results. In our approach, voxelization is applied after the data fusion process to the accumulated point clouds. A voxel grid filter is used to retain only the mean point within each voxel, ensuring that the most representative points are preserved, leading to improved localization accuracy. Additionally, the map itself is voxelized, ensuring consistency between the point clouds and the map during the localization process. The structured grid of voxels simplifies subsequent computations, and voxel-based models are widely utilized in autonomous systems for representing and navigating unexplored environments (Oleynikova et al., 2017).

### 2.3 Localization Approach

The localization approach is best described as a predictive point cloud odometry method, which leverages a constant velocity model to estimate a system's motion between consecutive point clouds. This approach offers a balance between computational efficiency and localization accuracy, making it suitable for real-time applications with limited computational resources.

The process begins with an initial position and corresponding transformation matrix, known from the starting point of the system. The constant velocity model is then used to predict the next transformation matrix based on the assumption that the system's translational and rotational velocities remain constant between consecutive time steps. This initial prediction provides an estimate of the system's next pose.

However, real-world motion is rarely perfectly linear or constant, leading to discrepancies between the predicted pose and the real movement. To correct these deviations, the system employs the Iterative Closest Point (ICP) algorithm, which aligns the newly captured point cloud with the pre-built reference map. ICP iteratively adjusts the transformation matrix by minimizing the difference between corresponding points in the current point cloud and the reference map. This refinement process ensures that the system maintains accurate localization, even when the initial prediction deviates from the actual path. This combination of predictive modeling and iterative refinement forms the core of the localization approach.

**2.3.1 Constant Velocity Model :** The constant velocity model assumes that the system's motion between consecutive time steps remains constant in both translation and rotation. This assumption simplifies the process of estimating the next pose, reducing the need for additional motion sensors such as IMUs or wheel encoders. This is particularly advantageous in low-cost systems where computational resources and hardware are limited.

Mathematically, the model uses the previous two poses to predict the next pose. The pose at time $t$ is represented by a homogeneous transformation matrix $\mathbf{T}_t$ , combining both rotation and translation:

$$\mathbf{T}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ 0 & 1 \end{bmatrix} \tag{1}$$

where $R_t$ and $t_t$ represent the robot's orientation and position, respectively. To predict the pose at this time step $T_{pred,t}$, we utilize the relative transformation observed between the previous two time steps:

$$\mathbf{T}_{\text{pred},t} = \begin{bmatrix} \mathbf{R}_{t-2}^{\top}\mathbf{R}_{t-1} & \mathbf{R}_{t-2}^{\top}(\mathbf{t}_{t-1} - \mathbf{t}_{t-2}) \\ \mathbf{0} & 1 \end{bmatrix} \tag{2}$$

This approach (Thrun et al., 2005), leverages the assumption that the system's velocity remains unchanged, providing a computationally efficient method to estimate the next pose. While this method is commonly used in applications such as visual odometry and SLAM, it requires refinement to correct for any deviations from the constant velocity assumption. In our system, this refinement is achieved through ICP, which iteratively aligns the predicted point cloud with the reference map.

**2.3.2 ICP Refinement :** For refinement, the predicted transformation matrix, derived from the constant velocity model, is used as the initial transformation matrix in the point-to-point ICP algorithm. This approach iteratively aligns the newly combined point cloud from the stereo cameras with the pre-built

map by minimizing the point-to-point distances. This alignment process corrects any deviations in the predicted transformation, thus enhancing localization accuracy.

In our system, we employ the point-to-point variant of the ICP algorithm. This variant minimizes the Euclidean distance between corresponding points in the two point clouds. While point-to-plane ICP is often more efficient for environments with planar surfaces (such as flat walls or floors), point-to-point ICP is better suited for environments with diverse and irregular structures, like the interior of a fair hall, where objects such as booths, banners, and varying structures are present.

The refinement process starts with the predicted transformation and iteratively updates it through ICP by minimizing the following error function:

$$\mathbf{T} = \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}} \|\mathbf{p} - \mathbf{T}\mathbf{q}\|^2 \tag{3}$$

Where:

- $\mathbf{T}$ is the transformation matrix (rotation and translation) being optimized,
- $\mathbf{p}$ are points from the reference map (LiDAR point cloud),
- $\mathbf{q}$ are points from the newly captured point cloud (stereo camera data),
- $\mathcal{K}$ is the correspondence set, representing pairs of points $(\mathbf{p}, \mathbf{q})$ between the reference map and the new point cloud.

