UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2025-098
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 10/10/2025 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *INFORMATIQUE*

by

Junlin SONG
Born in China

CALIBRATION FOR VISION-BASED MULTI-SENSOR SYSTEM

Dissertation defence committee

Dr Miguel Angel OLIVARES MENDEZ, dissertation supervisor
Professor, Université du Luxembourg

Dr Holger VOOS, Chairman
Professor, Université du Luxembourg

Dr Javier GONZALEZ-JIMENEZ
Professor, UNIVERSITY OF MALAGA

Dr Pedro SANCHEZ-CUEVAS
SENER / Sevilla / SPAIN

Dr Jose Luis SANCHEZ LOPEZ
Research scientist, Université du Luxembourg

Acknowledgements

Firstly, I would like to express my gratitude and appreciation to my supervisor, Miguel. He provided me with an amazing opportunity to conduct doctoral research at the University of Luxembourg. In the early stage of doctoral research, he encouraged me to explore various research directions that I was interested in. When my paper was rejected, he always gave me warm encouragement and actively supported my research. Recalling the PhD interview four years ago, I was asked why I wanted to pursue a PhD. My answer was that I wanted to publish excellent research in top-tier conferences and journals and become an expert. This wish is continuously being fulfilled or approached with Miguel's support.

Secondly, I would like to thank Pedro and Antoine. Pedro provided many valuable research suggestions and actively participated in the discussions of my research during the first year of my doctoral studies. I still remember he told me to 'protect the doctoral research' and 'focus on the work and avoid interference'. Antoine guided my research after Pedro left. I am grateful for his useful suggestions, such as clearly explaining motivation and experimental phenomena in writing. Like Miguel, he often encouraged me in the meeting, which increased my confidence when my work was rejected. Thanks to all SpaceR colleagues and friends that I met in Luxembourg. You accompanied me through this journey.

Lastly, I would like to thank my family for their love and companionship during the lonely but meaningful journey. They always support and encourage me. The love from my family is an endless source of strength for me on the road to achieve my goals.

Abstract

Calibration is an indispensable prerequisite for a wide range of multi-sensor fusion applications. In this thesis, the main focus is on vision-based multi-sensor system. Cameras have the favorable characteristics of low cost, small size, and low power consumption. Moreover, cameras can perceive rich semantic information in the environment. These appealing advantages have made cameras popular in robotics, AR/VR, and planetary exploration. Thus, visual localization becomes ubiquitous. In order to evaluate the accuracy of visual localization, an additional sensor with higher reference precision needs to be introduced. Meanwhile, to improve the accuracy and robustness of visual localization, cameras are usually tightly coupled with other complementary sensors. These demands require solving the problem of multi-sensor calibration, which is the core topic of this thesis.

The first part of this thesis considers the calibration problem in vision-based relative localization applications. The focus is on how to use a high-precision global pose sensor to evaluate visual localization accuracy. To achieve this goal, the spatial-temporal calibration parameters between the camera and the global pose sensor must be calibrated. Two novel calibration algorithms are proposed, including target-based and target-less methods. The principles of these two methods can be applied to any calibration task.

The second part of this thesis investigates the calibration problem in multi-sensor localization with the participation of GPS. More specifically, the observability issue is addressed for the GPS-VIO system. The existing analysis based on linear observability theory points out that the rotational extrinsic parameters between GPS and VIO is unobservable. However, experiments indicate that this is not true. In order to address the discrepancy between theory and experiment, a novel nonlinear observability analysis is proposed, highlighting the theoretical contribution of this research.

The third part of this thesis revisits the online extrinsic calibration for the VIO system. For the common-seen and fundamental pure translational straight line motion, an issue with respect to the existing observability conclusion is identified. Contrary to the existing conclusion, novel proof shows that this motion can lead to the unobservability of the rotational extrinsic parameter between IMU and camera (at least one degree of freedom). By correct-

ing the existing conclusion, this novel theoretical finding disseminates more precise principle to the research community and provides explainable calibration guideline for practitioners.

Lastly, this thesis advances the calibration efficiency for IMU-Camera system. Most existing offline target-based calibration algorithms adopt the continuous-time state representation based on the B-spline. Although these methods can accurately calibrate spatial-temporal parameters, they suffer from high computational costs. To address this limitation, an extremely efficient calibration algorithm that unleashes the power of discrete-time state representation is designed, which achieves up to 1000x speedup compared to the most popular calibration toolbox (Kalibr).

Overall, this thesis deepens the calibration research for vision-based multi-sensor systems. These investigations provide more solid foundations for the state estimation of multi-sensor systems, from the perspective of system development and theoretical support.

Index

List of Figures

List of Tables

Chapter 1

Introduction

Visual localization has gained great popularity in the last few decades. Compared to other exteroceptive sensors such as LiDAR, the camera has the advantages of being low cost, lightweight, and low power consumption. As a fundamental technology, visual localization has supported a wide range of intelligent applications such as AR/VR [1, 2, 3, 4], robotics [5, 6, 7, 8], and planetary exploration [9, 10, 11], as shown in Fig. 1.1. For AR/VR applications, real-time accurate visual localization for headsets or glasses is crucial for rendering virtual scenes to real-world environments. The mismatch between virtual and real caused by incorrect localization could have adverse impacts on user experience and interaction. For autonomous robotic and planetary exploration applications without human intervention, visual localization provides critical inputs for downstream modules, for example, high-level decision-making, planning, and low-level control.

Current visual localization applications can be divided into three categories. The first type is vision-based relative localization for two agents, which means that one agent uses a camera to localize another agent and track their relative position expressed in the camera frame. The second type is multi-sensor localization with the participation of GPS. The third type is multi-sensor localization without the participation of GPS. A research topic closely related to localization is calibration, which determines the spatial-temporal relationships between different sensors. Therefore, calibration is an essential prerequisite for multi-sensor fusion. This thesis focuses on the calibration problems that need to be solved for these three

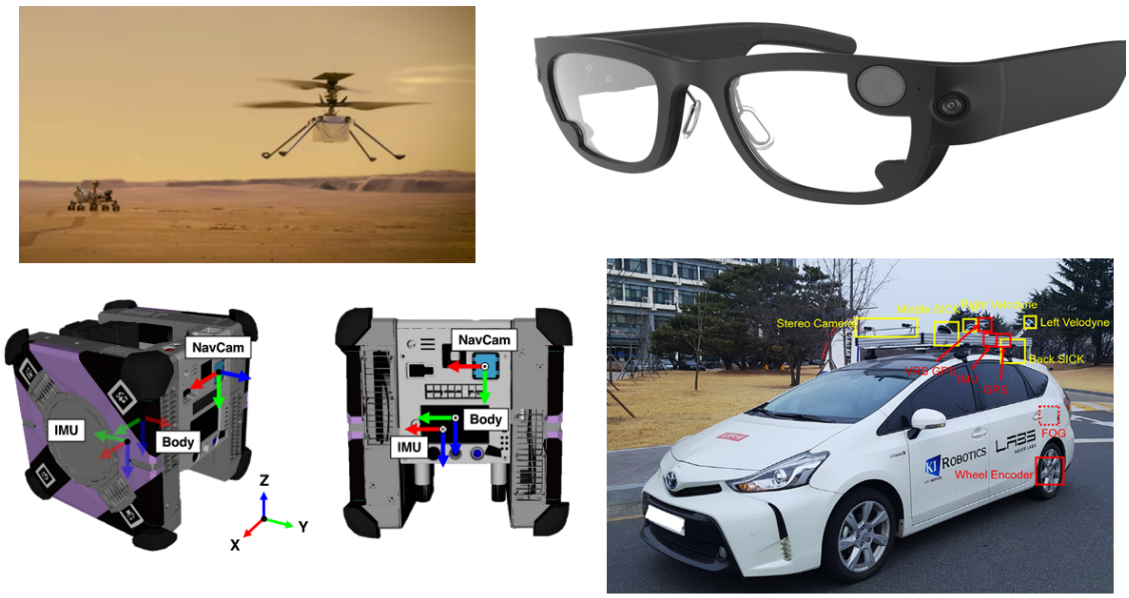


Figure 1.1: Top left: Ingenuity Mars Helicopter [11]. Top right: Wearable AR glass [12]. Bottom left: The Astrobee freeflying robots that operate inside the International Space Station (ISS) [13]. Bottom right: Autonomous driving vehicle [14].

types of visual localization application mentioned above.

In the following sections, I will clearly describe the necessity of calibration and different research objectives one by one.

1.1 Necessity of Calibration

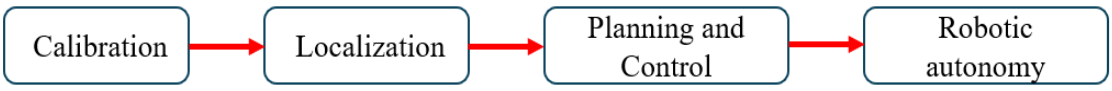


Figure 1.2: The necessity of calibration for autonomous robotic system.

The importance of calibration for robots is presented in Fig. 1.2. Calibration is the starting point for the development of autonomous robots. For example, Skydio R1 has 12 naviga-

tion cameras[1], as shown in Fig. 1.3. The spatial-temporal calibration parameters between 12 cameras and the built-in IMU must be known before bootstrapping multi-sensor fusion, which is the cornerstone for a wide range of high-level applications[2], which are depicted in Fig. 1.4. Another example is autonomous mapping with a drone [15]. Sensors used for autonomous 3D colored mapping include IMU, camera, and LiDAR, as shown in Fig. 1.5. Before operation, the spatial-temporal calibration parameters between these sensors must be obtained for path tracking and colored mapping (see Fig. 1.6). The path tracking module requires localization information depending on the sensor calibration. Colored mapping needs accurate extrinsics between camera and LiDAR so that color information from RGB camera could be aligned to point cloud from LiDAR.

Figure 1.3: Navigation cameras for Skydio R1.

Any multi-sensor localization system, such as Visual-Inertial Odometry (VIO), Lidar-Inertial Odometry (VIO), or Lidar-Visual-Inertial Odometry (LVIO), will malfunction without calibration parameters. The necessity of calibration for these localization systems is shown in Tab. 1.1. Therefore, the localization system cannot work without calibration. The failure of the localization system can lead to robot crashes. For example, the recent failure of the RESILIENCE lunar lander from ispace is likely due to the delay of the laser rangefinder[3]. If

[1] https://www.youtube.com/watch?v=gsfkGlSajHQ
[2] https://www.youtube.com/watch?v=Us6h9Q0-B2k
[3] https://ispace-inc.com/news-en/?p=7664

Figure 1.4: Autonomy engine of Skydio R1.

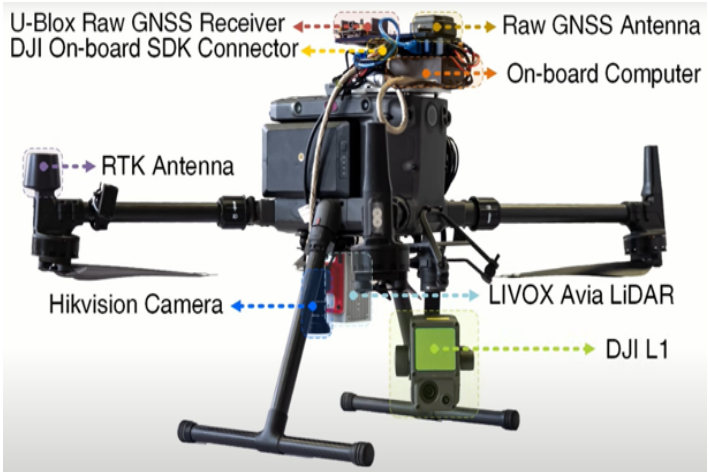


Figure 1.5: Sensors for aerial mapping.

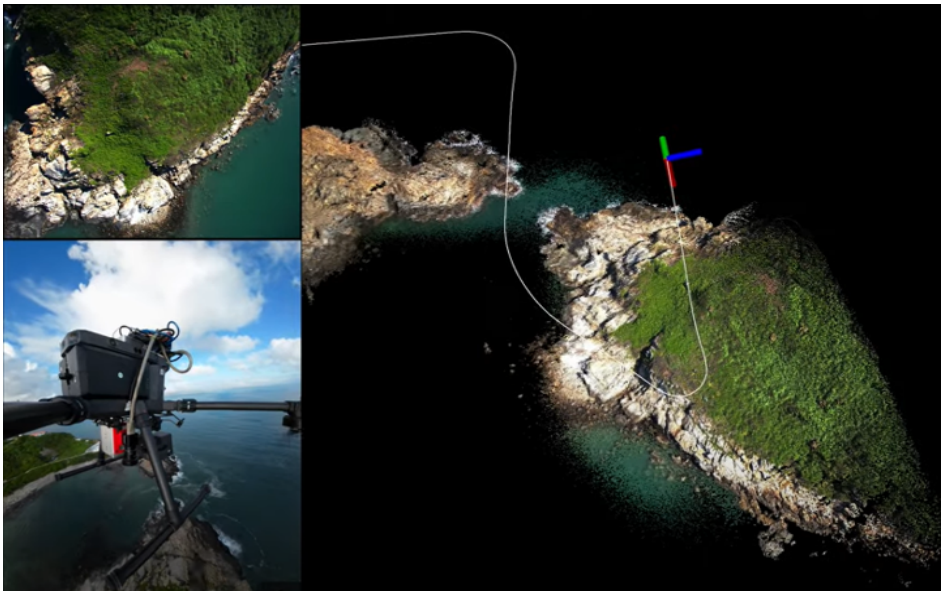


Figure 1.6: Autonomous 3D colored mapping with a drone.

Table 1.1: The necessary calibration parameters in different localization systems.

Localization system	Sensors	The necessary calibration parameters
VIO [16]	Camera and IMU	Spatial-temporal calibration parameters between IMU and Camera; intrinsics
LIO [17]	LiDAR and IMU	Spatial-temporal calibration parameters between IMU and LiDAR; intrinsics
LVIO [18]	LiDAR, Camera and IMU	Spatial-temporal calibration parameters between IMU, Camera and LiDAR; intrinsics

the measurement delay can be compensated through temporal calibration, then perhaps the accident would not have occurred. For multi-sensor systems, temporal calibration is critical, especially when the hardware-synchronization can not be guaranteed.

