10.2195/lj_proc_krishnamurthy_en_202410_01

Analysing visual-inertial odometry algorithms for the localization of industrial autonomous mobile robots in intralogistics and manufacturing

Analyse visueller Odometrie-Algorithmen für die Lokalisierung von autonomen mobilen Industrierobotern in der Intralogistik und Fertigung

Aishwarya Krishnamurthy[2]
Asan Adamanov[1]
Adithya Kumar Chinnakkonda Ravi[2]
Hendrik Rose[1]
Philipp Braun[1]
David Küstner[2]

[1]*Institute of Technical Logistics, Hamburg University of Technology, Hamburg*

[2]*Synergeticon GmbH, Hamburg*

The use of Autonomous Mobile Robots (AMR) plays a significant role in the automation of intralogistics processes. For safe operation and navigation, high localization accuracy is required. Common AMR systems rely on cost-intensive sensors such as LIDAR scanners. To enable widespread use of AMRs the industry alternative solutions are required. This study explores stereo camera-based visual SLAM as a cost-effective alternative to conventional 3D LIDAR-based localization solutions for an industrial robot application. Using Stereolabs ZED 2I and Intel RealSense D455 cameras with ORB-SLAM3 and OpenVINS algorithms, we evaluated Mean Absolute Pose Error (APE) and Root Mean Square Pose Error (RPE). The highest accuracy was achieved with the ZED 2I with OpenVINS with an APE of 0.17 m and an RPE of 0.02 m while the use of a RealSense D455 showed an APE of 0.33 m with an RPE of 0.02 m.

[V-SLAM, ORB-SLAM3, VIO, AGVs, localization]

Der Einsatz von Autonomen Mobilen Robotern ist ein wichtiger Teil der Automatisierung von intralogistischen Prozessen. Dabei wird stets eine hohe Lokalisierungsgenauigkeit vorausgesetzt. Gängige AMR-Systeme basieren auf kostenintensiven Sensoren wie LIDAR-Scannern. Um einen breiten Einsatz von AMR in der Industrie zu ermöglichen, werden alternative Lösungen benötigt. In dieser Studie wird stereokamerabasiertes visuelles SLAM als kostengünstige Alternative zu herkömmlichen 3D-LIDAR-basierten Lokalisierungslösungen für eine industrielle Anwendung untersucht. Unter Verwendung von Stereolabs ZED 2I und Intel RealSense D455 Kameras mit ORB-SLAM3 und OpenVINS Algorithmen wurde der mittlere absolute Posenfehler und der mittleren quadratischer Posenfehler bewertet. Der ZED 2I mit OpenVINS erreichte einen APE von 0,17 m und einen RPE von 0,02 m, während ORB-SLAM3 mit RealSense D455 einen APE von 0,33 m und einen RPE von 0,02 m erreichte.

[V-SLAM, ORB-SLAM3, VIO, FTS, Lokalisierung]

1 INTRODUCTION

The increasing use of Autonomous Mobile Robots (AMRs) transporting goods or inspecting industrial environments underscores the critical need for precise and cost-effective localization solutions. While traditional 3D LIDAR-based methods offer high accuracy, their substantial cost and power consumption pose significant challenges, particularly for power-constrained industrial AMRs. Meanwhile, affordable sensors like 2D lasers (e.g., RPLIDAR A2M12) are commonly used due to cost constraints [1], and they often fall short of delivering the necessary accuracy for industrial applications. Even when supplemented with wheel encoder or IMU data, these systems suffer from issues like error accumulation and drift [2], challenging precise localization. Therefore, enhancing localization performance, potentially through the integration of camera-based visual odometry algorithms, is crucial for ensuring that AMRs can navigate and operate reliably in complex industrial settings.

Stereo camera-based visual SLAM (V-SLAM) uses two cameras to capture images from slightly different perspectives [3]. The cameras capture rich visual information, coupled with their ability to provide depth perception. This results in the ability to estimate the distance to objects which is critical for accurate localization, especially in dynamic industrial environments [3]. Therefore, stereo cameras present an opportunity to enhance localization accuracy and robustness. This research is motivated by the challenges with indoor localization faced in deploying autonomous mobile robots (AMRs), such as the robotic research platform at Synergeticon GmbH, shown in Figure 1, which is designed for autonomous environmental scanning and 3D data collection.

The accuracy limitations of affordable sensor-based localization methods (e.g. 2D laser, wheel encoder, IMU), particularly in dynamic industrial environments, highlight the need for a more robust and adaptive solution. Research has shown that the successful implementation of stereo camera-based visual SLAM could significantly enhance the autonomy, reliability, and efficiency of AMRs like our robot [4]. However, the selection of the optimal visual SLAM algorithm and stereo camera system for industrial AMR localization remains an open question due to the significant variation in the type of cameras and their image processing ways, which directly affects V-SLAM performance [4]. Common image-capturing methods, such as Global Shutter, Rolling Shutter, Time-of-Flight (ToF), and High Dynamic Range (HDR), along with factors like frame rate, resolution, field of view, and the integration of additional sensors (e.g., IMU), can significantly influence the performance of selected V-SLAM algorithms [5]. This research aims to address those challenges by conducting a comprehensive evaluation of different V-SLAM algorithms and stereo camera systems to determine how these factors impact performance in industrial AMR robots.

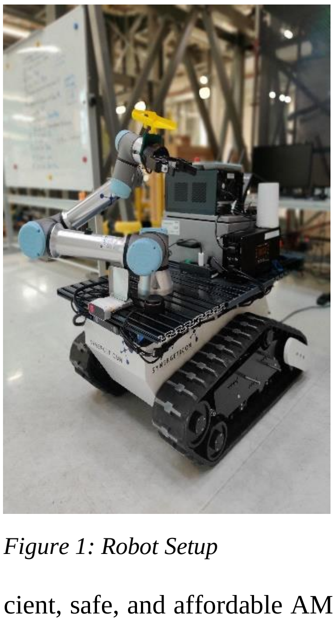

Figure 1: Robot Setup

The central research question is: How can stereo camera-based V-SLAM be optimized to serve as a viable alternative to 3D LIDAR for accurate and cost-effective localization in industrial AMRs? The study will focus on identifying the most suitable algorithm-camera combination that delivers the highest localization accuracy to integrate into the figure 1 Robotic research platform setup. The findings will contribute to the development of more efficient, safe, and affordable AMR deployment strategies in intralogistics and manufacturing. The paper is organized as follows: after this introduction, we begin with a review of the state of the art. Next, we present the system overview, where the selected algorithm and its key implementation factors are discussed. Afterwards, we describe the validation methodology. Finally, we discuss the results and their implications for industrial applications.