The approach, originally formulated by Besl and McKay (1992), provides a robust method to iteratively reduce the localization error. Once the ICP refinement is complete, the resulting transformation matrix is used to predict the next set of frames, ensuring continuous and precise localization.

## 3. Experimental Results

To demonstrate the effectiveness of the proposed multi-stereo camera system for indoor localization, data was collected over a two-minute period, during which the system continuously collected and processed point cloud data.

Initially, point clouds from the three stereo cameras, facing different directions, were merged to enhance environmental coverage and density. This spatial fusion ensured that sufficient structural information was captured from different perspectives, even when certain cameras had limited or no visible features due to the short range of stereo vision. To further improve robustness and detail, temporal fusion was performed by aggregating 30 consecutive frames captured over one-second intervals, resulting in dense point clouds suitable for accurate localization tasks.

In the context of improving both the accuracy and efficiency of the localization process, the point clouds generated by a stereo camera were downsampled to a resolution of 30 cm. This downsampling was crucial for two main reasons: 1) it simplified the point cloud by reducing the density of points, which can have the effect of smoothing out small-scale variations and inconsistencies in the data, and 2) it significantly accelerated the localization process by decreasing the number of points that needed to be processed.

For the localization process, each temporally accumulated point cloud was aligned with a pre-built map using our predictive point cloud odometry method. The method employed a constant velocity model to generate an initial pose estimate, which was then refined through the point-to-point ICP algorithm. This iterative refinement corrected deviations and ensured accurate alignment between the measured data and the map. Following each localization step, the refined pose was used to predict the position of the subsequent merged point cloud, facilitating continuous localization throughout the data collection period.

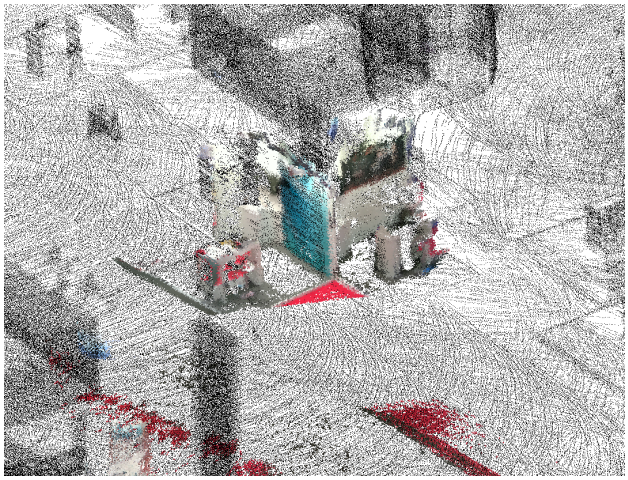
To generate the ground truth trajectory, we initially used the ICP algorithm to align each frame with the pre-built map. However, due to the presence of noisy or distant objects not well-captured by the stereo cameras, the ICP algorithm often misaligned frames by attempting to match all points globally, including unreliable features, instead of focusing on closer, more reliable structures. To improve accuracy, we manually refined the transformations of each frame, emphasizing alignment with nearby features. While this approach introduces some subjectivity, it allowed us to obtain a more reliable trajectory for our evaluation given our current constraints.

The outcomes of this approach are illustrated in Figure 5, which shows the alignment results at two different time instances during the data collection. The colored points represent the merged point clouds from the stereo cameras, while the gray points depict the pre-built map. These visualizations demonstrate the effectiveness of the point cloud fusion and localization process, highlighting the accurate alignment achieved through our approach. The dense and comprehensive nature of the merged point clouds contributed to consistent alignment precision, validating the system's capability to achieve reliable indoor localization in real-world scenarios. The trajectory plot in Figure 6 illustrates the strong alignment between the estimated trajectory (red line) and the ground truth (blue line), highlighting the consistent and accurate localization achieved throughout the process.