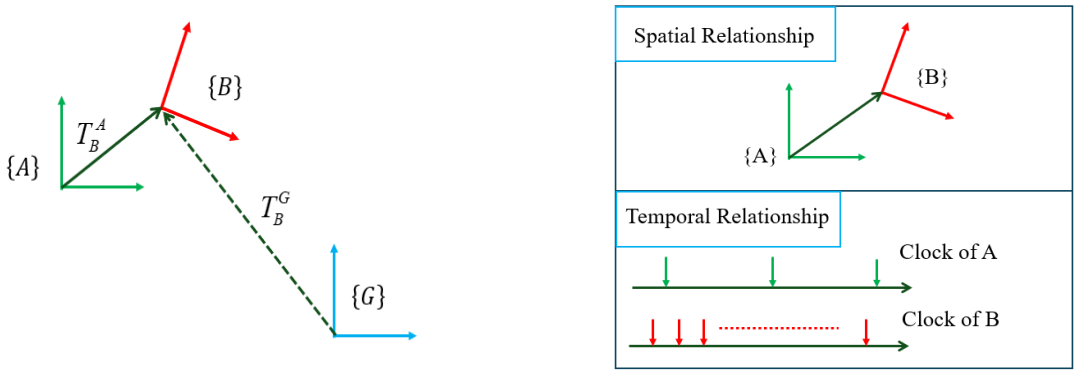
For a general localization system, there is typically a base sensor and other auxiliary sensors, as shown in Fig. 1.7. The base sensor can be an IMU, which provides instantaneous angular velocity and acceleration measurements. The auxiliary sensor can be a camera or a LiDAR, etc. The measurement model of the auxiliary sensor is abstracted as

$$y = h\left(x_b, x_c\right) + n_y \tag{1.1}$$

Where x_b represents the motion state of the base sensor, which may include position, attitude, and velocity at different timestamps. x_c is the set of calibration parameters for the base sensor and the specific auxiliary sensor.

$$x_c = \{T_B^A, t_B^A, i_B, i_A\} \tag{1.2}$$

Where T_B^A represents spatial calibration parameter, which is a transformation matrix from frame $\{B\}$ to $\{A\}$. t_B^A represents temporal calibration parameter, which is a time offset between the sensor $\{B\}$ and the sensor $\{A\}$. If the sensors are hardware-synchronized, the temporal calibration parameter can be ignored. However, the spatial calibration parameter is still required to ensure the measurements of the auxiliary sensor are valid. i_B and i_A are intrinsic parameters for the base sensor and the auxiliary sensor, respectively. An example description of the IMU intrinsic parameters[4] and camera intrinsic parameters[5] for the VIO system can be found in Open-VINS [16].



(a) (b)

Figure 1.7: (a) A general localization system with a base sensor (B) and an auxiliary sensor (A). $\{G\}$ represents the global coordinate frame. (b) The spatial-temporal relationship between B and A.

If the calibration parameters are missing, it is equivalent to the failure of the auxiliary sensor because Eq. (1.1) becomes invalid. If only the base sensor can work, then the localization system would diverge in a short period of time. After receiving incorrect localization

[4] https://docs.openvins.com/propagation_analytical.html
[5] https://docs.openvins.com/update-feat.html

information, the planning and control module would output incorrect commands to actuators, ultimately leading to malfunction and damage to the robot.

If the calibration parameters are inaccurate, it is equivalent to the permanent systematic error occurred in the measurement model of the auxiliary sensor, which can lead to inaccurate localization information. Although the safety of robots can be ensured to some extent through robust control and other redundant schemes, risks still exist in the system. In order to reduce risks from the source and provide optimal performance, it is better to recalibrate sensors or perform online calibration (see Fig. 1.8) to adapt to changes caused by the environment or its own structure. It should be noted that online self-calibration is not always effective as the observability of calibration parameters may be affected by different motion profiles [19, 20, 21]. Therefore, it is necessary to rigorously and carefully explore the observability of calibration parameters to provide theoretical support for online calibration.

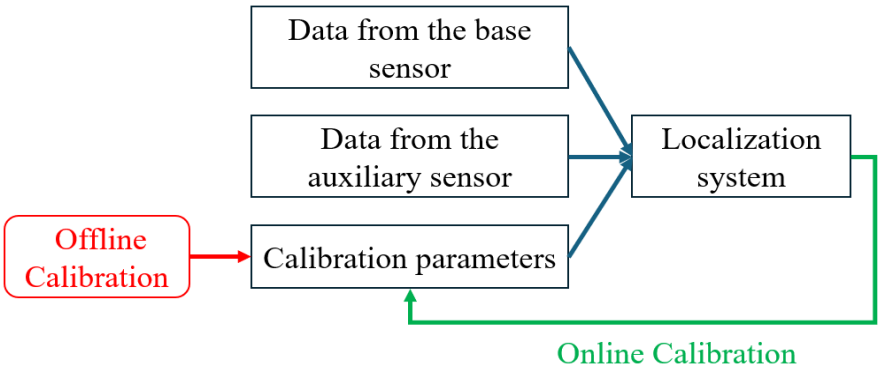


Figure 1.8: Offline and online calibration for localization system.

There are extensive calibration methods for different sensors in the literature. Taking IMU-Camera calibration as an example, a summary and comparison of IMU-Camera calibration methods are provided in Tab. 1.2

Table 1.2: Summary of IMU-Camera calibration research.

Method	Year	Continuous?	Target-based?	Offline?	Optimization-based?
[22]	2008	Discrete	Target-based	Online	Filter-based
[23]	2011	Discrete	Target-based	Online	Filter-based
[24]	2014	Discrete	Target-less	Online	Filter-based
[25, 26]	2013,2016	Continuous	Target-based	Offline	Optimization-based
[27]	2020	Continuous	Target-based	Offline	Optimization-based
[28]	2022	Continuous	Target-based	Offline	Optimization-based
[29]	2024	Discrete	Target-based	Offline	Optimization-based
[30]	2025	Continuous	Target-less	Offline	Optimization-based
[31]	2025	Discrete	Target-based	Offline	Optimization-based

1.2 Research Objective I

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor." 2024 International Conference on 3D Vision (3DV 2024).

The first research tackles the calibration question about how to evaluate the accuracy of vision-based relative localization with an auxiliary sensor. To evaluate the relative localization accuracy, an additional high-accuracy sensor is required, for example, a motion capture system. How to link the camera to the marker frame built by the motion capture system, which is treated as a global pose sensor? The key is the spatial-temporal calibration parameters of the camera and the global pose sensor. In this research, two novel solutions are proposed for the calibration of these two sensors.

1.3 Research Objective II

- Song, Junlin, Pedro J. Sanchez-Cuevas, Antoine Richard, Raj Thilak Rajan, and Miguel Olivares-Mendez. "GPS-VIO Fusion with Online Rotational Calibration." 2024 IEEE International Conference on Robotics and Automation (ICRA 2024).

The second research studies the calibration problem in multi-sensor localization with the participation of GPS, more specifically GPS-aided VIO. This research objective is targeted at outdoor GPS-aided applications, while Research Objective I is aimed at indoor applications with the support of a motion capture system. Firstly, the motivation for the GPS-VIO fusion is introduced. Secondly, an issue regarding the observability of the spatial transformation to couple both the GPS and the VIO reference frame is discussed. Lastly, a novel analysis is presented to address this issue, which highlights the theoretical contribution of this research for multi-sensor localization.

1.4 Research Objective III

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion." (Accepted by IROS 2025).

The third research revisits the online extrinsic calibration problem for the VIO system. Research Objective II aims at the calibration of the extrinsics between GPS and VIO. VIO is treated as a subsystem in Research Objective II. While this research objective is targeted at the calibration of VIO system itself. Firstly, the motivation for the online extrinsic calibration of VIO is introduced. Secondly, the issue of the existing observability conclusion is identified. Lastly, a novel proof is presented to address this issue, which highlights the theoretical contribution of this research to online extrinsic calibration.

1.5 Research Objective IV

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Spatial-Temporal Calibration." arXiv preprint arXiv:2509.12846 (2025).

The fourth research advances the offline target-based calibration efficiency for the IMU-Camera system. As shown in Research Objective III, some degrees of freedom are unobservable for online calibration, which means that offline target-based calibration is still necessary due to the potential risk of online calibration. Firstly, the motivation for the visual-inertial fusion is introduced. Secondly, the low efficiency of existing calibration methods is discussed. Lastly, a novel efficient solution is designed to fill this gap, which highlights the system contribution of this research for multi-sensor calibration.

1.6 Structure of the thesis

The calibration research in this thesis is summarized in TABLE 1.3.

- Chapter 1 introduces research objectives in this thesis and lists all research papers that have been accepted or are being submitted during the author's doctoral studies.
- Chapter 2 provides literature review for each research objective.
- Chapter 3 achieves Research Objective I described in Sec. 1.2.
- Chapter 4 achieves Research Objective II described in Sec. 1.3.
- Chapter 5 achieves Research Objective III described in Sec. 1.4.
- Chapter 6 achieves Research Objective IV described in Sec. 1.5.

Table 1.3: Summary of calibration research in this thesis.

Chapter	Sensors	Target-based?	Offline?	Optimization-based?
3	Camera and Global pose sensor	Target-based and Target-less	Offline and Online	Optimization-based and Filter-based
4	IMU, Camera and GPS	Target-less	Online	Filter-based
5	IMU, Camera and Global pose sensor	Target-less	Online	Filter-based
6	IMU and Camera	Target-based	Offline	Optimization-based

- Chapter 7 summarizes this thesis and provides an overview of the future research directions.

1.7 Publication and competition

Here are all the research papers that have been accepted or are submitting during the author's doctoral studies. The corresponding research paper for each chapter will be indicated at the beginning of each chapter.

Moreover, a SLAM competition is listed here. In the competition, my submission achieved good performance compared to other strong academic and industrial teams from around the world.

1.7.1 Research out in this thesis

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Spatial-Temporal Calibration." arXiv preprint arXiv:2509.12846 (2025).
- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion." (Accepted by IROS 2025).

- Song, Junlin, Pedro J. Sanchez-Cuevas, Antoine Richard, Raj Thilak Rajan, and Miguel Olivares-Mendez. "GPS-VIO Fusion with Online Rotational Calibration." 2024 IEEE International Conference on Robotics and Automation (ICRA 2024).

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor." 2024 International Conference on 3D Vision (3DV 2024).

1.7.2 Other research out not included in this thesis

- Song, Junlin and Miguel Olivares-Mendez. "Structureless VIO." (Accepted by the SLAM Workshop at RSS 2025).

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Improving Monocular Visual-Inertial Initialization with Structureless Visual-Inertial Bundle Adjustment." 2025 IEEE International Conference on Robotics and Automation (ICRA 2025).

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "An Accurate Filter-based Visual Inertial External Force Estimator via Instantaneous Accelerometer Update." Accepted by the 40th Anniversary of the IEEE Conference on Robotics and Automation (ICRA@40).

- Song, Junlin, Pedro J. Sanchez-Cuevas, Antoine Richard, and Miguel Olivares-Mendez. "GPS-aided Visual Wheel Odometry." 2023 IEEE International Conference on Intelligent Transportation Systems (ITSC 2023).

- Song, Junlin, Pedro J. Sanchez-Cuevas, and Miguel Olivares-Mendez. "Towards Online System Identification: Benchmark of Model Identification Techniques for Variable Dynamics UAV Applications." 2022 International Conference on Unmanned Aircraft Systems (ICUAS 2022).

1.7.3 ICRA 2022 HILTI SLAM CHALLENGE

- Designed a tightly-coupled Lidar-Visual-Inertial Odometry (LVIO) based on Multi-State Constraint Kalman Filter (MSCKF). Supported multiple cameras. Accelerated Lidar measurement update with QR decomposition.
- Ranked 3rd from 100 teams in this competition (see Fig. 1.9). Check my submission (SpaceR-Junlin) in Hilti SLAM challenge 2022[6].

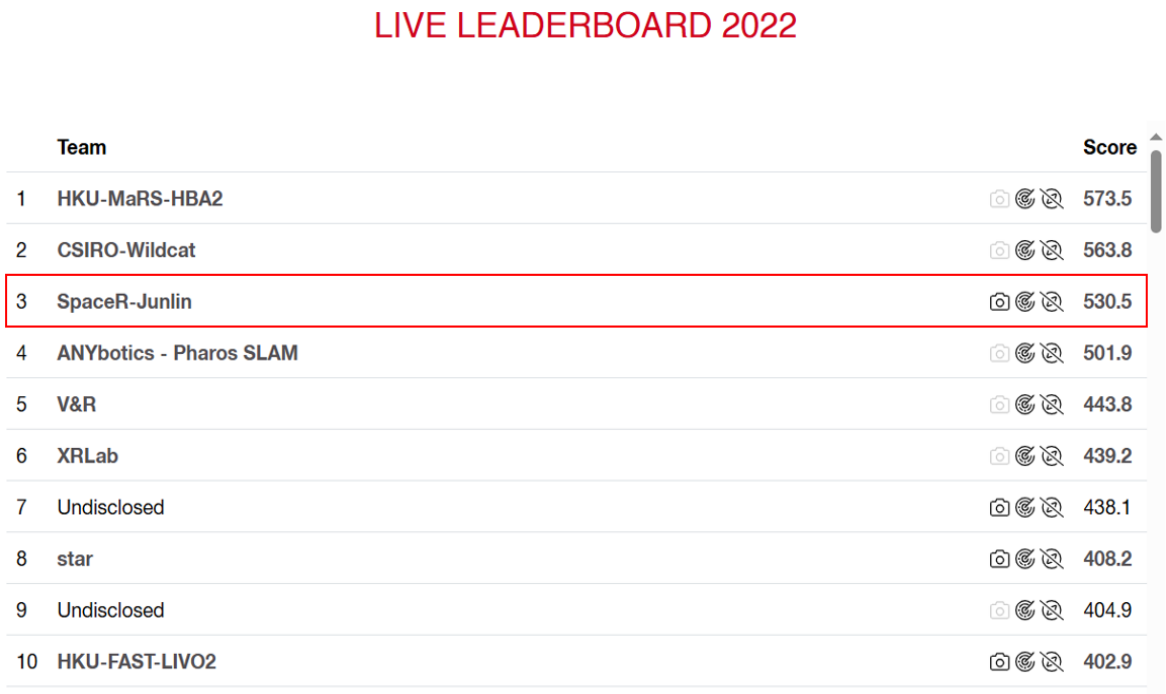


Figure 1.9: My current ranking in Hilti SLAM challenge 2022.

[6] https://hilti-challenge.com/leader-board-2022.html

Chapter 2

Literature Review

2.1 Literature Review for Research Objective I

In the literature, the methods to solve the spatial-temporal calibration are divided into two categories: target-based methods and target-less methods. The target-based methods are more accurate than the target-less methods, benefiting from the prior knowledge of the calibration target. Target-based methods are widely used in multi-sensor calibration tasks [25, 26, 32]. Target-based spatial-temporal hand-eye calibration was first presented in [33]. The spatial-temporal parameters are calibrated by aligning the motion capture trajectory with the camera trajectory, which is obtained by the Perspective-n-Point (PnP) algorithm, with the calibration target. The camera's intrinsic parameters are assumed to be fixed. Therefore, the accuracy of [33] is limited by the PnP algorithm, employed on every single image. After the PnP process, all raw pixels measurements are discarded. The isolation processing of the motion capture sequence and camera sequence cannot uncover the inherent correlation between raw pixel measurements and motion capture measurements. Unlike our target-based calibration algorithm which fully utilizes all the raw sensor data to optimize the spatial-temporal parameters, camera intrinsic and trajectory simultaneously.