2 State of the Art

The localization of AMR using SLAM is implemented in various approaches using different sensors and algorithms (see figure 2).

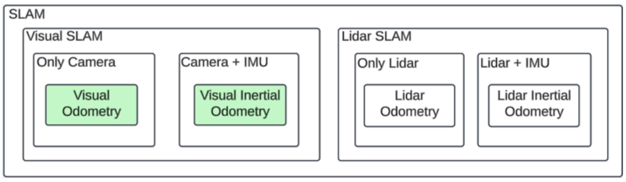

Figure 2. V-SLAM vs LIDAR SLAM

The SLAM approach can be primarily divided into two techniques based on the input device used: Visual SLAM, which relies on cameras, and LIDAR SLAM, which utilizes laser devices. This work focuses on camera-based systems, specifically on Visual SLAM, which can be further categorized into Visual Odometry (VO) and Visual Inertial Odometry (VIO). VO relies solely on camera input for motion estimation, making it simpler but potentially less accurate in complex environments. In contrast, VIO enhances accuracy by integrating camera data with an IMU, offering improved precision, especially in dynamic or challenging conditions.

Visual Odometry is defined as the process of estimating the robot’s motion (translation and rotation with respect to a reference frame) by observing a sequence of images of its environment [6]. It serves as the foundation for visual SLAM, providing incremental motion estimates that enable the robot to track its position and build a map of its surroundings. The accuracy and robustness of VO directly impact the overall performance of the visual SLAM system. VO methods can be divided into monocular [8] and stereo-camera methods [9]. These methods are further divided into feature matching (matching features over a number of frames) [10] and feature tracking [11] (matching features in adjacent frames). [12] proposed methods for obtaining camera motion from visual input in both monocular and stereo systems. [13] proposed a stereo VO system for outdoor navigation in which the sparse flow obtained by feature matching was separated into a flow based on close features and a flow based on distant features. The rationale for the separation is that small changes in camera translations do not visibly influence points that are far away.

While VO only estimates the ego-motion of an agent using images, V-SLAM is a process in which a robot is required to localize itself in an unknown environment and build a map of this environment at the same time without any prior information with the aid of external sensors (or a single sensor). The key contribution in the field of solving the problem of V-SLAM was made by the ORB-SLAM3 [14, 15] algorithm, which currently represents a state-of-the-art approach, it integrates visual and inertial data to enhance accuracy and robustness, especially in challenging visual conditions. ORB (Oriented FAST and Rotated BRIEF)-SLAM has three parallel processes: The first process is the construction of the local camera trajectory by matching the observed key points to the local map. The second process builds a local map and solves the local bundle adjustment problem, and the last process finds loop closures. Moreover, it can trigger a full optimization of the entire camera trajectory.

Other than VO methods Visual Inertial Odometry (VIO) approaches are slightly different from SLAM approaches. They focus on local consistency and aim to incrementally estimate the path of the camera/ robot pose after pose, and possibly perform local optimization. SLAM aims to obtain a globally consistent estimate of the camera/ robot trajectory and map and contain loop closure. Some VIO approaches also implement loop closure but are mostly used for accurate pose estimation.

Open Visual Inertial Navigation System (OpenVINS) [16] is a Multi-State Constraint Kalman Filter (MSCKF) based on a VIO estimator. It uses IMU in the propagation step of the filter and camera data in the update step. It also incorporates a loosely coupled loop closure thread based on VINS-Fusion. VINS-Mono [17] and VINS-Fusion [16] are graph-based VIO approaches. VINS-Fusion is an extension of VINS-Mono and supports multiple visual-inertial sensor types (mono camera and IMU, stereo cameras and IMU, even stereo cameras only). It also has support for global sensors like GPS and Barometer and has a global graph optimization module.

The author [18] evaluates various open-source Visual SLAM and Visual-Inertial Odometry algorithms. The authors' emphasis on the lack of a universal "out-of-the-box" solution and the necessity for algorithm tuning and data pre-processing further underscored the importance of conducting a focused comparison tailored to the project’s specific requirements. Similarly, [19, 20] emphasize that there is no single best open-source solution performed in all scenarios, reinforcing the need for careful selection and adaptation based on the task at hand.

The state of the art in Visual SLAM has seen significant advancements, yet several critical issues persist that hinder its widespread adoption in real-world applications. One of the most pressing concerns is the reproducibility problem, which has become a significant barrier to further innovation. As authors [18, 21, 22] highlight, researchers often struggle to replicate published results due to poorly documented code, insufficient examples, and the necessity for extensive algorithm tuning and data pre-processing. This issue not only limits the ability to build upon existing work but also undermines the credibility of research findings. Addressing this problem is essential for the continued progress of the field, requiring improved transparency, documentation, and accessibility of research outputs.

Another significant challenge is the absence of a universal V-SLAM solution that performs optimally across diverse environments and scenarios. Current algorithms are highly specialized, with performance heavily dependent on factors such as lighting conditions, sensor setup, and the specific characteristics of the environment. For instance, as authors [10] note, challenging lighting conditions, including variations in illumination, can severely degrade the accuracy of feature detection and tracking. Similarly, sensor limitations, such as the resolution, frame rate, and field of view of cameras, play a crucial role in determining the robustness of V-SLAM systems. Additionally, monocular V-SLAM suffers from inherent scale ambiguity due to the inability of single cameras to measure depth directly, often requiring additional sensors or techniques to resolve this issue [19].

Given that the choice of an algorithm depends heavily on the specific environment, the sensor setup, and desired performance trade-offs, this study embarks on a detailed investigation to identify the optimal Visual SLAM solution for our robotic platform, which is equipped with two cameras. The goal is to ensure robust and accurate indoor navigation capabilities. To achieve this, the study compares four different setups, using two distinct algorithms and sensor configurations, to determine the most effective approach for this particular application.

3 System Overview

This section gives a detailed overview of selected cameras, criteria for selecting the algorithm, validation setup, and obtaining a ground truth to validate the selected algorithms.

3.1 Selection of Camera for V-SLAM

The choice of the camera is critical for the performance of V-SLAM systems. Stereo cameras offer significant advantages over monocular setups, particularly for VIO. There are some key benefits of using stereo cameras. Firstly, scale ambiguity resolution wherein stereo cameras eliminate scale ambiguity by using the baseline distance between the cameras, leading to greater accuracy. Secondly, stereo cameras provide direct depth information, enhancing accuracy during large motions and rapid image changes, which is crucial for 3D mapping and navigation. Another primary importance is the Enhanced Feature Tracking. Stereo setups improve feature tracking, especially in texture-less environments, by triangulating positions in 3D space. These advantages make stereo cameras a superior choice for V-SLAM, offering enhanced robustness and accuracy. Therefore, to perform visual-based localization, the two most popular stereo cameras used for AI and robotic applications ZED2 [23] and Intel RealSense D455 [24] were chosen.