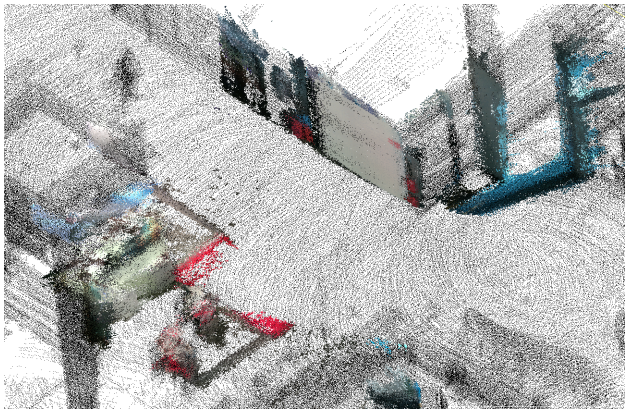
### 3.1 Translation and Rotation Error Analysis

To evaluate the accuracy of the localization approach, both translation and rotation errors between the estimated positions and the ground truth positions have been computed. The translation error has been calculated as the Euclidean distance between the estimated and ground truth positions for each timestamp. This calculation has been performed by comparing the translation vectors extracted from the homogeneous transformation matrices representing the pose at each timestamp. The histogram of translation errors (Figure 7 - left) reveals that the majority of translation errors are concentrated around 13 cm, with a standard deviation of 10 cm. This clustering indicates consistent small deviations from the ground truth, while occasional larger errors beyond 40 cm are likely due to challenges in feature-less areas or remaining sensor noise. However, these instances are relatively rare, suggesting that the system is robust for most indoor environments.

For rotation errors, the angular deviation between the estimated and ground truth rotation matrices has been computed. The histogram of rotation errors (Figure 7 - right) shows that the majority of rotation errors are clustered around 1.01 degrees, with a standard deviation of 1.58 degrees, indicating strong rotational accuracy throughout the localization process. Occasional outliers in both translation and rotation errors can be attributed to factors such as limited distinguishable features, sensor noise,



(a)



(b)

Figure 5. Localization results showing the alignment of the merged point cloud with the pre-built map in two different locations.

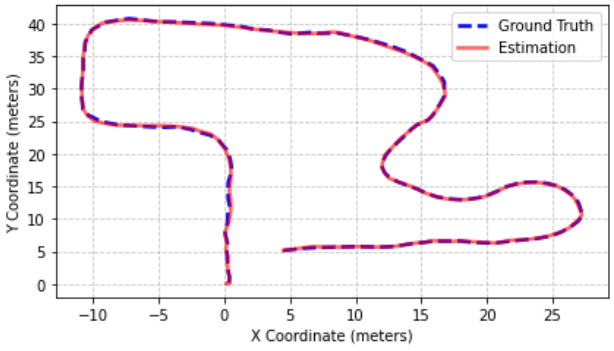



Figure 6. Comparison of the estimated trajectory (red) and the ground truth (blue)

and the accumulation of minor deviations over time. Despite these, the system remains robust, with the majority of errors well-controlled due to temporal fusion, and ICP refinement. The close alignment of the estimated and ground truth trajectories (Figure 6) confirms the system's reliability and accuracy in large indoor spaces.

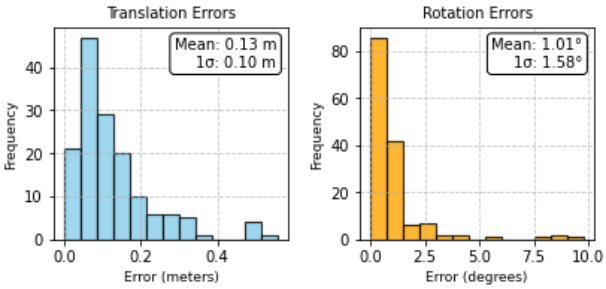



Figure 7. Histograms of translation (left) and rotation (right) errors derived from ground truth comparisons.

## 4. Conclusion

This study presents a novel multi-stereo camera system designed to enhance indoor localization by generating robust point clouds in large indoor environments. The proposed system successfully integrates point clouds from multiple stereo cameras, facing different directions, and applies temporal fusion techniques to overcome the limitations of stereo cameras, such as short range and limited coverage. The experimental results demonstrate the effectiveness of this approach, showing that the temporal data fusion significantly improves the density of point clouds, making low-cost stereo cameras a viable alternative to high-cost LiDAR systems for indoor localization.