However, these methods are only suitable for offline non-real-time calibration and require significant amounts of manual effort. Markers attached to the camera may be removed during experiments, therefore changing the spatial calibration parameter. Moreover, the tem-

poral calibration parameter would also change due to different clocks, transmission delays, data jam, jitter, and skew [34]. Therefore, online target-less calibration method is also worth exploiting, saving human effort and improving the ease of application.

In recent years, online target-less calibration has attracted significant attention in visual-inertial navigation systems (VINS) [24, 35, 36]. Among them, the EKF-based methods are the most popular thanks to their computational efficiency. [24] pointed out that given sufficient motion excitation, the spatial-temporal calibration parameters of VINS are observable. However, under specific motion profiles, some degrees of freedom of the calibration parameters would be unobservable [19]. Identifying potential motion degradation, and avoiding such motion, is crucial to reliably apply these types of algorithms.

2.2 Literature Review for Research Objective II

Sensor fusion of camera and IMU is a well studied topic [37, 6]. Visual-inertial fusion algorithms can be broadly classified into two categories i.e., optimization-based methods and filter-based methods. Optimization-based methods achieve higher theoretical accuracy, which include VINS-Mono [38], Basalt [39] and ORB-SLAM3 [40]. Their high computational cost is a major disadvantage. In contrast, sliding-window filter-based methods, such as the Multi-State Constraint Kalman Filter (MSCKF) [41, 42, 16], are more resource efficient and achieve comparable accuracy.

The combination of a camera and an IMU can only generate relative pose estimation, resulting in the unobservability of global position and absolute yaw [23]. Therefore, pure VIO systems tend to drift over time [43]. Recent works have employed GPS measurement to eliminate this drift. These methods can be divided into loosely-coupled methods and tightly-coupled methods. VINS-Fusion is a loosely-coupled approach, which fuses GPS position measurements and output pose of VIO subsystem [44]. However, the fusion algorithm is unable to improve the VIO subsystem. Therefore, the inner correlations of all measurements are discarded, causing suboptimal localization results. Gomsf is a similar loosely-coupled work [45].

Tightly-coupled methods fully exploit the complementary merits of multi-sensor data, and are promising to further improve the accuracy. A tightly-coupled estimator based on sliding window optimization is proposed in [46]. The rotation between the GPS reference frame and the VIO reference frame is included in the state vector, but the non-synchronization between GPS timestamp and VIO system timestamp is neglected. [47] describes another tightly-coupled optimization-based approach. The comparative experiments with VINS-Fusion have demonstrated that tightly-coupled methods are superior to loosely-coupled methods. However, the transformation between GPS reference frame and VIO reference frame is not estimated in [47]. The closest to our work is [48], which is a tightly-coupled estimator based on MSCKF. The extrinsic parameters between the GPS reference frame and the VIO reference frame are inserted into the state during initialization, however, marginalized after all states are transformed from the VIO reference frame to the GPS reference frame [48].

Consequently, the approach of [48] does not estimate the extrinsic parameters between GPS-VIO online, as they show the extrinsic parameters are unobservable with linear observability analysis. However, linear observability analysis maybe unreliable for a nonlinear system. A locally observable system is sure to be globally observable, but a locally unobservable system maybe globally observable [49, 50, 51]. Our main contribution in this work is to point out that rotational extrinsic parameter is globally observable using nonlinear observability analysis. This novel observability conclusion is similar to our recent accepted work [52], termed as GPS-VWO. The difference between GPS-VIO and GPS-VWO lies in the different kinematic equations, with the former driven by IMU while the latter driven by wheel odometer. As the reference frame of VIO is gravity aligned, the rotational extrinsic parameter of GPS-VIO only has yaw component. Unlike GPS-VIO, the rotational extrinsic parameter of GPS-VWO is 3DoF in general. To analyze the observability of extrinsic parameter, Lie derivative is employed for GPS-VIO, like GPS-VWO [52], considering the nonlinearity of this system.

The unavoidable errors caused by imposing fixed extrinsic parameters after GPS-VIO initialization lead to miss-calculations of the fusion algorithms in long distances. Without online calibration, the estimation error of rotational extrinsic parameter at the start will dete-

riorate the localization accuracy, especially when the GPS noise is relatively large. [46, 53, 54] adopt explicitly online calibration of the rotational extrinsic parameter to improve localization accuracy. To simplify the state estimation complexity, [54] disables the online estimation once the rotational extrinsic parameter is converged. However, neither of them provide a theoretical observability analysis. In this paper, we prove the rotational extrinsic parameter is observable; hence, including it in the state vector is a promising and theoretically guaranteed mean to improve the accuracy of the state estimator.

2.3 Literature Review for Research Objective III

The success of online extrinsic parameter calibration depends on the observability. Remarkable works have studied the observability of extrinsic parameter between IMU and camera. With the help of artificial visual features on the calibration target board, [22] conclude that extrinsic parameter is observable if the moving platform undergoes at least 2DoF rotational excitation. An interesting corollary from [22] is that the observability of extrinsic parameter is independent of translational excitation. However, the conclusion of [22] is limited by the usage of calibration board, and cannot be applied to real operating environments without calibration board. [23] further extend the calibration of extrinsic parameter with target-less approach, and the conclusion is updated. The moving platform should undergo at least 2DoF motion excitation for both rotation and translation, to ensure the observability of extrinsic parameter.

The above-mentioned observability studies miss the analysis of degenerate motion profiles, which could be occurred and unavoidable in practice. As a supplement, [19] thoroughly explore the possible degenerate motion primitives and analyze the impact of degenerate motion on the observability of calibration parameters. We note that the rotational extrinsic parameter is summarized as observable for all identified degenerate motions (see Table I in [19]), except for no motion. However, by observing the top subplot of Fig. 2a in [19], we found that the rotational calibration results exhibit unexpected large RMSE (greater than 1 degree) for the case of pure translational straight line motion, which is clearly different from

other motion cases. Actually, this distinct curve is an indicator for unobservability.

The inconsistency between the observability conclusion and the calibration results motivates the following research question as the main purpose of this work:

Is the rotational extrinsic parameter of (global-pose aided) VIO observable under pure translational straight line motion?

2.4 Literature Review for Research Objective IV

To address the offline calibration problem for IMU and cameras, extensive studies have been conducted in the literature, from theory to practice. Currently, almost all IMU-Camera calibration methods employ a continuous-time state representation based on the B-spline. This type of method can obtain accurate and consistent calibration results with the aid of a calibration board, and the representative work is termed **Kalibr**, developed by [25]. However, Kalibr suffers from high computational cost due to its B-spline based state representation. To reduce computational complexity, [27] further derive a novel and efficient derivative calculation method for the B-spline on Lie groups [55].

Except for continuous-time state representation, discrete-time state representation can also be applied to the spatial-temporal IMU-Camera calibration task. Surprisingly, there has been rare exploration in this direction in the decade after the release of the Kalibr toolbox. Many researchers believe that discrete-time state representation is difficult or inferior for temporal calibration [25, 56, 57]. For example, authors of Kalibr, [25] are concerned that discrete-time state representation requires a new state at each measurement time, which could be challenging for the utilization of high-frequency IMU measurements, and subsequent estimator design for temporal calibration.

In fact, this concern can be addressed by aggregating IMU measurements over a short period of time. Inspired by IMU preintegration [58, 59], we propose a novel optimization-based IMU-Camera calibration method with discrete-time state representation. Several IMU measurements between two consecutive images are aggregated as one pseudo-measurement, thus greatly reducing the state dimensions that need to be optimized.

MVIS [29] is another discrete-time calibration method based on IMU preintegration, with appealing full calibration capability. However, due to the use of a gravity-aligned reference frame, MVIS sacrifices efficiency by introducing 3D feature positions in the state vector. Instead, our method eliminates features from the state vector by adopting a reference frame similar to Kalibr and Basalt, thus fully unleashing the efficiency power of discrete-time calibration.

Chapter 3

Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor

In robotics, motion capture systems have been widely used to measure the accuracy of localization algorithms. Moreover, this infrastructure can also be used for other computer vision tasks, such as the evaluation of Visual (-Inertial) SLAM dynamic initialization, multi-object tracking, or automatic annotation. Yet, to work optimally, these functionalities require having accurate and reliable spatial-temporal calibration parameters between the **camera** and the **global pose sensor**. In this study, we provide two novel solutions to estimate these calibration parameters. Firstly, we design an offline target-based method with high accuracy and consistency. Spatial-temporal parameters, camera intrinsic, and trajectory are optimized simultaneously. Then, we propose an online target-less method, eliminating the need for a calibration target and enabling the estimation of time-varying spatial-temporal parameters. Additionally, we perform detailed observability analysis for the target-less method. Our theoretical findings regarding observability are validated by simulation experiments and provide explainable guidelines for calibration. Finally, the accuracy and consistency of two proposed methods are evaluated with hand-held real-world datasets where traditional hand-eye calibration method do not work.

3.1 Related paper

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor." 2024 International Conference on 3D Vision (3DV 2024).

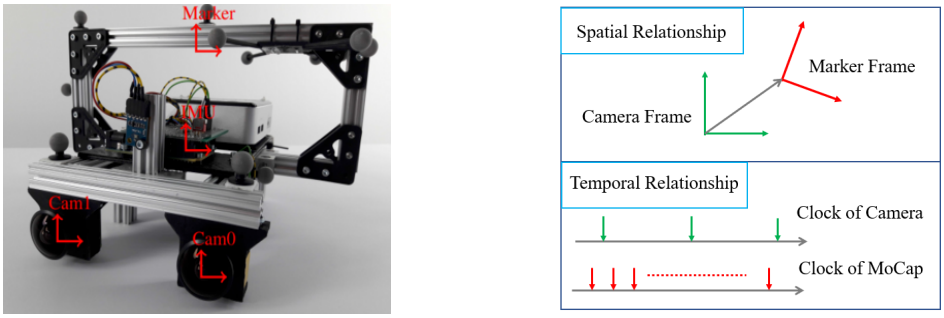
3.2 Relationships to other chapters

This chapter provides two solutions to estimate interested calibration parameters. Both target-based and target-less methods are considered. These principles can be applied to any calibration task. For Chapter 4 and Chapter 5, target-less methods are adopt. For Chapter 6, it is essentially a target-based method.

3.3 Introduction

Nowadays, motion capture systems are widely used to perform 6DoF pose tracking thanks to their high accuracy (sub-millimeter). In odometry and SLAM research, most datasets leverage these to provide the ground truth pose [60, 61, 62]. The collection platform from [61] shown in Fig. 3.1a displays some passive markers typically associated with motion capture systems. Aside from its application to localization methods, the potential of motion capture systems in the field of computer vision has not been fully exploited. The key is the spatial-temporal calibration parameters of the camera and the global pose sensor (see Fig. 3.1b).

For instance, in Fig. 3.2b, we assume a target tracking or automatic labeling task, performed with the motion capture system. The camera $\{C\}$ is rigidly linked with the marker frame $\{M\}$ tracked by the motion capture system. The target is regarded as a point f. The motion capture system provides ${}^{G}p_f$ and $\left\{ \begin{smallmatrix} G \\ M \end{smallmatrix} q \quad {}^{G}p_M \right\}$. Given the spatial-temporal



(a) (b)

Figure 3.1: (a) Photo of the sensor setup, taken from [61]. (b) The spatial-temporal relationship between the camera measurements and the global pose measurements.

calibration parameters linked $\{M\}$ and $\{C\}$, the image coordinates of f can be obtained automatically via rigid body link ($f \rightarrow G \rightarrow M \rightarrow C$).

The above example illustrates the benefits of having a spatial-temporal calibration between a camera and a global pose sensor. In the literature, the methods to solve the spatial-temporal calibration are divided into two categories: target-based methods and target-less methods. The target-based methods are more accurate than the target-less methods, benefiting from the prior knowledge of the calibration target. Target-based methods are widely used in multi-sensor calibration tasks [25, 26, 32]. Target-based spatial-temporal hand-eye calibration was first presented in [33]. The spatial-temporal parameters are calibrated by aligning the motion capture trajectory with the camera trajectory, which is obtained by the Perspective-n-Point (PnP) algorithm, with the calibration target. The camera's intrinsic parameters are assumed to be fixed. Therefore, the accuracy of [33] is limited by the PnP algorithm, employed on every single image. After the PnP process, all raw pixels measurements are discarded. The isolation processing of the motion capture sequence and camera sequence cannot uncover the inherent correlation between raw pixel measurements and motion capture measurements. Unlike our target-based calibration algorithm which fully utilizes all the raw sensor data to optimize the spatial-temporal parameters, camera intrinsic and trajectory simultaneously.

However, these methods are only suitable for offline non-real-time calibration and require significant amounts of manual effort. Markers attached to the camera may be removed

during experiments, therefore changing the spatial calibration parameter. Moreover, the temporal calibration parameter would also change due to different clocks, transmission delays, data jam, jitter, and skew [34]. Therefore, online target-less calibration method is also worth exploiting, saving human effort and improving the ease of application.

In recent years, online target-less calibration has attracted significant attention in visual-inertial navigation systems (VINS) [24, 35, 36]. Among them, the EKF-based methods are the most popular thanks to their computational efficiency. [24] pointed out that given sufficient motion excitation, the spatial-temporal calibration parameters of VINS are observable. However, under specific motion profiles, some degrees of freedom of the calibration parameters would be unobservable [19]. Identifying potential motion degradation, and avoiding such motion, is crucial to reliably apply these types of algorithms.

The contributions of this work are summarized as:

- To our knowledge, this is the first work to simultaneously calibrate spatial-temporal parameters of the camera and the global pose sensor, with raw monocular camera pixel measurements and global pose measurements.
- We propose two novel approaches to estimate the spatial-temporal parameters. Both target-based and target-less methods are considered.
- We provide detailed observability analysis for the proposed target-less calibration method and identify the degenerated motions that may occur in practice, causing partial calibration parameters unobservable.
- We verify the degenerate motions in simulation and evaluate the accuracy and consistency of two proposed algorithms with hand-held real-world datasets.
- We demonstrate the applicability of online calibration time-varying spatial-temporal parameters for the target-less method.

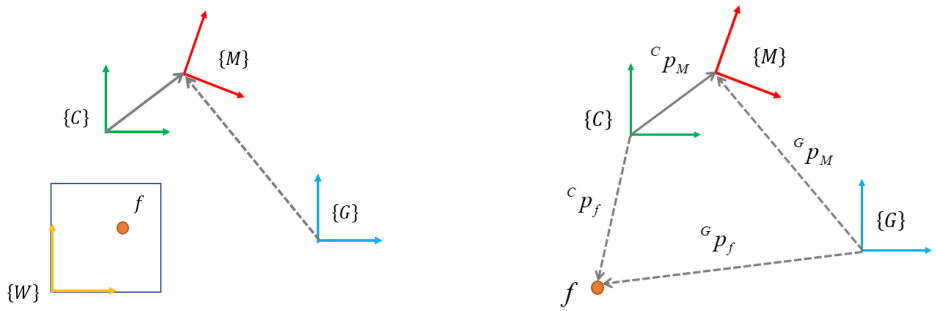


Figure 3.2: (a) Coordinate frames for the target-based method. (b) Coordinate frames for the target-less method.