3.2 Algorithm Selection

Selecting a suitable localization algorithm requires meeting key industrial requirements. Accuracy and precision are a key requirement. For precise navigation, centimetre-level accuracy is to be ensured. Real-time performance enables instant decision-making and adaptive localization, especially for encoder drifts. Compatibility with different cameras, sensors, and environments is essential, which can be termed scalability and flexibility. Finally, the practicality and usability of the algorithms, such as the availability of documentation, examples on popular datasets, the convenience of the interface, the ability to change the parameters of algorithms, and the presence of Docker/ROS wrappers are of importance.

The paper [18] shows the comparison of various open-source V-SLAM algorithms based on practicality, different popular datasets, CPU, and memory usage. After evaluating open-source algorithms based on the above-mentioned criteria, including ROS1/2 compatibility and community support, two VIO, algorithms OpenVINS and ORB-SLAM3 were selected.

To sum up, OpenVINS and ORB-SLAM3 emerge as two strong contenders, particularly when considering key industrial requirements. OpenVINS shows stability in feature-sparse environments by effectively leveraging inertial data [16]. Its reliance on pre-integrated IMU values significantly reduces tracking loss, although it requires a comprehensive IMU initialization. This robustness ensures reliable performance supporting real-time decision-making crucial for adaptive localization and mitigating encoder drifts. On the other hand, ORB-SLAM3, known for its faster initialization and immediate pose estimation, demonstrates strong accuracy, particularly with global shutter cameras like the D455, although it is more sensitive to feature quality and IMU input frequency [18].

From a practical perspective, OpenVINS is user-friendly, with extensive documentation, an active supportive community, and ease of integration, which aligns with the need for scalability and flexibility across different sensors and environments. ORB-SLAM3, while more complex to implement due to its original ROS version dependency and the need for Docker, remains one of the most stable, popular, and industrially used SLAM algorithms available. However, both algorithms require careful parameter tuning for optimal performance. Both algorithms, despite their differences, provide options for real-world applications and are very adaptable to different camera sensors, with OpenVINS being more resilient and ORB-SLAM3 offering quick, precise localization under ideal conditions.

As this paper mainly focuses on the practical implementation of real-life data, based on the key requirements mentioned above and the comparison from the paper [18] OpenVINS and ORB-SLAM3 were the most suited algorithms to perform visual-inertial localization. Some of the other alternatives mentioned in the paper [18] include Basalt, Kimera, and OpenVSLAM.

3.3 Experimental Setup

To evaluate the performance of the two localization algorithms, a test environment was set up at the Institute of Technical Logistics at the Hamburg University of Technology, utilizing twelve motion capture cameras. Motion capture (Mocap) [25] is a technology that tracks the movement of an object or a fixed marker. Its high frequency and accuracy, with a precision of up to 0.05-0.11 mm, allow us to compare the proposed algorithm against the ground truth obtained from Mocap. A trolley served as the base platform, with a camera attached to one end of it. Four motion capture markers were positioned on the trolley as shown in figure 3.

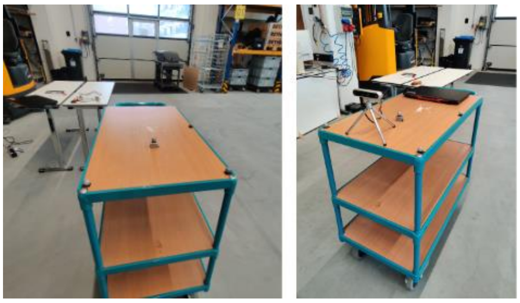

Figure 3. Trolley setup with motion capture markers(left), D455 on a tripod at one end of trolley(right)

Two cameras, namely ZED2 and D455, were used to test the algorithms. These cameras were mounted on one end of the trolley, as depicted in figure 3. The software Qualisys was utilized to define the rigid body system using motion capture markers. Additionally, this software provides a ROS wrapper that can be launched alongside the cameras to test the algorithms. This emulates as a moving base of an AGV, enabling the camera to traverse the hall and assess the performance of the algorithms.

3.4 Calibration

Initially, calibrating both cameras ZED2 and D455 is required, therefore, the open-source calibration toolbox called Kalibr [26] was used to obtain camera intrinsic properties. Afterward, it is essential to perform a visual-inertial calibration, i.e., a spatial and temporal calibration of an IMU w.r.t a camera system along with IMU intrinsic parameters. Before performing Visual Inertial Calibration, it is important to know how noisy the IMU unit of the camera is. To obtain the intrinsic parameters of the IMU (e.g., scales, axis misalignment), it must first undergo calibration, and the necessary corrections should then be applied to the raw measurements. Achieving accurate calibration is crucial, as the IMU errors related to the gyroscope and accelerometer should remain within acceptable limits.

3.5 Evaluation Metrics

The performance evaluation metrics are divided into quantitative and qualitative evaluation.

Quantitative evaluation includes Mean Absolute Pose Error (APE) and Root Mean Square Pose Error (RPE).

- *APE* is the absolute pose error and is a metric for investigating the global consistency of a SLAM trajectory. APE is based on the absolute relative pose between two poses *Pref_i*, *Pest_i* at timestamp i.
- *RPE* is the relative pose error and is a metric for investigating the local consistency of a SLAM trajectory. RPE compares the relative poses along the estimated and the reference trajectory. This is based on the delta pose difference between the estimated pose and ground truth.

Qualitative Evaluation:

The qualitative evaluation in this paper is performed based on three factors. The first factor is to evaluate based on *Scenario Handling*, which describes how each algorithm performs in environments with few distinguishable features. The second qualitative factor is *User Experience.* Feedback from developers who integrated OpenVINS and ORB-SLAM3 into their projects, highlighting the ease of integration and any challenges faced. The final qualitative factor evaluates based on *Visual Quality* which depicts a side-by-side comparison of the trajectories produced by both algorithms in a complex environment, with annotations discussing the differences in detail and accuracy.

4 Approach and Implementation

This section discusses the implementation of the two selected algorithms, OpenVINS and ORB-SLAM3 that are used for comparison to perform visual inertial localization in detail.