The integration of the enhanced point clouds into a predictive point cloud odometry method further validated the system's ability to achieve accurate and continuous localization within a pre-built map. Future work could focus on enhancing the fusion process to be more dynamic and adaptive, allowing for the selection of an optimal number of frames based on the movement and environmental conditions. Additionally, efforts could be made to further improve sensor robustness and reduce computational overhead in real-time applications. This research highlights the potential of low-cost stereo cameras in high-precision indoor localization for applications in areas such as robotics, indoor mapping, and augmented reality systems.

## 5. Acknowledgements

We acknowledge valuable discussions with C. von Viebahn, J. Rohde, and P. O. Gottschewski-Meyer. This work wass financially supported by the German Federal Ministry of Digital and Transport (BMDV), project 5GAPS (grant no. 45FGU121).

## 6. References

Agus, M., Gobbetti, E., Guitián, J., Iglesias A., Marton, F., 2010. Split-Voxel: A Simple Discontinuity-Preserving Voxel Representation for Volume Rendering. IEEE/ EG Symposium on Volume Graphics. The Eurographics Association, 978-3-905674-23-1.http://dx.doi.org/10.2312/VG/VG10/021-028.

Besl, P.J. and McKay, N.D., 1992, April. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures* (Vol. 1611, pp. 586-606). Spie.

Caselitz, T., Steder, B., Ruhnke, M. and Burgard, W., 2016, October. Monocular camera localization in 3D LiDAR maps. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1926-1931). IEEE.

Chajdas, M. G., 2015. A voxel-based visualization pipeline for high-resolution geometry. Diss. Technische Universität München.

Han, D., Zou, Z., Wang, L. and Xu, C.Z., 2019, December. A robust stereo camera localization method with prior LiDAR map constrains. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (pp. 2001-2006). IEEE.

Kim, Y., Jeong, J. and Kim, A., 2018, October. Stereo camera localization in 3D LiDAR maps. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1-9). IEEE.

Koketsu, K., Fujiwara, H. and Ikegami, Y., 2004. Finite-element simulation of seismic ground motion with a voxel mesh. *Pure and Applied Geophysics*, *161*(11), pp.2183-2198.

Kuzminykh, A., Rohde, J., Gottschewski-Meyer, P. O. and Ahlers, V., 2023. Stereo vision and LiDAR based point cloud acquisition for creating digital twins in indoor applications. In *Proceedings of the 2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS 2023)*. (pp. 947–952). IEEE.

Ma, J., Tong, J., Wang, S., Zhao, W., Duan, Z. and Nguyen, C., 2021. Voxelized 3d feature aggregation for multiview detection. *arXiv preprint arXiv:2112.03471*.

Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H. and Xu, C., 2021. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3164-3173).

Mortazavi, F.S., Shkedova, O., Feuerhake, U., Brenner, C. and Sester, M., 2024. Voxel-based point cloud localization for smart spaces management. *arXiv preprint arXiv:2406.15110*.

Oleynikova, H., Taylor, Z., Fehr, M., Siegwart, R. and Nieto, J., 2017, September. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1366-1373). IEEE.

Pantaleoni, J., 2011. VoxelPipe: a programmable pipeline for 3D voxelization. In Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics (HPG '11). Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/2018323.2018339.

Ruchti, P., Steder, B., Ruhnke, M. and Burgard, W., 2015, May. Localization on OpenStreetMap data using a 3D laser scanner. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 5260-5265). IEEE.

Segal, A., Haehnel, D. and Thrun, S., 2009, June. Generalized-ICP. In *Robotics: science and systems* (Vol. 2, No. 4, p. 435).

Suzuki, T., Kitamura, M., Amano, Y. and Hashizume, T., 2010, October. 6-DOF localization for a mobile robot using outdoor 3D voxel maps. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 5737-5743). IEEE.

Thrun, S., Burgard, W. and Fox, D., 2005. Probabilistic Robotics. Cambridge.

Xie, H., Yao, H., Sun, X., Zhou, S. and Tong, X., 2018, August. Weighted Voxel: a novel voxel representation for 3D reconstruction. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service* (pp. 1-4).

Xu, Y., John, V., Mita, S., Tehrani, H., Ishimaru, K. and Nishino, S., 2017, June. 3D point cloud map based vehicle localization using stereo camera. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 487-492). IEEE.