3.4 Notation

As shown in Fig. 3.2, $\{G\}$ represents the global reference frame of the motion capture system. $\{M\}$ and $\{C\}$ represent the marker frame and the camera frame respectively. In this paper, **"marker"** is an equivalent term of **"global pose sensor"**, as the 6DoF movement of frame $\{M\}$ could be tracked by the motion capture system. The 6DoF rigid body transformation between $\{M\}$ and $\{C\}$, ${}^{C}_{M}T$, is the spatial calibration parameter. In our formulation, the camera time clock is treated as the time reference in the estimators. The time offset between the marker clock and the camera clock is the temporal calibration parameter t_d. If the timestamp at the camera clock is t_C, then the corresponding timestamp at the marker clock is:

$$t_M = t_C + t_d \tag{3.1}$$

We use ${}^{G}(\bullet)$ to represent a physical quantity in the frame $\{G\}$. The position of a point M in the frame $\{G\}$ is expressed as ${}^{G}p_M$. The velocity of a point M in the frame $\{G\}$ is expressed as ${}^{G}v_M$. The local angular velocity of $\{M\}$ is denoted as ω. A Unit quaternion is employed to represent the rotation of a rigid body [63]. ${}^{M}_{G}q$ represents the orientation of the frame $\{M\}$ with respect to the frame $\{G\}$, and its corresponding rotation matrix is ${}^{M}_{G}R$. $[\bullet]_\times$ is denoted as the skew symmetric matrix corresponding to a three-dimensional vector. The transpose of a matrix is $[\bullet]^T$.

3.5 Target-based Calibration

A target-based calibration method which adopts offline full-batch nonlinear least squares optimization is designed to provide high accurate and consistent solutions for calibration parameters.

We use a grid of AprilTag [64] as the calibration target, as shown in Fig. 3.3b. The coordinate frames involved in target-based method are depicted in Fig. 3.2a. Compared with Fig. 3.2b, additional frame $\{W\}$ is built and fixed on the calibration target.

Suppose that the timestamp of the ith image is t_i. The image coordinate of the jth AprilTag corner f_j detected in the ith image is u_{ij}. Its associated 3D coordinates ${}^{W}p_{f_j}$ in $\{W\}$ is known. The optimization variables are defined as:

$$\chi = \left\{ \begin{array}{cccccccc} {}^{W}_{C_1}T & \cdots & {}^{W}_{C_N}T & {}^{G}_{W}T & {}^{C}_{M}T & t_d & \varsigma \end{array} \right\} \tag{3.2}$$

Where N is the image numbers. χ includes the all camera poses ${}^{W}_{C_i}T, i = 1 \cdots N$, the rigid body transformation between $\{W\}$ and $\{G\}$, the spatial-temporal calibration parameters $\left\{ \begin{array}{cc} {}^{C}_{M}T & t_d \end{array} \right\}$, and the vector of camera intrinsic parameters ς. By integrating all raw image pixel measurements and global pose measurements, we formulate the least squares optimization as:

$$\begin{aligned} \chi &= \arg\min \left\{ \sum_{i=1}^{N} \sum_{j=1}^{K} \rho\left(r_{ij}\right) + \sum_{i=1}^{N} \rho\left(r_{gi}\right) \right\} \\ r_{ij} &= \pi\left({}^{C_i}_{W}T {}^{W}p_{f_j}, \varsigma\right) - u_{ij} \\ r_{gi} &= Log\left({}^{M}_{G}T\left(t_i + t_d\right) {}^{G}_{W}T {}^{W}_{C_i}T {}^{C}_{M}T\right) \end{aligned} \tag{3.3}$$

Where K is the corner numbers for each image. $\rho\left(\bullet\right)$ is a robust kernel function [65]. $\pi\left(\bullet, \bullet\right)$ is a fixed camera projection function [66, 67]. $Log\left(\bullet\right)$ maps the element on a Lie group to the tangent space vector [55].

${}^{M}_{G}T\left(t_i + t_d\right)$ is the interpolated global pose measurement. To calculate ${}^{M}_{G}T\left(t_i + t_d\right)$, we find two closet timestamps over all global pose measurements, t_a and t_b, which subject to $t_a \leq t_i + t_d < t_b$. Two corresponding pose measurements are ${}^{M_a}_{G}T$ and ${}^{M_b}_{G}T$ respectively. Using linear interpolation with two bounding poses, the synthetic measurement at $t_i + t_d$ is

expressed as:

$$
\begin{aligned}
&{}^{M}_{G}T\left(t_i+t_d\right)=Exp\left(\lambda Log\left({}^{M_b}_{G}T{}^{M_a}_{G}T^{-1}\right)\right){}^{M_a}_{G}T\\
&\lambda=\left(t_i+t_d-t_a\right)/\left(t_b-t_a\right)
\end{aligned}
\tag{3.4}
$$

$Exp\left(\bullet\right)$ is the inverse operation of $Log\left(\bullet\right)$ [55].

Jacobians of residuals in Eq. (3.3) with respect to the optimization variables χ are calculated according to the chain rule and provided in Sec. 3.10 of supplementary material. The Levenberg-Marquardt algorithm is adpot to minimize Eq. (3.3) and update the optimal estimation iteratively.

Differentiate from [33], the proposed target-based method is able to optimize and refine the spatial-temporal calibration parameters, the transformation between $\{W\}$ and $\{G\}$, the camera intrinsic ς and trajectory ${}^{W}_{C_i}T, i=1\cdots N$ simultaneously, without information loss.

3.6 Target-less Calibration

To alleviate the need for calibration target and enable time-varying parameters calibration during the operation, we provide an alternative online EKF-based target-less calibration method. Coordinate frames are shown in Fig. 3.2b.

3.6.1 State Vector

The EKF state vector inspired by MSCKF [41] includes the marker state, the spatial-temporal calibration parameters, the camera intrinsic parameters, augmented N marker states and up to L augmented features:

$$
\begin{aligned}
x&=\begin{bmatrix} x_M^T & x_{calib}^T & x_c^T & x_f^T \end{bmatrix}^T\\
x_M&=\begin{bmatrix} {}^{M}_{G}q^T & {}^{G}p_M^T & \omega^T & {}^{G}v_M^T \end{bmatrix}^T\\
x_{calib}&=\begin{bmatrix} {}^{C}_{M}q^T & {}^{C}p_M^T & t_d & \varsigma \end{bmatrix}^T\\
x_c&=\begin{bmatrix} x_{c_1}^T & \cdots & x_{c_N}^T \end{bmatrix}^T \qquad x_{c_i}=\begin{bmatrix} {}^{M_i}_{G}q^T & {}^{G}p_{M_i}^T \end{bmatrix}^T\\
x_f&=\begin{bmatrix} {}^{G}p_{f_1}^T & \cdots & {}^{G}p_{f_L}^T \end{bmatrix}^T
\end{aligned}
\tag{3.5}
$$

Where x_M is the current marker state at the camera clock. Calibration parameter x_{calib} includes the 6DoF transformation $\left\{ \begin{smallmatrix} C \\ M \end{smallmatrix} q \quad {}^C p_M \right\}$, the time offset t_d and the camera intrinsic parameters ς. x_c is the augmented marker states, which is obtained by cloning the first two physical quantities of x_M at different image times. N is the sliding window size, a fixed parameter. The pose clones in the sliding window are utilized to triangulate environmental feature points. ${}^G p_{f_j}$ is an augmented feature, or termed as a SLAM feature [68, 69, 16].

Angular and linear velocity (ω and ${}^G v_M$) are included to predict the motion because the measurements provided by motion capture system may be intermittent. Moreover, they are needed to estimate time offset (see Eq. (3.9)).

3.6.2 Constant Velocity Propagation

Referring to previous study on trajectory estimation [70, 71], a constant-velocity motion prior is applied. x_M is propagated forward based on the constant velocity motion model. The kinematic model can be described as:

$$\begin{aligned} {}^M_G \dot{q} &= \tfrac{1}{2} \Omega(\omega) {}^M_G q, \quad {}^G \dot{p}_M = {}^G v_M \\ \dot{\omega} &= n_\omega, \quad {}^G \dot{v}_M = n_v \end{aligned} \tag{3.6}$$

$\Omega(\omega) = \begin{bmatrix} -[\omega]_\times & \omega \\ -\omega^T & 0 \end{bmatrix}$. $n_{[\bullet]}$ represents the zero mean Gaussian noise of $[\bullet]$, which is a hyperparameter. These hyperparameters can be determinated in advance using existing approaches [70, 72].

By linearizing Eq. (3.6) at the current state estimation, the state transition matrix from

time t_0 to time t_k can be analytically calculated as follows:

$$\Phi_M(t_k, t_0) = \begin{bmatrix} A & 0_3 & B & 0_3 \\ 0_3 & I_3 & 0_3 & I_3\Delta t \\ 0_3 & 0_3 & I_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & I_3 \end{bmatrix} \tag{3.7}$$
$$A = {}^{M_k}_{G}R\, {}^{M_0}_{G}R^T$$
$$B = {}^{M_k}_{G}R\, {}^{M_0}_{G}R^T J_r(-\omega\Delta t)\Delta t$$

Where $J_r(\bullet)$ is the right Jacobian of SO(3) [73].

3.6.3 Visual Measurement Update

For a new coming image with the timestamp t, we clone the latest marker pose and augment it to the state vector x to track the camera pose. According to Eq. (3.1), the corresponding marker timestamp is $t + t_d$. The new cloned marker pose is:

$$x_{c_{new}} = \begin{bmatrix} {}^{M}_{G}q(t + t_d) \\ {}^{G}p_M(t + t_d) \end{bmatrix} \tag{3.8}$$

The state augmentation Jacobian with respect to $\begin{bmatrix} {}^{M}_{G}q^T & {}^{G}p_M^T & t_d \end{bmatrix}^T$ is calculated as:

$$H_{aug} = \begin{bmatrix} I_3 & 0_3 & \omega \\ 0_3 & I_3 & {}^{G}v_M \end{bmatrix} \tag{3.9}$$

After the state augmentation is completed, we check the sliding window size and marginalize the oldest clone state if the window size exceeds N. The carefully selected feature points are used to update the poses over the sliding window and the position of the feature points. The feature measurement model can be written as:

$$z_f = \pi\left({}^{C}p_f, \varsigma\right)$$
$${}^{C}p_f = {}^{C}_{M}R\, {}^{M}_{G}R\left({}^{G}p_f - {}^{G}p_M\right) + {}^{C}p_M \tag{3.10}$$

The subset of state variables related to z_f is noted as[1]:

$$x_s = \begin{bmatrix} {}^M_G q^T & {}^G p_M^T & {}^C_M q^T & {}^C p_M^T & {}^G p_f^T \end{bmatrix}^T \tag{3.11}$$

The feature measurement Jacobian is calculated as:

$$\begin{aligned} H_f &= \frac{\partial z_f}{\partial {}^C p_f} {}^C_M R {}^M_G R \begin{bmatrix} J_1 & -I_3 & J_2 & {}^G_M R {}^M_C R & I_3 \end{bmatrix} \\ J_1 &= \left[({}^G p_f - {}^G p_M) \right]_\times {}^G_M R \\ J_2 &= \left[({}^G p_f - {}^G p_M) \right]_\times {}^G_M R {}^M_C R \end{aligned} \tag{3.12}$$

More details about feature detection, tracking, outlier rejection, triangulation, sliding window update scheme and covariance management can be found in [16].

3.6.4 Global Pose Measurement Update

The timestamp of the global pose measurements t, provided at the marker clock, are shifted by t_d, $t - t_d$. The corrected global pose measurement is used to update x_M. The global pose measurement model can be written as:

$$z_g = \begin{bmatrix} {}^M_G q \\ {}^G p_M \end{bmatrix} \tag{3.13}$$

The global pose measurement Jacobian with respect to $\begin{bmatrix} {}^M_G q^T & {}^G p_M^T \end{bmatrix}^T$ is calculated as:

$$H_g = \begin{bmatrix} I_3 & 0_3 \\ 0_3 & I_3 \end{bmatrix} \tag{3.14}$$

3.7 Observability Analysis

System observability plays an important role in state estimation. To study the potential calibration failures, we perform observability analysis for the linearized system [74] derived

[1]The camera intrinsic ς is omitted here because it does not affect the subsequent observability analysis in Sec. 3.7.

in the target-less calibration. To the best of our knowledge, this is the first time that a paper studies the observability of the spatial-temporal parameters between the camera and the marker.

Since the state vector couples both motion variables and calibration parameters together by covariance matrix. It is expected that the success of calibration depends on motion profiles. Identifying the potential degenerate motion profiles that adversely affect the calibration accuracy can guide the calibration process in practice.

To concise the presentation, we do not consider clone states in the state vector. ω and ${}^{G}v_{M}$ are also neglected as their observablity property is consistent with the marker pose. And only one SLAM feature is kept. The results can be extended to general cases [75, 76]. The system state vector becomes:

$$x = \begin{bmatrix} {}^{M}_{G}q^{T} & {}^{G}p_{M}^{T} & {}^{C}_{M}q^{T} & {}^{C}p_{M}^{T} & t_d & {}^{G}p_{f}^{T} \end{bmatrix}^{T} \tag{3.15}$$

The state transition matrix becomes:

$$\Phi\left(t_k, t_0\right) = \begin{bmatrix} A & & \\ & I_{13} & \end{bmatrix} \tag{3.16}$$

A is defined in Eq. (3.7).

H_{aug} in Eq. (3.9), H_f in Eq. (3.12) and H_g in Eq. (3.14) are stacked to construct the general Jacobian of the state:

$$H_k = \begin{bmatrix} I_3 & 0_3 & 0_3 & 0_3 & \omega & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & {}^{G}v_{M} & 0_3 \\ J_1 & -I_3 & J_2 & {}^{G}_{M}R^{M}_{C}R & 0_{3\times1} & I_3 \\ I_3 & 0_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \end{bmatrix} \tag{3.17}$$

The common factor $\frac{\partial z_f}{\partial {}^{C}p_f}{}^{C}_{M}R^{M}_{G}R$ in Eq. (3.12) is ignored here because it does not affect

the observability analysis. Now the observability matrix would be constructed as [74]:

$$
\begin{aligned}
O &= \begin{bmatrix} \cdots & O_k^T & \cdots \end{bmatrix}^T \\
O_k &= H_k \Phi\left(t_k, t_0\right) \\
&= \begin{bmatrix}
A & 0_3 & 0_3 & 0_3 & \omega & 0_3 \\
0_3 & I_3 & 0_3 & 0_3 & {}^G v_M & 0_3 \\
J_1 A & -I_3 & J_2 & {}^G_M R^M_C R & 0_{3\times 1} & I_3 \\
A & 0_3 & 0_3 & 0_3 & 0_{3\times 1} & 0_3 \\
0_3 & I_3 & 0_3 & 0_3 & 0_{3\times 1} & 0_3
\end{bmatrix}
\end{aligned}
\tag{3.18}
$$

We note that for generic motions, O is a time varying matrix, whose columns are linearly independent. At this point, we state that the spatial-temporal calibration parameters are observable with fully excited 6DoF motions.