4.1 OpenVINS Implementation

OpenVINS is an open platform designed to help researchers and engineers quickly develop new capabilities for visual-inertial systems. It offers a robust foundation with out-of-the-box support for key features commonly needed in visual-inertial estimation. These features include an on-manifold sliding window Kalman filter, online calibration for both camera intrinsic and extrinsic parameters, camera-to-inertial sensor time offset calibration, SLAM landmarks with multiple representations, and consistent First-Estimates-Jacobian (FEJ) treatments [16]. In addition to these technical capabilities, OpenVINS places a strong emphasis on detailed documentation and derivations, making it a valuable resource for both development and research within the community. The OpenVINS consists of these key functionalities:

- *ov core* – Contains 2D image sparse visual feature tracking; linear and GaussNewton feature triangulation methods; visual-inertial simulator for arbitrary number of cameras and frequencies.
- *ov eval* – Contains trajectory alignment; plotting utilities for trajectory accuracy and consistency evaluation, Monte-Carlo evaluation of different accuracy metrics, and utility for recording ROS topics to file.
- *ov msckf* – Contains the extendable modular Extended Kalman Filter (EKF)-based sliding-window visual inertial estimator with a–manifold type system for flexible state representation.

The main feature that is provided by this algorithm estimate the current state of a camera-IMU pair. The capability of constructing sparse Jacobians reduces the computational complexity of adding new features and serves as an advantage of the OpenVINS algorithm. Instead of constructing a Jacobian for all state elements, the “sparse” Jacobian needs to only include the state elements that the measurement is a function of. This algorithm was used in this work due to its ease of use and computation, out-of-the-box testing, and ability to add and improvise on features that can be easily added. Figure 4 shows an example of how the algorithm performs with tracking features from a stereo-inertial camera.

One of the most important steps in implementing the algorithm is to fine-tune the parameters. With the camera placed on our robotic research platform and giving initialization time, three main tuning parameters need to be changed to get minimal drift and maximal accuracy of the robot’s positions. The three important fine-tuning parameters are firstly, *Camera calibration* parameters which can be accessed while obtaining offline calibration. These parameters are obtained from the calibration data of the camera using the Kalibr toolbox *Initialization window time* and *maximum features* can be updated based on the number of features to capture in an initial sliding window. The initialization window time was set to 1.5 and the maximum number of features to track during initialization was set to 50.

Finally, the feature tracking parameters, to update the number of tracking features which was set to 450 points as the number of points to track and type of descriptors used.. These three tuning parameters were crucial to perform accurate localization. The configuration file also includes other tuning parameters [30] that can further be optimized for preferred accuracy and application.

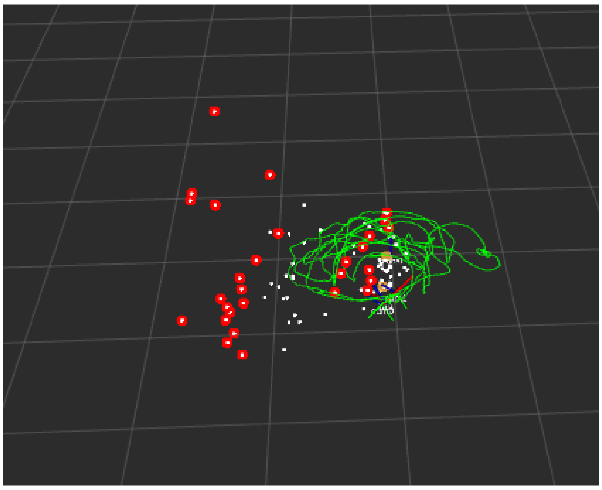

Figure 4 OpenVINS localization from stereo-inertial cameras.

OpenVINS was tested for both ZED2 and D455 on stereo-inertial mode. While stereo images for ZED2 are RGB images, for D455, the images in stereo-inertial mode are infrared and this was used to test OpenVINS.

4.2 ORB-SLAM3 Implementation

ORB-SLAM3 is a full multi-map and multi-session system able to work in pure visual or visual-inertial modes with monocular, stereo, or RGB-D sensors, using pinhole and fisheye camera models. ORB-SLAM3 provides a fast and accurate IMU initialization, technique, and an open-source SLAM library capable of monocular-inertial and stereo-inertial SLAM [27].

In the first part of Visual Inertial SLAM for ORB-SLAM3, the IMU measurements are taken between consecutive visual frames, i and i+1. The pre-integrated rotation, velocity, and position measurements are obtained for a whole measurement vector. Combining inertial and visual residual terms, visual-inertial SLAM can be posed as a keyframe-based minimization problem.

The second part is the IMU initialization. The goal of this step is to obtain good initial values for the inertial variables: body velocities, gravity direction, and IMU biases. The IMU initialization is considered as a Maximum-a-Posteriori (MAP) estimation problem which is split into three steps: Vision-only MAP estimation, inertial-only MAP estimation, and visual-inertial MAP estimation. Once there is a good estimation for inertial and visual parameters, a joint visual inertial optimization is performed further.

The third part involves tracking which solves a simplified visual-inertial optimization where only the states of the last two frames are optimized, while map points remain fixed. The visual-inertial system enters a visually lost state when less than 15-point maps are tracked and achieves robustness in two stages:

- *Short-term loss*: The current body state is estimated from IMU readings and map points are projected in the estimated camera pose and searched for matches within a large image window. The resulting matches are included in visual-inertial optimization. In most cases, this allows to recover visual tracking. Otherwise, after 5 seconds, we pass to the next stage.

- *Long-term loss*: A new visual-inertial map is initialized as explained above, and it becomes the active map. If the system gets lost within 15 seconds after IMU initialization, the map is discarded. This prevents to accumulation of inaccurate and meaningless maps.

It is important to note that this algorithm requires an IMU frequency of at least 100 Hz to perform localization [28]. The IMU frequency was set to 200 Hz for both D455 and ZED2. Upon increasing the IMU frequency, induced more noise which affected the tracking of the features and noise in the trajectory. Figure 5 shows an example of how the path is tracked and mapped on the GUI from ORB-SLAM3.

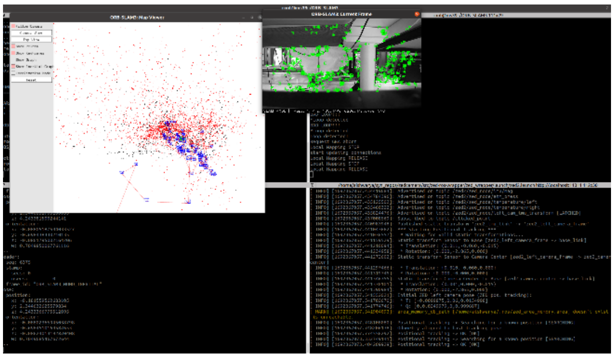

Figure 5 ORB-SLAM3 from stereo-inertial cameras.

On a reduced frequency, the camera system does not require an initial jerk to start the system to initialize the features and IMU. The algorithm automatically identifies once the command is launched. It also improved the quality of tracking by fine-tuning two important parameters from the configuration.

The first important parameter is the *ORBextractor.features*. It defines the number of features to be extracted. The more the number of features to track, the more accurate the system. It is also sensitive to correlate with IMU input. The number of features to track was set to 1200. The second important parameter is the *IMU frequency*. This plays a

10.2195/lj_proc_krishnamurthy_en_202410_01

very important role in stereo-inertial mode. Upon increasing the *IMU frequency* more noise is induced in the accuracy of localization. This was set to 200 Hz. These two fine-tuning parameters are very crucial for accurate positional localization for this study. Other fine-tuning parameters can be found in the configuration file, and they can be optimized for suitable cameras, applications, and required accuracy.

5 Results

This section presents the results of the evaluation of two algorithms that were used for localization, ORB-SLAM3 and OpenVINS. To perform the tests, two cameras were used for comparison, Intel RealSense D455 and ZED2.

5.1 Qualitative (Non-Numeric) Results

This section delves into the qualitative, non-numeric assessment of the algorithms, focusing on their scenario handling, user experience, and visual quality. The performance nuances of OpenVINS and ORB-SLAM3 are discussed in various operational contexts, highlighting their strengths and limitations.

5.1.1 Scenario Handling

OpenVINS's reliance on inertial data compensated for the lack of visual features, maintaining a very stable trajectory without relocalization. However, when using the D455 stereo camera, the algorithm's accuracy is reduced due to the infrared images produced by the camera, which are not supported by the algorithm. In feature-sparse settings, OpenVINS showed resilience by effectively utilizing inertial measurements. ORB-SLAM3 performed tracking without any loss with the D455, due to its infrared imaging capabilities which were accommodated in the algorithm. The global shutter of the D455 ensured clear and undistorted images, which enhanced feature matching and overall localization accuracy in ORB-SLAM3. However, with the ZED2 camera, performance is poor due to the rolling shutter nature of the camera, which causes motion blur and occasionally compromises image quality. Also, when the algorithm loses features, it quickly leads to tracking loss. This was very often observed in rolling shutter cameras [29] as mentioned in this paper.

5.1.2 User Experience

The OpenVINS algorithm is easy to integrate and test out of the box and the documentation is very detailed. The repository has a very active community answering issues and questions related to the algorithm, making it very user-friendly. While implementing ORB-SLAM3, the algorithm was complex due to its original implementation on ROS1 compared to the implementation of OpenVINS. This required docker and was not actively updated. The repository was not very active, although this algorithm is one of the most stable SLAM algorithms. Optimizing ORB-SLAM3 for diverse environmental conditions required careful parameter tuning.

5.1.3 Visual Quality

The visual quality of the trajectories from OpenVINS and ORB-SLAM3 can be observed in figure 6. OpenVINS is not a SLAM algorithm, it does not produce a map upon localization. ORB-SLAM3 is a SLAM algorithm, which means it can also produce a map and focus on loop closure. Since this paper focuses only on VIO, the visual quality is restricted and compared between trajectories obtained from localization. With IMU integrated along with visual input, the tracking of the trajectory is smooth in both OpenVINS and ORB-SLAM3 although tracking loss and relocalization were experienced higher on ORB-SLAM3.

5.2 Quantitative Results

Table 1 shows an overview of the overall performance of ORB-SLAM3 and OpenVINS with D455 and ZED2 cameras. Table 1 (c,f) shows ZED2 with the OpenVINS algorithm performing with a Maximum APE of 0.17 m and a Mean APE of 0.41 m. From the table, the positional accuracy of ZED2 with OpenVINS is 0.16 m less than D455 in ORB-SLAM3

The first tests were conducted using stereo-only mode using ORB-SLAM3. From Table 1. (a,b), it is shown that the maximum APE of D455 is 2.17 m less compared to the maximum APE error of ZED2. While the mean APE is almost 4.12 m for ZED2, it is much lesser for D455 which is around 0.70 m.

A similar evaluation was conducted between the D455 and ZED2 cameras to assess the performance of ORB-SLAM3 in stereo-inertial mode. During this test, ORB-SLAM3 with the D455 camera yielded successful results. However, results for the ZED2 camera could not be obtained as it failed to localize throughout the entire trajectory. This issue was also observed with the D455 initially, but upon reducing the frequency of the IMU, results were achieved. This behavior can be attributed to the fact that higher IMU frequencies induce more noise in the trajectory. The same approach did not yield positive results for the ZED2 camera. It continuously struggled with position relocalization, preventing the acquisition of meaningful data for analysis. Therefore, the following data only shows the performance of ORB-SLAM3 using D455 in -inertial mode. Table 1. (a,c) shows that in stereo-inertial mode, D455 performs 0.15 m less and better catering to the rotational peaks that were encountered in the stereo-only mode. The IMU input for localization proves to be an important aspect while performing rotational trajectories around the environment Table 1 (a,c) shows.that both APE and RPE are imporved in stereo-inertial mode compared

Metrics (m)	ORB-SLAM3				OpenVINS	
	Stereo mode		Stereo-Inertial mode		Stereo-Inertial mode	
	D455 (a)	ZED2 (b)	D455 (c)	ZED2 (d)	D455 (e)	ZED2 (f)
Max APE	0.48	2.65	0.33	-	0.70	**0.17**
Mean APE	0.70	4.12	0.53	-	1.24	**0.41**
Max RPE	0.03	0.04	0.02	-	0.01	**0.02**
Mean RPE	0.82	1.91	0.18	-	0.16	**0.53**

Table 1: Overall performance comparison for ORB-SLAM3 and OpenVINS in stereo, stereo-inertial modes for D455 and ZED2

to the stereo mode of D455.The overall dip in performance for ZED2 with ORB-SLAM3 in stereo see Table 1. (b) and stereo-inertial mode see Table 1. (d) is because of the rolling shutter property of the camera. This is explained in detail in the Discussion section.