However, under the special motion situation, the linear independent relationship is no longer maintained, resulting in some degrees of freedom of the calibration parameters becoming unobservable.

Lemma 3.7.1. *If the frame $\{M\}$ performs pure translation (no rotation) motion, ${}^C p_M$ is unobservable. The corresponding right null space of O is:*

$$
N_1 = \begin{bmatrix} 0_{3\times 9} & I_3 & 0_{3\times 1} & -\left({}^G_M R^M_C R\right)^T \end{bmatrix}^T \tag{3.19}
$$

Proof. The fact that N_1 is indeed the right null space of O can be verified by multiplying O_k with N_1. $O_k N_1 = 0$ is hold for any k. And we note that N_1 is a constant matrix. Since there is no rotation, ${}^G_M R$ is a constant matrix. Hence, N_1 belongs to the right null space of O. N_1 indicates that the unobservable direction is ${}^C p_M$. □

Lemma 3.7.2. *If the the frame $\{M\}$ rotates around a constant axis ω_2 during the generic translation motion, the unobservable directions depend on the projection of ω_2 in the frame $\{C\}$, and the corresponding right null space of O is:*

$$
N_2 = \begin{bmatrix} 0_{1\times 9} & \left({}^C_M R \omega_2\right)^T & 0 & -\left({}^G_M R \omega_2\right)^T \end{bmatrix}^T \tag{3.20}
$$

Proof. Similarly, we verify that $O_k N_2 = 0$ is hold for any k. Since ω and ω_2 are parallel at this setting, for any given ω_2, the time derivative of ${}^G_M R\omega_2$ is given by:

$$\frac{d\left({}^G_M R\omega_2\right)}{dt} = \left(\frac{d\left({}^G_M R\right)}{dt}\right)\omega_2 = {}^G_M R[\omega]_\times \omega_2 = 0 \tag{3.21}$$

This proves that N_2 is a constant matrix and belongs to the right null space of O. N_2 indicates that the unobservable directions are from ${}^C p_M$, and dependent on the non-zero components of ${}^C_M R\omega_2$, or ${}^C_M R\omega$. □

There could be some other degeneration motion primitives that have not been considered, such as constant angular and linear velocities, constant angular velocity and linear accelerations. We can find these two are special cases for Lemma 3.7.2. In this paper, we do not derive all degeneration cases where the full column rank condition of O breaks.

As a final remark, we note that the translation calibration parameter ${}^C p_M$ is more sensitive to different motions, compared to the rotation and temporal calibration parameter. These theory findings are important for the calibration, as these degenerate motions are likely to occur in practice, such as the planer motion of wheeled robot and the pure translation of flying robot. We run real-world experiments on random generic trajectories with full excitation to avoid these potential specific degenerate trajectories.

3.8 Experiments

We state again that the inputs of two proposed calibration methods are global pose measurements and monocular image stream. Firstly, the observability analysis in Sec. 3.7 is verified by generating these measurements in the simulation environment. Then the real-world datasets are used to test the calibration accuracy and consistency. The target-based method requires the calibration target to be located in the field of view of the image and geometric prior about the calibration target. Finally, an example of calibrating time-varying spatial-temporal parameters is presented with the online target-less method.

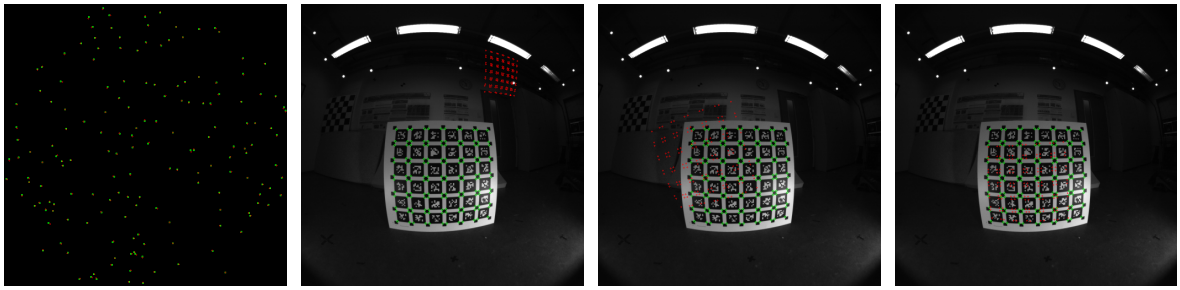
3.8.1 Validation of the Observability Analysis

The simulated environment includes randomly generated 3D points to be captured by images. The characteristics of the simulated sensors are consistent with those of the actual sensors used in the real-world. Global pose measurements are reported in 120Hz. Images are received in 20Hz. The Gaussian noises of the sensors are generated and added into the synthetic measurements. Fig. 3.3a shows the synthetic feature points and the corresponding reprojected points in one simulated image during the visual update process. The translation motion of the marker frame is simulated as a sinusoidal trajectory, which is widely used in calibration tasks [24, 36, 77].

To validate the observability assertion in Sec. 3.7, we set ${}^{C}_{M}R$ as I_3, and design five rotation motion cases.

- Case1: $\omega = \begin{bmatrix} 0.4\cos(1.5t) & 0.4\sin(t) & 0 \end{bmatrix}^T$.
- Case2: $\omega = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$.
- Case3: $\omega = \begin{bmatrix} 0.4 & 0 & 0 \end{bmatrix}^T$.
- Case4: $\omega = \begin{bmatrix} 0 & 0.5 & 0.6 \end{bmatrix}^T$.
- Case5: $\omega = \begin{bmatrix} 0.1 & 0.2 & 0.3 \end{bmatrix}^T$.

The calibration results of these cases are presented in Fig. 3.4. The initial rotation error is $\begin{bmatrix} 20^\circ & 20^\circ & -20^\circ \end{bmatrix}^T$. The initial translation error is $\begin{bmatrix} -5 & 15 & -10 \end{bmatrix}^T$ cm. The initial time offset error is 50 ms. Case1 corresponds to the generic motion with full excitation. It is clear that the estimation errors of all calibration parameters converge perfectly to near zero within 10s. All calibration parameters are observable in this case. Case 2 corresponds to a pure translation (no rotation) motion. The estimation error of the translation calibration parameter and its 1σ bound can not approach 0, thus this parameter is unobservable. While the rotation and temporal calibration parameters are still observable. Case3, Case4, and Case5 correspond to the constant axis rotational motion. The non-zero components of this

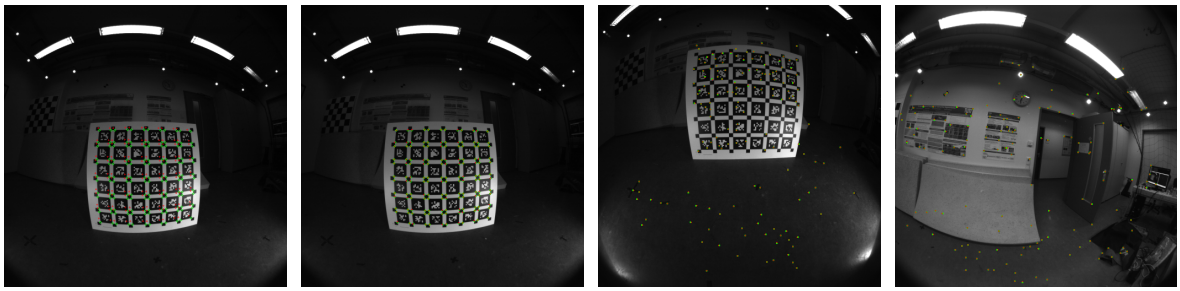


(a) Tracked features and reprojected features in the simulation.

(b) The 1st iteration.

(c) The 2nd iteration.

(d) The 3rd iteration.



(e) The 4th iteration.

(f) The 5th iteration.

(g) Image update in the environment with calibration target.

(h) Image update in the environment without calibration target.

Figure 3.3: Expected feature positions (green) and predicted feature positions (red) in the image.

axis indicate the unobservable directions. For example, the rotation axis of Case3 only has non-zero component in the x-axis. Thus, the x-direction of the translation calibration parameter is unobservable, yet y and z direction are still observable, as shown in Fig. 3.4. The similar analysis also applies to Case 4 and Case 5.

3.8.2 Real-World Experiments

Firstly we present the rationale of dataset selection for real-world experiments. For the target-less method, the simulation experiments in Sec. 3.8.1 show that it is advised to choose the fully excited 6DoF trajectory. The experiments in [36] also inspire us to utilize the fully excited hand-held TUM-VI Dataset [61] instead of under-actuated dataset, such as EuRoC MAV Dataset [60]. TUM-VI Dataset contains multiple sequences with or without

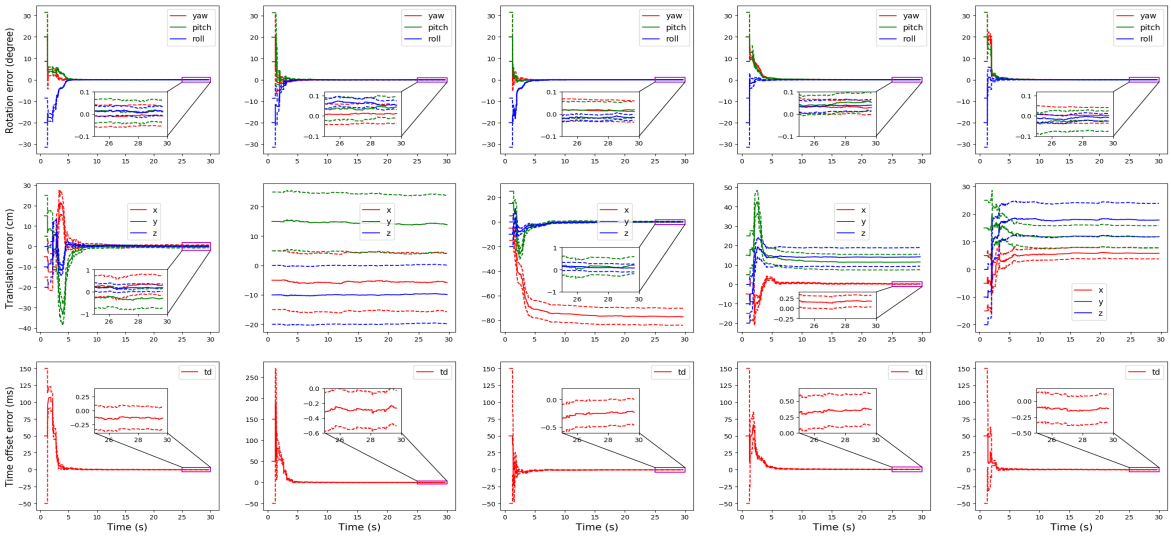


Figure 3.4: Errors (solid lines) and 1σ bounds (dashed lines) of the spatial-temporal calibration parameters. x-axis represents time in seconds. Left to right corresponds to Case1 to Case5 in Sec. 3.8.1. The estimation error of the rotation and temporal calibration parameters perfectly approach to zero for any cases. While the convergence results of the translation calibration parameter are varied from case to case.

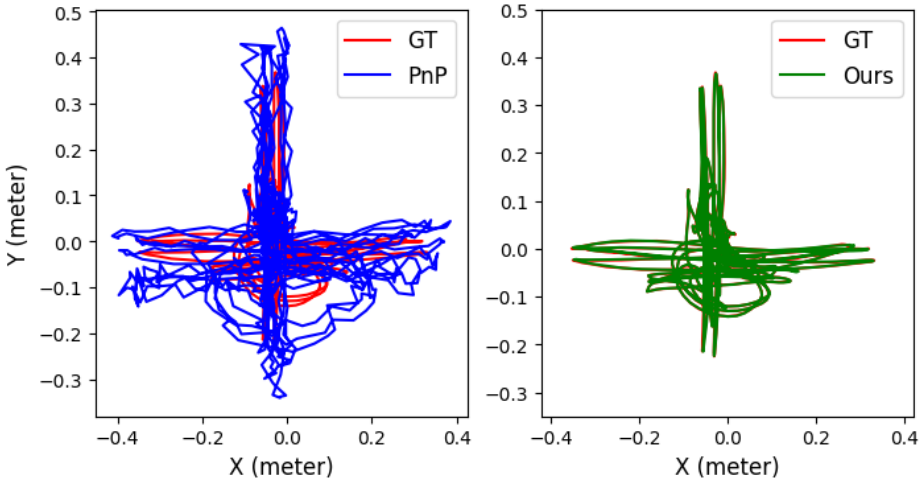


Figure 3.5: $imu1$ is used. GT: groundtruth trajectory output from motion capture system. PnP: camera trajectory output from PnP algorithm. Ours: refined camera trajectory ${}^{W}_{C_i}T, i = 1 \cdots N$.

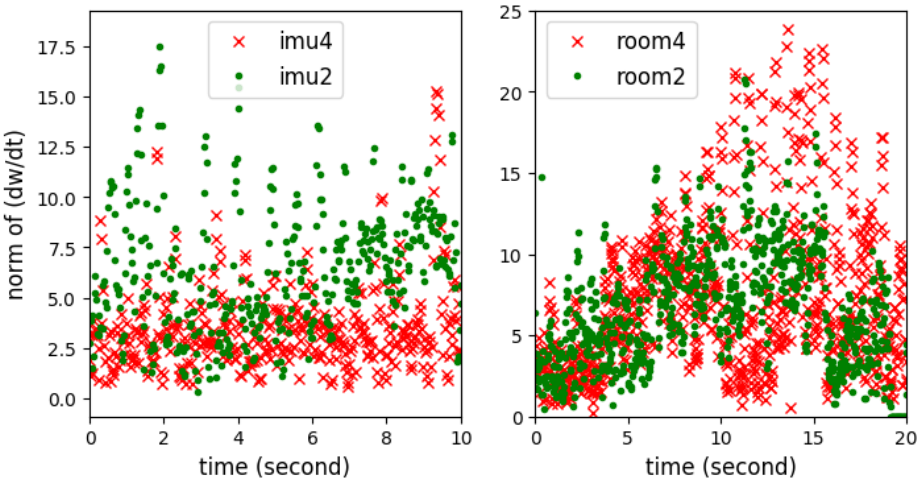


Figure 3.6: Norm of $d\omega/dt$.