Finally, the tests using the OpenVINS algorithm were conducted with ZED2 and D455. Open-VINS algorithm is a VIO algorithm which means that it will not work without an inertial input. To test the OpenVINS algorithm, it is required to provide an initial excitation in all axes of the camera to initialize the algorithm. Although loop closure is achieved by ZED2, D455 had a lot of sharp lags and jerks while following the trajectory and was not able to perform loop closure. Table 1. (f,e) shows that the mean APE of ZED2 is 0.83 m less compared to the mean APE of D455. While the maximum APE for D455 is 0.53 m bigger than for ZED2 which is around 0.17 m. Comparing the mean RPE of both camera systems, the mean RPE of D455 is 0.37 m less than that of ZED2. Although the mean RPE of D455 is lesser, it can be observed that the trajectory loop formed by D455 on OpenVINS is inside the trajectory followed by the motion capture leading to a negative offset in position throughout the loop and has a lot of sharp jerks and edges. This shows position values 0.37 m less than that given by the ground truth which leads to lower RPE. This is due to the stereo images from D455 being infrared and this affects the features that are used for tracking in OpenVINS, which causes the drift in the path. OpenVINS works with RGB images and D455 has a mono RGB camera on it. Having evaluated both ORB-SLAM3 and OpenVINS with the D455 and ZED2 camera systems individually, the next step involves comparing their respective performances to determine the superior algorithm-camera combination. While ORB-SLAM3 demonstrates robust performance with global shutter cameras

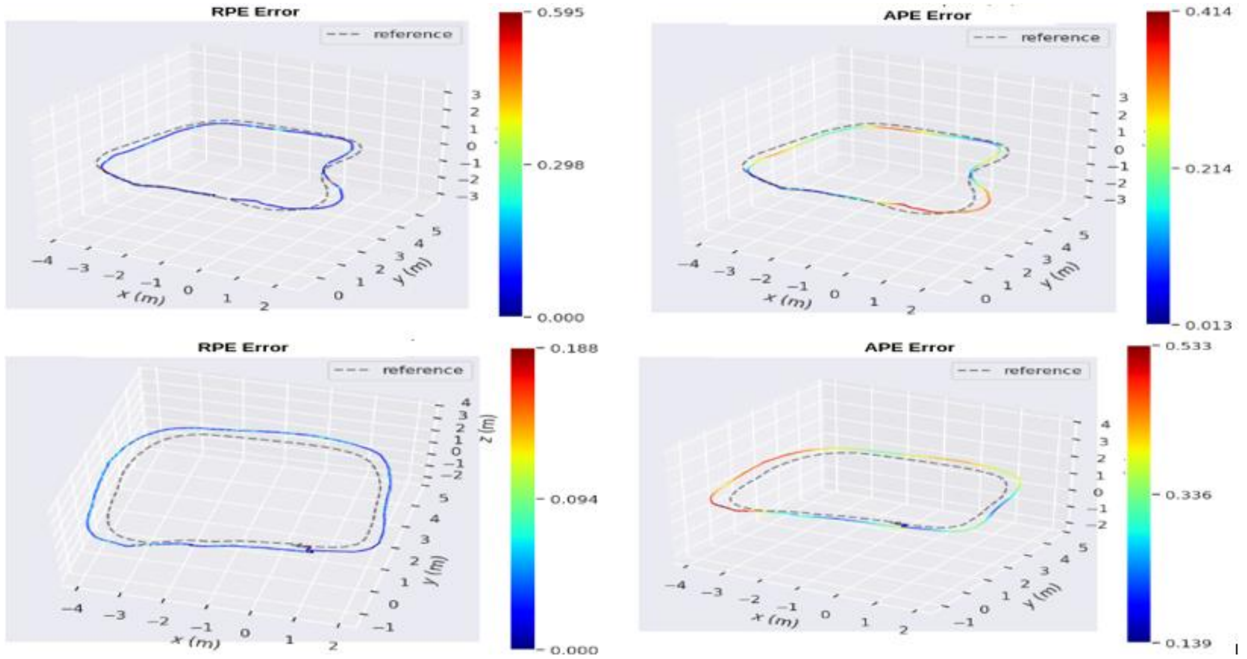

Figure 6 Stereo-Inertial performance with ORB-SLAM3 (D455-top) and OpenVINS (ZED2-bottom)

like the D455, OpenVINS showcases its strengths in odometry-based approaches, with IMU integration, and is particularly evident in its compatibility with stereo-inertial modes.

Figure 6 shows the performance of stereo-inertial mode in ORB-SLAM and OpenVINS respectively. Figure 6 shows that both ORB-SLAM3 and OpenVINS perform loop closure with respect to the reference motion capture. While both performances of ORB-SLAM3 on D455 and OpenVINS on ZED2 seem very promising, Table 1 (c,f) shows that the mean APE of ZED2 on OpenVINS is much lesser compared to D455 on ORB-SLAM3. While ORB-SLAM3 performs well in global shutter cameras, OpenVINS does not accommodate infrared images to perform localization. Since stereo-inertial mode in D455 provides infrared images, this causes the tediousness performance from D455 in OpenVINS.

Based on the results, comparing the performance of algorithms with D455 (ORB-SLAM3) from (a,c) and ZED2 (OpenVINS) from (b,f), OpenVINS was selected as the localization algorithm to implement on our research platform.

6 DISCUSSION

This section discusses the findings derived from the tests and analysis conducted on the algorithms to perform visual-inertial localization and gives a clear understanding of some of the common issues and problems faced while implementing the algorithms.

Critical Reflection

From the results of localization in the previous chapter, it is evident that ZED2 showcased the best performance using OpenVINS in stereo-inertial mode with a maximum APE of 0.17 m and a maximum RPE of 0.02 m. This superior performance can be attributed to high-resolution images from ZED2 and the ability of OpenVINS to extract features from high-resolution color images. While the Intel RealSense D455 demonstrated poor performance with OpenVINS because of the infrared nature of the stereo images in stereo-inertial mode, it performed promisingly with ORB-SLAM3 in both stereo and stereo-inertial modes. Table 1. highlights that D455 follows a smooth trajectory that adheres closely to the ground truth. The robustness of ORB-SLAM3 to different lighting conditions, combined with the scale and rotation invariance of ORB features, contributes to its efficacy, particularly with infrared images.

The performance discrepancy between ZED2 and D455 using ORB-SLAM3 can be primarily attributed to the nature of the cameras. ZED2, being a rolling shutter camera, suffers from motion blur in fast-moving environments, especially in the stereo-inertial mode where IMU input discrepancies accumulate quickly. Conversely, the D455, which uses a global shutter, captures the entire image simultaneously, preventing the distortions typically caused by rolling shutters in high-speed scenarios. ORB-SLAM3 also, as an algorithm is very sensitive to the camera being used and based on their calibration parameters. The contrary can be seen in OpenVINS wherein, ZED2 being a rolling shutter camera performs very well in OpenVINS due to high frame rates that can mitigate the rolling shutter distortions. Further, the integration of IMU at high frequency compensates for this distortion, leading to good performance [31].