Table 3.1: Average RMSE of the calibration results (mean value $\pm$ standard deviation) over 50 Monte-Carlo trials. Method1: target-less method. Method2: target-based method. L: left camera is used. R: right camera is used.

Sequence	Rotation (deg)		Translation (cm)		Time offset (ms)	
	Method1	Method2	Method1	Method2	Method1	Method2
imu1 (L)	0.124 $\pm$ 0.051	0.032 $\pm$ 4.74e-05	0.572 $\pm$ 0.126	0.103 $\pm$ 1.65e-05	0.543 $\pm$ 0.128	0.339 $\pm$ 0.00e-05
imu2 (L)	0.142 $\pm$ 0.043	0.035 $\pm$ 4.63e-07	0.336 $\pm$ 0.076	0.090 $\pm$ 0.00e-07	0.149 $\pm$ 0.059	0.300 $\pm$ 0.00e-07
imu3 (L)	0.074 $\pm$ 0.038	0.048 $\pm$ 0.00e-07	0.686 $\pm$ 0.141	0.146 $\pm$ 0.00e-07	0.088 $\pm$ 0.069	0.757 $\pm$ 0.00e-07
imu4 (L)	0.083 $\pm$ 0.053	0.065 $\pm$ 3.91e-07	1.014 $\pm$ 0.115	0.125 $\pm$ 0.00e-07	1.156 $\pm$ 0.144	0.960 $\pm$ 0.00e-07
imu1 (R)	0.075 $\pm$ 0.024	0.027 $\pm$ 9.97e-07	1.040 $\pm$ 0.228	0.085 $\pm$ 0.00e-07	0.432 $\pm$ 0.132	0.335 $\pm$ 0.00e-07
imu2 (R)	0.180 $\pm$ 0.044	0.034 $\pm$ 0.00e-07	0.465 $\pm$ 0.270	0.075 $\pm$ 0.00e-07	0.161 $\pm$ 0.082	0.305 $\pm$ 0.00e-07
imu3 (R)	0.125 $\pm$ 0.051	0.038 $\pm$ 0.00e-07	0.719 $\pm$ 0.101	0.136 $\pm$ 0.00e-07	0.091 $\pm$ 0.096	0.766 $\pm$ 0.00e-07
imu4 (R)	0.087 $\pm$ 0.039	0.050 $\pm$ 3.22e-07	1.077 $\pm$ 0.119	0.132 $\pm$ 0.00e-07	1.449 $\pm$ 0.147	0.955 $\pm$ 0.00e-07

calibration target. Each sequence provides images at 20Hz, global pose measurements at 120Hz. These raw measurements together with IMU measurements are post-processed to ensure time-synchronization. Thus it is convenient to set the time offset by manually shifting the timestamps of the global pose measurements with a certain value. The shifted time offset is the reference value of the temporal parameter. As [61] has leveraged IMU to align the marker frame to the IMU frame, the transformation from IMU to camera [27], is also the reference value of the interested spatial parameter.

For each selected dataset, we run the specific calibration method multiple times to examine the statistical properties. Reference value is perturbed to perform a Monte-Carlo trial. The perturbed calibration parameters are set as initial calibration guess. Random errors drawn from zero-mean Gaussian distributions are added to reference values. For rotation and translation parameter, 1σ values of the error distribution along each axis are 20° and 10 cm respectively. For temporal parameter, the 1σ value is set as 50 ms.

Environments with target

Sequence $\{imu1 \sim imu4\}$ is selected because the environments of these datasets contain the calibration target.

[33] can not work for these sequences due to the relatively large trajectory noise output by PnP algorithm, as shown in Fig. 3.5. The absolute trajectory error (ATE) of the PnP trajectory is 7.29 cm, while the optimized trajectory of our target-based method has an ATE of only 0.28 cm. Clearly, the accuracy of camera trajectory has significantly improvement by fully utilizing the raw measurements. Additional comparison results are provided in Sec. 3.11 of supplementary material.

To visualize the estimation accuracy of the calibration parameters of the target-based method, the predicted feature position linked with calibration parameters is defined as:

$$\begin{aligned} z &= \pi\left({}^{C}p_f, \varsigma\right) \\ {}^{C}p_f &= {}^{C}_{M}T\,{}^{M}_{G}T\,(t+t_d)\,{}^{G}_{W}T\,{}^{W}p_f \end{aligned} \tag{3.22}$$

Where f denotes the AprilTag corner. t is the image timestamp. ${}^{G}_{W}T$, ${}^{C}_{M}T$, t_d, and ς are variables from Eq. (3.2).

For a specific run of the target-based method, the iterative update results are visualized from Fig. 3.3b to Fig. 3.3f. After 5 iterations, all predicted feature positions are perfectly close to expected feature positions. Fig. 3.3g shows the feature points update of the target-less method. The predicted feature position is obtained via Eq. (3.10).

When using the left camera, the RMSE of the calibration results are shown in Tab. 3.1. As expected, the calibration accuracy and consistency of the target-based method are better than the target-less method. When using the right camera, the corresponding results are also shown in Tab. 3.1. Both calibration methods demonstrate similar accuracy and consistency for left and right camera.

Compared with the target-based method, the target-less method's accuracy is affected by imperfect visual feature tracking and numerical precision of the triangulation process of visual landmarks. In addition, the target-less method is an online estimator, which can not use all available measurements simultaneously.

It is worth noting that the dataset itself or the trajectory characteristic has impacts on the calibration accuracy for both methods. For example, the estimation accuracy of the translation calibration parameter of $imu2$ is better than that of $imu4$. Inspired by the observability analysis in Sec. 3.7 and Sec. 3.8.1, it is reasonable to examine the rotation excitation to reveal the behind reason. Fig. 3.6 depicts the norm of the angular velocity difference. $imu2$ has more sufficient rotation excitation, improving the observability of the translation calibration parameter.

Environments without target

To eliminate the impact of the calibration target on the accuracy of the target-less method, we conduct experiments on the sequence $\{room1 \sim room6\}$ without calibration target. The target-based method can not work at this setting.

The calibration results of the target-less method are shown in Tab. 3.2. Compared with the sequence with calibration target (see Tab. 3.1), the estimation of the calibration parame-

Table 3.2: Average RMSE (L / R) of the calibration results over 50 Monte-Carlo trials. L: left camera. R: right camera. The units for rotation, translation and time offset are in deg, cm and ms.

Sequence	Rotation	Translation	Time offset
room1	0.033 / 0.056	0.681 / 0.584	0.101 / 0.073
room2	0.136 / 0.136	0.860 / 0.758	0.957 / 0.930
room3	0.036 / 0.057	0.657 / 0.550	1.298 / 1.264
room4	0.042 / 0.043	0.315 / 0.385	0.633 / 0.588
room5	0.033 / 0.067	0.566 / 0.484	0.398 / 0.411
room6	0.161 / 0.180	0.765 / 0.708	0.601 / 0.696

ter does not incur loss of performance without the calibration target in the field of view. The calibration accuracy is still impacted by the trajectory itself. For example, the estimation accuracy of the translation calibration parameter of $room4$ is better than that of $room2$. Fig. 3.6 shows that $room4$ has more sufficient rotation excitation.

For all the results presented so far, the spatial-temporal parameters are assumed to be constant, which is also the most common scenario in practice. Considering the vibration or morphology change of the robot platform [78] and clock drift during the running, it is also worth investigating the calibration of time-varying spatial-temporal parameters, a more challenge scenario. $room4$ is used here for test. To construct time-varying spatial parameters, the global pose measurements are perturbed. ${}^{M'}_{M}T$ is the designed perturbation. The spatial parameters are changed accordingly.

$$ {}^{M'}_{G}T = {}^{M'}_{M}T\,{}^{M}_{G}T \qquad {}^{M'}_{C}T = {}^{M'}_{M}T\,{}^{M}_{C}T \tag{3.23} $$

The time-vary temporal parameter is constructed more straightforward by changing the timestamps of the global pose measurements with designed time-vary values.

The target-based method can not work as it includes constant calibration parameters in state vector. And the requirement of facing the calibration target makes it impractical during the large change of calibration parameters. While EKF-based target-less method could handle dynamic change of state naturally, even without the prior knowledge about such change. As shown in Fig. 3.7, the time-varying quantity of spatial-temporal parameter is designed to

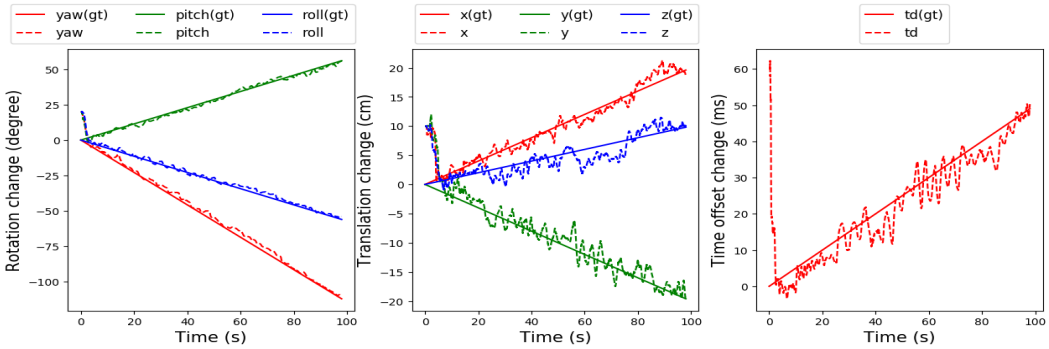


Figure 3.7: Groundtruth (solid lines) and estimation (dashed lines) of the time-varying change of the spatial-temporal parameters.

change linearly with time. The initial rotation and translation errors along each axis are $20°$ and 10 cm respectively. The initial time offset error is 60 ms. Despite the significant estimation errors at the beginning, the target-less method could quickly converge to the groundtruth value and accurately track the time-varying change. After 10s, the average tracking RMSE of the rotation change, the translation change and the time offset change are $1.754°$, 1.346 cm and 4.151 ms respectively. Once dynamic change stage is over, these small errors mean that good initial guess is provided for follow-up constant parameters calibration.

3.9 Conclusion

In this work, we propose two novel calibration methods to estimate the spatial-temporal parameters between the camera and the global pose sensor. One is a target-based method, it adopts offline full-batch nonlinear least squares optimization. Another is a target-less method based on an online EKF estimator. The observability analysis of the target-less method shows that the calibration parameters are observable when the system is fully excited by 6DoF movements. Real-world experiments demonstrate both methods provide accurate and reliable calibration results when traditional hand-eye calibration fails to work. Moreover, the ability of capturing time-varying parameters, rarely studied in literature, is verified successfully for the target-less method. Proposed methods can be easily extended to other global pose sensors besides motion capture system, and different camera models. In

the future, we plan to improve the accuracy of the target-less method using sliding window optimization.

Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor

Supplementary Material

3.10 Analytical on-manifold Jacobians for the target-based method

The optimization function (Eq. (3.3)) contains two types of measurement residual, namely pixel measurement residual and global pose measurement residual. The Jacobians of these residuals with respect to the optimization variables are provided here. On-manifold formulation of the optimization variables, like SE(3) transformations, allows us to easily calculate analytical Jacobian which is more accurate and computational efficient than numerical differentiation.

3.10.1 Jacobians of pixel measurement residual

Firstly, we analyze the Jacobians involved in the pixel measurement residual r_{ij}:

$$\begin{aligned} r_{ij} &= \pi\left({}^{C_i}p_{f_j}, \varsigma\right) - u_{ij} \\ {}^{C_i}p_{f_j} &= {}^{C_i}_{W}T\,{}^{W}p_{f_j} \end{aligned} \tag{3.24}$$

The subset of optimization variables related to r_{ij} is noted as:

$$\chi_{s_1} = \left\{ \begin{array}{cc} {}^{W}_{C_i}T & \varsigma \end{array} \right\} \tag{3.25}$$

The Jacobians of the pixel residual r_{ij} with respect to the 3D point in camera frame ${}^{C_i}p_{f_j}$ and the camera intrinsic ς are $\frac{\partial r_{ij}}{\partial {}^{C_i}p_{f_j}}$ and $\frac{\partial r_{ij}}{\partial \varsigma}$ respectively. Both are determined by the camera projection model [66, 67]. The Jacobian of the pixel residual r_{ij} with respect to the

camera pose ${}^{W}_{C_i}T$ is:

$$\begin{aligned}
\frac{\partial r_{ij}}{\partial {}^{W}_{C_i}T} &= \frac{\partial r_{ij}}{\partial {}^{C_i}p_{f_j}} \frac{\partial {}^{C_i}p_{f_j}}{\partial {}^{C_i}_{W}T} \frac{\partial {}^{C_i}_{W}T}{\partial {}^{W}_{C_i}T} \\
\frac{\partial {}^{C_i}p_{f_j}}{\partial {}^{C_i}_{W}T} &= \left({}^{C_i}_{W}T \, {}^{W}p_{f_j} \right)^{\odot} \\
\frac{\partial {}^{C_i}_{W}T}{\partial {}^{W}_{C_i}T} &= -I
\end{aligned} \tag{3.26}$$

Where $\odot$ is an operator for the homogeneous coordinate [73, Sec. 7.1.8].

In summary, the Jacobians of the pixel measurement residual r_{ij} with respect to χ_{s_1} can be computed via Eq. (3.26) and $\frac{\partial r_{ij}}{\partial \varsigma}$.

3.10.2 Jacobians of global pose measurement residual

Next, we analyze the Jacobians involved in the global pose measurement residual r_{gi} (Eq. (3.3)). To simplify the description, we define the following intermediate quantities:

$$\begin{aligned}
{}^{M}_{G}\hat{T} &\triangleq {}^{M}_{G}T \left(t_i + t_d \right) \\
{}^{W}_{C}T &\triangleq {}^{W}_{C_i}T \\
{}^{M_b}_{M_a}\theta &\triangleq Log \left({}^{M_b}_{G}T \, {}^{M_a}_{G}T^{-1} \right)
\end{aligned} \tag{3.27}$$

Therefore

$$\begin{aligned}
r_{gi} &= Log \left({}^{M}_{G}\hat{T} \, {}^{G}_{W}T \, {}^{W}_{C}T \, {}^{C}_{M}T \right) \\
{}^{M}_{G}\hat{T} &= Exp \left(\lambda {}^{M_b}_{M_a}\theta \right) {}^{M_a}_{G}T \\
\lambda &= (t_i + t_d - t_a)/(t_b - t_a)
\end{aligned} \tag{3.28}$$

The subset of optimization variables related to r_{gi} is noted as:

$$\chi_{s_2} = \left\{ \begin{array}{cccc} {}^{W}_{C_i}T & {}^{G}_{W}T & {}^{C}_{M}T & t_d \end{array} \right\} \tag{3.29}$$

The Jacobian of r_{gi} with respect to ${}^{C}_{M}T$ is:

$$\frac{\partial r_{gi}}{\partial {}^{C}_{M}T} = J_r^{-1} \left(r_{gi} \right) \tag{3.30}$$

Where $J_r \left(\bullet \right)$ is the right Jacobian of SE(3) [73].

The Jacobian of r_{gi} with respect to ${}^{W}_{C}T$ is:

$$\frac{\partial r_{gi}}{\partial {}^{W}_{C}T} = J_r^{-1}\left(r_{gi}\right) Ad\left({}^{C}_{M}T^{-1}\right) \tag{3.31}$$

Where $Ad\left(\bullet\right)$ is the adjoint of SE(3) [73].

The Jacobian of r_{gi} with respect to ${}^{G}_{W}T$ is:

$$\frac{\partial r_{gi}}{\partial {}^{G}_{W}T} = J_r^{-1}\left(r_{gi}\right) Ad\left(\left({}^{W}_{C}T\,{}^{C}_{M}T\right)^{-1}\right) \tag{3.32}$$

The Jacobian of r_{gi} with respect to ${}^{M}_{G}\hat{T}$ is:

$$\frac{\partial r_{gi}}{\partial {}^{M}_{G}\hat{T}} = J_r^{-1}\left(r_{gi}\right) Ad\left(\left({}^{M}_{G}\hat{T}\,{}^{G}_{W}T\,{}^{W}_{C}T\,{}^{C}_{M}T\right)^{-1}\right) \tag{3.33}$$

The Jacobian of ${}^{M}_{G}\hat{T}$ with respect to λ is:

$$\frac{\partial {}^{M}_{G}\hat{T}}{\partial \lambda} = Ad\left(Exp\left(\lambda {}^{M_b}_{M_a}\theta\right)\right) J_r\left(\lambda {}^{M_b}_{M_a}\theta\right) {}^{M_b}_{M_a}\theta \tag{3.34}$$

The Jacobian of λ with respect to t_d is:

$$\frac{\partial \lambda}{\partial t_d} = \frac{1}{t_b - t_a} \tag{3.35}$$

Finally, through the chain rule, the Jacobian of r_{gi} with respect to t_d is calculated as:

$$\frac{\partial r_{gi}}{\partial t_d} = \frac{\partial r_{gi}}{\partial {}^{M}_{G}\hat{T}} \frac{\partial {}^{M}_{G}\hat{T}}{\partial \lambda} \frac{\partial \lambda}{\partial t_d} \tag{3.36}$$

In summary, the Jacobians of the global pose measurement residual r_{gi} with respect to χ_{s_2} can be computed via Eq. (3.30), Eq. (3.31), Eq. (3.32) and Eq. (3.36).

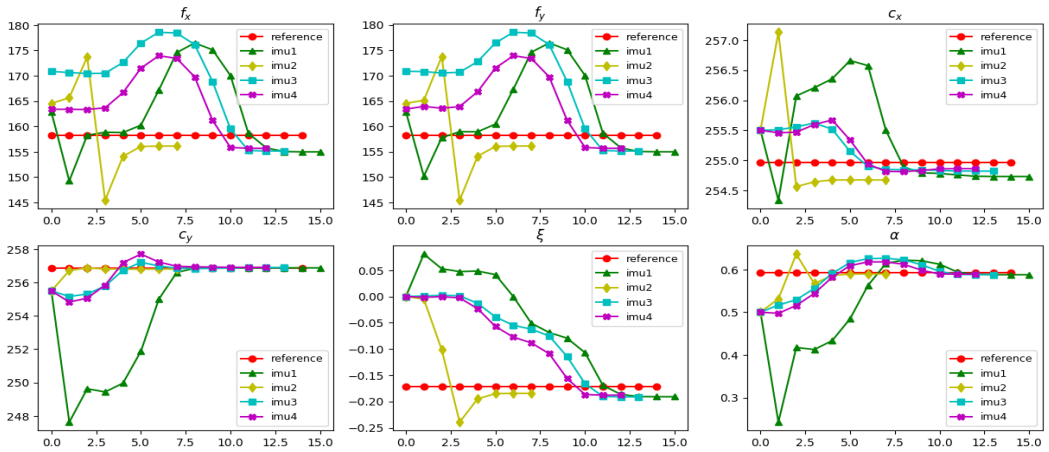


Figure 3.8: Iterative process of calibrating left camera intrinsic from scratch. x-axis represents iteration steps.

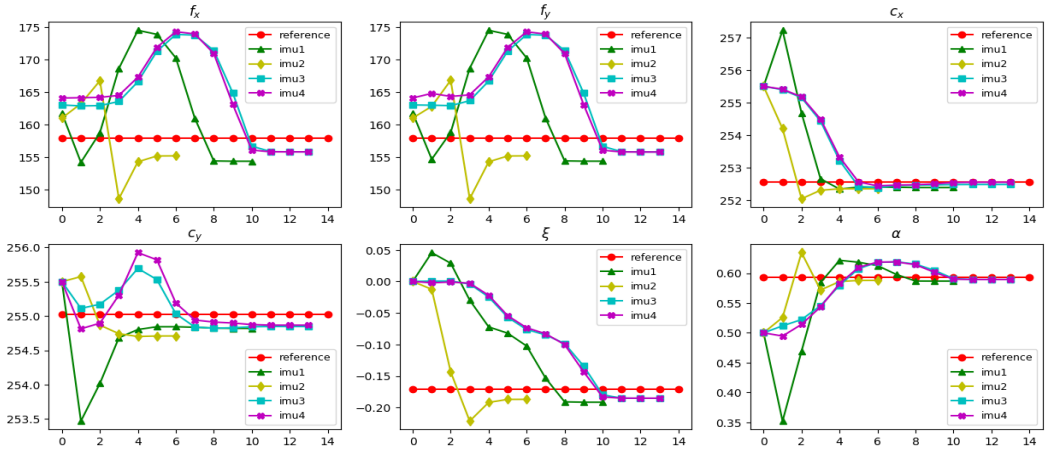


Figure 3.9: Iterative process of calibrating right camera intrinsic from scratch. x-axis represents iteration steps.

3.11 Additional comparison results

Compared to [33], our proposed target-based method has another benefit, in addition to iterative optimization of camera trajectory. Prior to perform spatial-temporal hand-eye calibration, [33] need to calibrate the camera intrinsic first. While our method does not require this step, as camera intrinsic is added to the optimization variables. This simultaneously calibration feature simplifies the calibration process. Moreover, [33] may suffer from the fixed camera intrinsic. Environmental influences and camera motions may lead to unmodelled errors for camera intrinsic. To address this issue, our method finds the optimal camera intrinsic parameters that best fit all available measurements for each sequence.

Fig. 3.8 shows the iterative process of calibrating monocular camera intrinsic from scratch with our target-based method. Left camera is used for the selected sequence $\{imu1 \sim imu4\}$ from TUM-VI Dataset [61], and double sphere camera model [66] is adopted. Regarding the initialization method and reference values for camera intrinsic parameters, we refer to [66]. In Fig. 3.8, all estimated intrinsic parameters converge near the reference values, with slightly difference for each sequence. When using the right camera, the corresponding results are shown in Fig. 3.9. Final average reprojection error and position error in Eq. (3.3) are smaller than 0.1 pixel and 0.1 cm for left and right camera from each sequence. Results from Fig. 3.8 and Fig. 3.9 demonstrate the ability of calibrating optimal camera intrinsic from scratch for each sequence with the target-based method.

Chapter 4

GPS-VIO Fusion with Online Rotational Calibration

Accurate global localization is crucial for autonomous navigation and planning. To this end, various GPS-aided Visual-Inertial Odometry (GPS-VIO) fusion algorithms are proposed in the literature. This paper presents a novel GPS-VIO system that is able to significantly benefit from the online calibration of the rotational extrinsic parameter between the GPS reference frame and the VIO reference frame. The behind reason is this parameter is observable. This paper provides novel proof through nonlinear observability analysis. We also evaluate the proposed algorithm extensively on diverse platforms, including flying UAV and driving vehicle. The experimental results support the observability analysis and show increased localization accuracy in comparison to state-of-the-art (SOTA) tightly-coupled algorithms.

4.1 Related paper

- Song, Junlin, Pedro J. Sanchez-Cuevas, Antoine Richard, Raj Thilak Rajan, and Miguel Olivares-Mendez. "GPS-VIO Fusion with Online Rotational Calibration." 2024 IEEE International Conference on Robotics and Automation (ICRA 2024).

4.2 Relationships to other chapters

Similarly to the principle of target-less method presented in Chapter 3, an online calibration approach is developed in this chapter to estimate the rotational extrinsic parameter between the GPS reference frame and the VIO reference frame. It is noted that these two calibration tasks have a key difference. The spatial calibration parameter in Chapter 3 is assumed to be rigid and is independent of where the system is initialized. However, in this chapter, the calibration parameter can be varied for each run, depending on where VIO is initialized. Here, we assume the spatial-temporal parameters for IMU-Camera system are known. The online and offline calibration for the IMU-Camera system are discussed in Chapter 5 and Chapter 6.

4.3 Introduction

The accuracy, robustness and reliability of the pose estimation are essential for the safe autonomous navigation of mobile robots. In the past few decades, the Global Positioning System (GPS) has been widely used for localization in outdoor scenes, since it offers a robust global localization solution without accumulating drift over time. However, due to the high-level noise of consumer grade GPS sensors, accurate positioning cannot be typically achieved by using GPS sensors alone. Moreover, in urban scenes, GPS signals are vulnerable to the interference of the local environment, such as signal occlusions, or bounces due to high-rise buildings, which further degrades the GPS positioning performance. In these GPS-degraded or -denied scenarios, Visual-Inertial Odometry (VIO) using IMU(Intertial Measurement Unit) and Camera(s), and Simultaneous Localization and Mapping (SLAM) algorithms are conventionally implemented.

VIO algorithms do not suffer from these interruptions, and provide high-precision and high-frequency local state estimation, however, these algorithms also have their inherent drawbacks. For instance, VIO systems cannot provide long-term drift-free localization and heading. In [23], the authors prove that VIO systems have four unobservable Degrees of

Freedom (DoF), namely the 3D positions and yaw. SLAM mitigates this drawback using simultaneous estimation of the localization, the map, along the execution of the loop-closure, and map alignment. These mechanisms allow SLAM techniques to decrease the localization uncertainty and the long-term drift. Unfortunately, SLAM techniques demand high computational and memory resources, which limit their applicability. In general, GPS positioning and visual inertial navigation system can be combined to provide an accurate, high frequency, robust localization and long-term drift-free localization. The fusion of the three sensors involved in this process, GPS, camera and IMU has produced promising results and can achieve locally accurate and globally drift-free localization.

GPS-aided VIO algorithms have been previously proposed in the literature [48, 47]. The spatial transformation to couple both the GPS and the VIO reference frame is shown to be unobservable with linear observability analysis [48, 79]. However, linearization implies that the derived observability is local and may be unreliable for the original nonlinear system [80, 49]. Thus, it is necessary to revisit the observable property with the tool of nonlinear observability analysis. More specifically, we aim to show in this paper that the rotational extrinsic parameter between GPS reference frame and VIO reference frame is observable.

In this paper, we propose a novel filter-based GPS-VIO system which is specially focused on including a reliable and accurate estimation of the rotation extrinsic between the GPS reference frame and the VIO reference frame. Our key contributions are summarized as:

- We propose a novel filter-based estimator to fuse GPS measurements and visual-inertial data, and simultaneously estimate the rotational extrinsic parameter between the GPS and VIO frames online.
- We prove that the rotational extrinsic parameter is observable via nonlinear observability analysis, and support our conclusion with simulations.
- We evaluate the localization accuracy of the proposed algorithm on multiple public datasets, including small scale flying datasets and large scale driving datasets, and show the superior performance of our algorithm.

4.4 Related work

Sensor fusion of camera and IMU is a well studied topic [37, 6]. Visual-inertial fusion algorithms can be broadly classified into two categories i.e., optimization-based methods and filter-based methods. Optimization-based methods achieve higher theoretical accuracy, which include VINS-Mono [38], Basalt [39] and ORB-SLAM3 [40]. Their high computational cost is a major disadvantage. In contrast, sliding-window filter-based methods, such as the Multi-State Constraint Kalman Filter (MSCKF) [41, 42, 16], are more resource efficient and achieve comparable accuracy.

The combination of a camera and an IMU can only generate relative pose estimation, resulting in the unobservability of global position and absolute yaw [23]. Therefore, pure VIO systems tend to drift over time [43]. Recent works have employed GPS measurement to eliminate this drift. These methods can be divided into loosely-coupled methods and tightly-coupled methods. VINS-Fusion is a loosely-coupled approach, which fuses GPS position measurements and output pose of VIO subsystem [44]. However, the fusion algorithm is unable to improve the VIO subsystem. Therefore, the inner correlations of all measurements are discarded, causing suboptimal localization results. Gomsf is a similar loosely-coupled work [45].

Tightly-coupled methods fully exploit the complementary merits of multi-sensor data, and are promising to further improve the accuracy. A tightly-coupled estimator based on sliding window optimization is proposed in [46]. The rotation between the GPS reference frame and the VIO reference frame is included in the state vector, but the non-synchronization between GPS timestamp and VIO system timestamp is neglected. [47] describes another tightly-coupled optimization-based approach. The comparative experiments with VINS-Fusion have demonstrated that tightly-coupled methods are superior to loosely-coupled methods. However, the transformation between GPS reference frame and VIO reference frame is not estimated in [47]. The closest to our work is [48], which is a tightly-coupled estimator based on MSCKF. The extrinsic parameters between the GPS reference frame and the VIO reference frame are inserted into the state during initialization, however, marginalized after all states

are transformed from the VIO reference frame to the GPS reference frame [48].

Consequently, the approach of [48] does not estimate the extrinsic parameters between GPS-VIO online, as they show the extrinsic parameters are unobservable with linear observability analysis. However, linear observability analysis maybe unreliable for a nonlinear system. A locally observable system is sure to be globally observable, but a locally unobservable system maybe globally observable [49, 50, 51]. Our main contribution in this work is to point out that rotational extrinsic parameter is globally observable using nonlinear observability analysis. This novel observability conclusion is similar to our recent accepted work [52], termed as GPS-VWO. The difference between GPS-VIO and GPS-VWO lies in the different kinematic equations, with the former driven by IMU while the latter driven by wheel odometer. As the reference frame of VIO is gravity aligned, the rotational extrinsic parameter of GPS-VIO only has yaw component. Unlike GPS-VIO, the rotational extrinsic parameter of GPS-VWO is 3DoF in general. To analyze the observability of extrinsic parameter, Lie derivative is employed for GPS-VIO, like GPS-VWO [52], considering the nonlinearity of this system.