Global Shutter vs Rolling Shutter

Global shutter cameras capture the entire image simultaneously, avoiding the distortions seen in rolling shutter cameras, which capture images row by row. This distinction is crucial in dynamic environments where fast-moving objects or the camera itself can introduce significant distortion. This difference underscores why ORB-SLAM performs better with the D455 compared to the ZED2, as ORB-SLAM relies heavily on accurate feature matching, which is compromised by rolling shutter distortions. Therefore, as shown in figure 7, the image captured with a rolling shutter, especially the rotary blades of the helicopter, appears deformed compared to the image captured using a camera with a global shutter. A similar effect will be observed when cameras are exposed to very high-speed events. This phenomenon can occur in two scenarios: either the camera is observing a high-speed object, or the camera itself is moving at a very high speed.

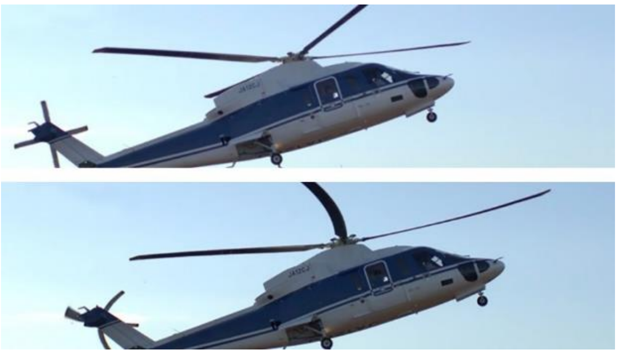

Figure 7 Global shutter(above) and Rolling shutter(below) [32]

In the context of visual localization for AGV, it can be observed that when the robot is moving, the environment changes dynamically and the camera perceives this change in the environment as it is in motion. This affects the speed of the scene that immediately changes due to the nature of the global shutter or rolling shutter of the camera in turn affecting the localization accuracy. Depending on the scene perceived by the camera in motion with AGV, the localization accuracy depends on the algorithm that is being used based on the images.

Stereo-Inertial Estimation

The results further indicate the superiority of stereo-inertial estimation over stereo estimation, particularly for differential robots like ours. Figure 6 shows that inertial measurements provide critical data about acceleration, angular velocity, and orientation, compensating for dynamic changes and maintaining robust performance across varying lighting conditions. It is important to perform accurate camera calibration in case of visual inertial localization since it relies solely on camera streams to extract features for localization which can be affected by the intrinsic, and extrinsic properties and the camera to IMU transformations. This capability is vital for avoiding the drifts and inaccuracies observed with vision-only based localization, ensuring reliable operation in diverse environments.

Limitations

Despite the promising results, several limitations were identified in the study. The first limitation is the camera type. The ZED2's reliance on a rolling shutter significantly hampers its performance in high-speed environments due to motion blur and the gradual accumulation of IMU input discrepancies. This limitation highlights the inherent challenges in using rolling shutter cameras for precise localization tasks in dynamic settings. Secondly, ORB-SLAM's sensitivity to rolling shutter distortions limits its effectiveness with rolling shutter cameras like the ZED2. Although at high frame rates and high IMU frequency rolling shutter distortions can be compensated, the reliance on accurate feature matching makes ORB-SLAM vulnerable to distortions that can degrade overall performance, emphasizing the need for algorithms that can better compensate for such distortions. Upon localizing while the robot is moving, this can also lead to motion blur. This remains a challenge during rapid rotational movements when relying solely on stereo vision. This blur can lead to incorrect pose estimates, underscoring the necessity of integrating inertial measurements for more accurate position estimation, which was seen in ORB-SLAM3. Although stereo-inertial estimation proved beneficial, its effectiveness can be influenced by the specific environmental conditions and the robot's operational context. Ensuring consistent performance across a wide range of scenarios requires further refinement and testing to address potential variability in results.

7 Conclusion and Outlook

This paper discusses the comparison and analysis of two different VIO methods that provide accurate tracking in industrial warehouses. ORB-SLAM3 and OpenVINS were compared using motion capture data in a warehouse and real-world environments. The results showed that OpenVINS, utilizing ZED2 stereo cameras, achieved superior localization accuracy with a mean absolute position error (APE) of 0.17 m and a mean relative position error (RPE) of 0.02 m, outperforming ORB-SLAM3 in the same conditions.

In conclusion, while the findings underscore the importance of stereo-inertial estimation for robust localization, particularly in differential robots, the limitations identified highlight the need for continued development and optimization of both hardware and algorithms to address the challenges posed by dynamic and diverse environments.

By adopting this visual-inertial localization onto the robot, the localization performance improved the accuracy of the robot’s real-time position and accounted for the drifts that were coming from the wheels of the robot. It showed no tracking loss due to active feature tracking on RGB images from the camera even in low lighting conditions. The result was very reliable for the robot to use the pose obtained from visual-inertial localization and to perform accurate motion planning and mapping as the next steps leading to autonomous navigation in industrial environments.

From the findings of this paper, the setup with ZED2 camera in combination with OpenVINS showed optimal performance and is recommended to use further as a VIO system for localization on the robot. Due to its positional accuracy as seen from the results in Table 5.1 and ease of use with clear documentation, active community support, and easy implementation, OpenVINS has an edge over ORB-SLAM3.

In the future, optimization of the absolute position error and relative position error of the localization algorithms can be achieved by integrating external sensors such as IMU and encoders.

8 Literature

[1] Lee, S., Lee, D., Choi, P. and Park, D., 2020. Accuracy–power controllable LIDAR sensor system with 3D object recognition for autonomous vehicle. Sensors, 20(19), p.5706.

[2] Shi, Y., Fang, L., Xue, Z. and Qi, Z., 2022. Research on Random Drift Model Identification and Error Compensation Method of MEMS Sensor Based on EEMD-GRNN. Sensors, 22(14), p.5225.

[3] Engel, J., Stückler, J. and Cremers, D., 2015, September. Large-scale direct SLAM with stereo cameras. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 1935-1942). IEEE.

[4] Yousif, K., Bab-Hadiashar, A. and Hoseinnezhad, R., 2015. An overview to visual odometry and visual SLAM: Applications to mobile robotics. Intelligent Industrial Systems, 1(4), pp.289-311.

[5] Taketomi, T., Uchiyama, H. and Ikeda, S., 2017. Visual SLAM algorithms: A survey from 2010 to 2016. IPSJ transactions on computer vision and applications, 9, pp.1-11.

[7] Bloesch, M., Czarnowski, J., Clark, R., Leutenegger, S. and Davison, A.J., 2018. Codeslam—learning a compact, optimisable representation for dense visual slam. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2560-2568).

[8] Khalid Yousif, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. “An overview to visual odometry and visual SLAM: Applications to mobile robotics”. In: Intelligent Industrial Systems 1.4 (2015), pp. 289–311 (cit. on pp. 11, 16).

[9] Jason Campbell et al. “A robust visual odometry and precipice detection system using consumer-grade monocular vision”. In: Proceedings of the 2005 IEEE International Conference on robotics and automation. IEEE. 2005, pp. 3421–3427 (cit. on p. 11).

[10] Larry Matthies and STEVENA Shafer. “Error modeling in stereo navigation”. In: IEEE Journal on Robotics and Automation 3.3 (1987), pp. 239–248 (cit. on p. 11).

[11] Ashit Talukder et al. “Real-time detection of moving objects in a dynamic scene from moving robotic vehicles”. In: Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453). Vol. 2. IEEE. 2003, pp. 1308–1313 (cit. on p. 11).

[12] Christian Dornhege and Alexander Kleiner. “Visual odometry for tracked vehicles”. In: (2006) (cit. on p. 11).

[13] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: International journal of computer vision 60 (2004), pp. 91–110 (cit. on p. 11).

[14] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. “ORB-SLAM: a versatile and accurate monocular SLAM system”. In: IEEE transactions on robotics 31.5 (2015), pp. 1147–1163 (cit. on pp. 6, 12, 24).

[15] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: 2011 International conference on computer vision. Ieee. 2011, pp. 2564–2571 (cit. on pp. 6, 12).

[16] Geneva, P., Eckenhoff, K., Lee, W., Yang, Y. and Huang, G., 2020, May. Openvins: A research platform for visual-inertial estimation. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4666-4672). IEEE.

[17] Qin, T., Li, P. and Shen, S., 2018. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE transactions on robotics, 34(4), pp.1004-1020.

[18] Sharafutdinov, D., Griguletskii, M., Kopanev, P., Kurenkov, M., Ferrer, G., Burkov, A., Gonnochenko, A. and Tsetserukou, D., 2023. Comparison of modern open-source visual SLAM approaches. Journal of Intelligent & Robotic Systems, 107(3), p.43.

[19] Servières, M., Renaudin, V., Dupuis, A. and Antigny, N., 2021. Visual and Visual-Inertial SLAM: State of the Art, Classification, and Experimental Benchmarking. Journal of Sensors, 2021(1), p.2054828

[20] Merzlyakov, A. and Macenski, S., 2021, September. A comparison of modern general-purpose visual SLAM approaches. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 9190-9197). IEEE.

[21] Kannan, K., Chakrabarty, A., Baculi, J.E., Kawamura, E., Holforty, W. and Ippolito, C.A., 2023. Comparison of visual and LIDAR SLAM algorithms using NASA Flight Test Data. In AIAA SCITECH 2023 Forum (p. 2679).

[22] Macario Barros, A., Michel, M., Moline, Y., Corre, G. and Carrel, F., 2022. A comprehensive survey of visual slam algorithms. Robotics, 11(1), p.24.

[23] Stereolabs, "ZED2 Documentation: API Reference, Tutorials, and Integration," Available: https://www.stereolabs.com/docs, accessed Aug. 23, 2024.

[24] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen and A. Bhowmik, "Intel(R) RealSense (TM) Stereoscopic Depth Cameras," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 1267-1276, doi: 10.1109/CVPRW.2017.167.

[25] Van der Kruk, E. and Reijne, M.M., 2018. Accuracy of human motion capture systems for sport applications; state-of-the-art review. European journal of sport science, 18(6), pp.806-819.

[26] Rehder, J., Nikolic, J., Schneider, T., Hinzmann, T. and Siegwart, R., 2016, May. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes. In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4304-4311). IEEE.

[27] Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M. and Tardós, J.D., 2021. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics, 37(6), pp.1874-1890.

[28] UZ-SLAMLab, "ORB-SLAM3," GitHub repository, https://github.com/UZ-SLAMLab/ORB_SLAM3, accessed Aug. 23, 2024.

[29] Norée Palm, C., 2023. Rolling shutter in feature-based Visual-SLAM: Robustness through rectification in a wearable and monocular context.

[30] RPNG, "Open-VINS," GitHub repository, https://github.com/rpng/open_vins, accessed Aug. 23, 2024.

[31] Fan, B., Dai, Y. and Li, H., 2022. Rolling shutter inversion: Bring rolling shutter images to high framerate global shutter video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), pp.6214-6230.

[32] J. Paul "Rolling Shutter vs Global Shutter: What's the difference?," PremiumBeat, Aug. 30, 2018. [Online]. Available: https://www.premiumbeat.com/blog/know-the-basics-of-global-shutter-vs-rolling-shutter/. Accessed: Aug. 23, 2024.

Krishnamurthy, Aishwarya, M.Sc., Junior Software Developer Synergeticon GmbH. She earned her master's degree in Mechatronics with a focus on Robotics and Intelligent Systems at Hamburg University of Technology in 2024. She focuses on developing a software stack of Manipulators and AGVs.

Adamanov, Asan, M.Sc., studied Theoretical Mechanical Engineering at Hamburg University of Technology. Since 2022 he has been working as Research Associate at the Institute of Technical Logistics. His main focus rely on building prototypes of robots and deploying them for industrial purposes.

Ravi, Adithya, M.Sc., studied Mechatronics at Hamburg University of Technology, and since 2022 he has been Robotics Engineer at Synergeticon GmbH.

Rose, Hendrik, M.Sc., studied Mechanical and Industrial Engineering at Hamburg University of Technology. Since 2021 he has been working as Research Associate and Chief Engineer (since 2023) at the Institute of Technical Logistics**.**

Braun, Philipp, M.Sc., studied Logistics and Industrial Engineering at Hamburg University of Technology. Since 2019 he has been working as Research Associate and Chief Engineer (since 2023) at the Institute of Technical Logistics.

Küstner, David, M.Sc., studied Industrial Engineering at RWTH Aachen. After his studies, he became co-Founder and Managing Director of Synergeticon GmbH, a company focused on developing AI-supported assistance systems, located at ZAL Future in Hamburg.