The unavoidable errors caused by imposing fixed extrinsic parameters after GPS-VIO initialization lead to miss-calculations of the fusion algorithms in long distances. Without online calibration, the estimation error of rotational extrinsic parameter at the start will deteriorate the localization accuracy, especially when the GPS noise is relatively large. [46, 53, 54] adopt explicitly online calibration of the rotational extrinsic parameter to improve localization accuracy. To simplify the state estimation complexity, [54] disables the online estimation once the rotational extrinsic parameter is converged. However, neither of them provide a theoretical observability analysis. In this paper, we prove the rotational extrinsic parameter is observable; hence, including it in the state vector is a promising and theoretically guaranteed mean to improve the accuracy of the state estimator.

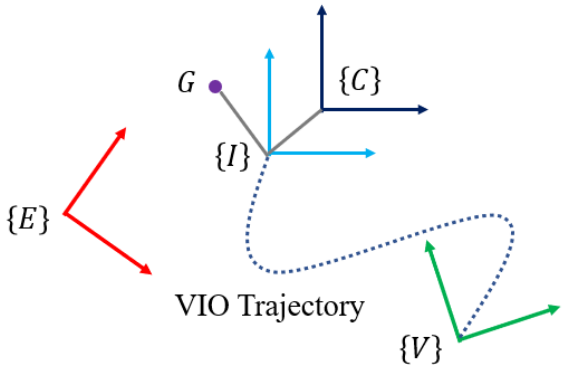


Figure 4.1: Coordinate systems, similar as Fig. 1a in [52].

4.5 Problem Formulation

4.5.1 Reference frames and Notation

The coordinate systems used in this work are shown in the Fig. 4.1. $\{E\}$ represents the East-North-Up (ENU) coordinate system, which is the reference frame of GPS position measurements. An arbitrary GPS measurement can be chosen as the origin of $\{E\}$ frame. $\{V\}$ is the reference frame of the VIO system. After the initialization of the VIO system, the orientation of this coordinate system is gravity aligned. $\{I\}$ and $\{C\}$ represent IMU coordinate frame and camera coordinate system, respectively. G is the position of the antenna of the GPS receiver.

We use the notation ${}^{V}(\bullet)$ to represent a quantity in the coordinate frame $\{V\}$. The position of point I in the frame $\{V\}$ is expressed as ${}^{V}p_I$. The velocity of frame $\{I\}$ in the frame $\{V\}$ is expressed as ${}^{V}v_I$. Furthermore, we use quaternion to represent the attitude of rigid body [63]. ${}^{I}_{V}q$ represents the orientation of frame $\{I\}$ with respect to frame $\{V\}$, and its corresponding rotation matrix is given by ${}^{I}_{V}R$. Similar notations also apply to the other reference frames. $[\bullet]_\times$ denotes the skew symmetric matrix corresponding to a three-dimensional vector, and $[\bullet]^T$ is used to represent the transpose of a matrix.

4.5.2 Classical MSCKF-based VIO structure

According to [48, 16], the classic MSCKF-based VIO algorithm usually defines the following states

$$\begin{aligned} x &= \begin{bmatrix} x_I^T & x_{c_1}^T & \cdots & x_{c_N}^T \end{bmatrix}^T \\ x_I &= \begin{bmatrix} {}^I_V q^T & {}^V p_I^T & {}^V v_I^T & {}^V p_f^T & b_\omega^T & b_a^T \end{bmatrix}^T \\ x_{c_i} &= \begin{bmatrix} {}^{I_i}_V q^T & {}^V p_{I_i}^T \end{bmatrix}^T \end{aligned} \tag{4.1}$$

where ${}^V p_f$ represents feature position, f, in the VIO reference frame $\{V\}$. To make the presentation concise, only one feature point is described here. b_ω and b_a are the biases of the IMU angular velocity and the linear acceleration measurements, respectively. x_I indicates the current IMU state. x_{c_i} is obtained by extracting the first two quantities of x_I at different image times. Then the state of the whole MSCKF system x can be constructed by augmenting N historical x_{c_i} in x_I.

After successful initialization and setting appropriate initial value and covariance for x, the VIO system follows the Kalman filter pipeline. IMU measurements are used for the propagation of x_I. Whenever a new image is received, x is augmented with the pose clone of the current x_I and the visual constraints between multiple pose clones are utilized to update the state. For more details of this part, we refer interested readers to Open-VINS [16].

4.5.3 GPS Measurement Update

Assuming that the first GPS position measurement as the origin of the $\{E\}$ frame, the subsequent GPS measurements are denoted as ${}^E p_G$. Each GPS observation can be formulated as[1]

$$\begin{aligned} z &= {}^E p_G + n_{gps} = {}^E p_V + {}^E_V R {}^V p_G + n_{gps} \\ &= {}^E p_V + {}^E_V R \left({}^V p_I + {}^I_V R^T {}^I p_G \right) + n_{gps} \end{aligned} \tag{4.2}$$

[1]Measurement equation (4.2) is used here just for the convenience of the observability analysis. In the implementation, we adopt the interpolation measurement equation as [48].

where n_{gps} is a white Gaussian noise. ${}^{I}p_G$ is the position of point G in the IMU frame $\{I\}$. This paper assumes that this quantity is known, since ${}^{I}p_G$ can be obtained from CAD model or calibrated before the system runs. ${}^{I}_{V}R$ and ${}^{V}p_I$ are quantities expressed in the VIO reference frame $\{V\}$. ${}^{E}p_V$ and ${}^{E}_{V}R$ are the transformations between frame $\{V\}$ and frame $\{E\}$. Since these two frames are gravity aligned, we can simply use the yaw angle to parameterize the rotation matrix between them. Therefore, ${}^{E}_{V}R$ can be expressed as

$$ {}^{E}_{V}R = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.3} $$

where ψ is the relative yaw angle between GPS reference frame and VIO reference frame.

To make the measurement equation usable, ψ and ${}^{E}p_V$ must be known. In [48], these are calculated in the initialization stage of the GPS-VIO system and are marginalized later. The main difference between our work and theirs is that we will provide a more suitable nonlinear observability analysis in Section 4.6.2, to decide the inclusion of observable quantities to the system state vector for potential online refinement.

To analyze the observability of all the extrinsic parameters between the frame $\{V\}$ and the frame $\{E\}$, ψ and ${}^{E}p_V$ are included in the state vector. Moreover, the time offset between the GPS and IMU should also be modeled for the sake of real world experiments. Thus, the new system state vector then becomes

$$ x = \begin{bmatrix} x_I^T & x_{c_1}^T & \cdots & x_{c_N}^T & \psi & {}^{E}p_V^T & {}^{I}t_G \end{bmatrix}^T \tag{4.4} $$

where ψ and ${}^{E}p_V$ represent the interested extrinsic parameters, and ${}^{I}t_G$, is the time offset between the GPS and IMU clock[2]. The subset of system state related to the GPS measurement equation is noted as x_s

$$ x_s = \begin{bmatrix} {}^{I}_{V}q^T & {}^{V}p_I^T & \psi & {}^{E}p_V^T \end{bmatrix}^T \tag{4.5} $$

[2]As time offset ${}^{I}t_G$ calibration is not the focus of this work, it is ignored in following analysis, but considered in real-world experiments to compensate non-synchronization (Section 4.7.3).

The measurement Jacobian H is expressed as

$$H = \frac{\partial \tilde{z}}{\partial \tilde{x}_s} = \begin{bmatrix} -{}^E_V R^I_V R^T \left[{}^I p_G\right]_\times & {}^E_V R & H_\psi {}^V p_G & I_3 \end{bmatrix} \tag{4.6}$$

$$H_\psi = \begin{bmatrix} -\sin\psi & -\cos\psi & 0 \\ \cos\psi & -\sin\psi & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.7}$$

To keep this paper focused and concise, we omit the description of the GPS-VIO system's initialization, as well as the proper handling of time-offsets among the different sensors, like ${}^I t_G$. Interested readers can refer to [48].

4.6 Observability Analysis

4.6.1 Comments on Linear Observability Analysis

The linear observability analysis of the GPS-VIO system has been investigated previously in [48] and detailed in [79]. However, the unobservable property obtained about extrinsic parameters in these works may be misleading, as they apply linear observability analysis for a typically nonlinear GPS-VIO system. As discussed in Section 4.4, a locally unobservable system maybe globally observable [49, 50, 51]. Moreover, no experiments were performed in [48, 79] to validate the observability conclusions regarding the extrinsic parameters. In this work, we employ a more appropriate nonlinear observability analysis for the nonlinear GPS-VIO system to obtain observable property and provide experiments to solidify the analysis.

4.6.2 Nonlinear Observability Analysis

We now conduct a nonlinear observability analysis, following the standard Lie derivatives method [51]. Removing the IMU bias and pose clones from the state vector (4.4), to simplify

the formulation, the state becomes

$$x = \begin{bmatrix} {}^{I}_{V}q^{T} & {}^{V}v_{I}^{T} & {}^{V}p_{I}^{T} & {}^{V}p_{f}^{T} & \psi & {}^{E}p_{V}^{T} \end{bmatrix} \tag{4.8}$$

Following immediately, we write the kinematic equations as

$$\begin{bmatrix} {}^{I}_{V}\dot{q} \\ {}^{V}\dot{v}_{I} \\ {}^{V}\dot{p}_{I} \\ {}^{V}\dot{p}_{f} \\ \dot{\psi} \\ {}^{E}\dot{p}_{V} \end{bmatrix} = \underbrace{\begin{bmatrix} 0_{4\times1} \\ g \\ {}^{V}v_{I} \\ 0_{3\times1} \\ 0 \\ 0_{3\times1} \end{bmatrix}}_{f_0} + \underbrace{\begin{bmatrix} \frac{1}{2}\Xi\left({}^{I}_{V}q\right) \\ 0_{3} \\ 0_{3} \\ 0_{3} \\ 0_{1\times3} \\ 0_{3} \end{bmatrix}}_{f_1} \omega + \underbrace{\begin{bmatrix} 0_{4\times3} \\ {}^{I}_{V}R^{T} \\ 0_{3} \\ 0_{3} \\ 0_{1\times3} \\ 0_{3} \end{bmatrix}}_{f_2} a \tag{4.9}$$

where g is denoted as the local gravity vector. ω and a are de-biased IMU angular velocity and linear acceleration measurements, respectively. Here, we use the time derivative property of quaternions

$$\dot{q} = \frac{1}{2}\Omega\left(\omega\right)q = \frac{1}{2}\Xi\left(q\right)\omega \tag{4.10}$$

The definition of $\Xi\left(q\right)$ can be found in [63].

Next, we list the usable measurement equations. The camera measurement equation is

$$h_1\left(x\right) = {}^{C}p_f = {}^{C}_{I}R^{I}_{V}Rp + {}^{C}p_I \tag{4.11}$$

where $p = {}^{V}p_f - {}^{V}p_I$. The norm constraint of the unit quaternion is also considered as a measurement equation

$$h_2\left(x\right) = {}^{I}_{V}q^{T}{}^{I}_{V}q - 1 = 0 \tag{4.12}$$

The measurement equation of the GPS is

$$h_3\left(x\right) = {}^{E}p_G = {}^{E}p_V + {}^{E}_{V}R^{V}p_I \tag{4.13}$$

where without loss of generality, we assume ${}^{I}p_G = 0_{3\times1}$ to simplify the expression.

Zeroth-Order Lie Derivatives

The zeroth-order Lie derivative of a function is itself.

$$\begin{aligned}
\pounds^0 h_1 &= {}^{C}p_I + {}^{C}_{I}R^{I}_{V}Rp \\
\pounds^0 h_2 &= {}^{I}_{V}q^{T}\,{}^{I}_{V}q - 1 \\
\pounds^0 h_3 &= {}^{E}p_V + {}^{E}_{V}R^{V}p_I
\end{aligned} \tag{4.14}$$

The gradients of zeroth-order Lie derivatives with respect to x are

$$\begin{aligned}
\nabla\pounds^0 h_1 &= \begin{bmatrix} X_1 & 0_3 & -{}^{C}_{I}R^{I}_{V}R & {}^{C}_{I}R^{I}_{V}R & 0_{3\times1} & 0_3 \end{bmatrix} \\
\nabla\pounds^0 h_2 &= \begin{bmatrix} 2\,{}^{I}_{V}q^{T} & 0_{1\times3} & 0_{1\times3} & 0_{1\times3} & 0 & 0_{1\times3} \end{bmatrix} \\
\nabla\pounds^0 h_3 &= \begin{bmatrix} 0_{3\times4} & 0_3 & {}^{E}_{V}R & 0_3 & H_{\psi}{}^{V}p_I & I_3 \end{bmatrix}
\end{aligned} \tag{4.15}$$

where X represents a quantity that does not need to be computed explicitly, as it does not affect the observability analysis.

First-Order Lie Derivatives

The first-order Lie derivative of h_1 with respect to f_0 is computed as

$$\pounds^1_{f_0} h_1 = \nabla\pounds^0 h_1 \bullet f_0 = -{}^{C}_{I}R^{I}_{V}R^{V}v_I \tag{4.16}$$

The gradient of $\pounds^1_{f_0} h_1$ with respect to x is

$$\nabla\pounds^1_{f_0} h_1 = \begin{bmatrix} X_2 & -{}^{C}_{I}R^{I}_{V}R & 0_3 & 0_3 & 0_{3\times1} & 0_3 \end{bmatrix} \tag{4.17}$$

The first-order Lie derivative of h_1 with respect to f_1 is computed as

$$\pounds^1_{f_1} h_1 = \nabla\pounds^0 h_1 \bullet f_1 = \frac{1}{2}X_1\Xi\left({}^{I}_{V}q\right) \tag{4.18}$$

where the gradient of $\pounds^1_{f_1} h_1$ with respect to x is

$$\nabla \pounds^1_{f_1} h_1 = \begin{bmatrix} X_3 & 0_{9\times 3} & X_4 & -X_4 & 0_{9\times 1} & 0_{9\times 3} \end{bmatrix} \tag{4.19}$$

The first-order Lie derivative of h_3 with respect to f_0 is computed as

$$\pounds^1_{f_0} h_3 = \nabla \pounds^0 h_3 \bullet f_0 = {}^E_V R^V v_I \tag{4.20}$$

and the gradient of $\pounds^1_{f_0} h_3$ with respect to x is

$$\nabla \pounds^1_{f_0} h_3 = \begin{bmatrix} 0_{3\times 4} & {}^E_V R & 0_3 & 0_3 & H_\psi{}^V v_I & 0_3 \end{bmatrix} \tag{4.21}$$

Observability analysis

By stacking the gradients of previously calculated Lie derivatives together, the following observability matrix is constructed

$$\mathcal{O} = \begin{bmatrix} \nabla \pounds^0 h_1 \\ \nabla \pounds^0 h_2 \\ \nabla \pounds^0 h_3 \\ \nabla \pounds^1_{f_0} h_1 \\ \nabla \pounds^1_{f_1} h_1 \\ \nabla \pounds^1_{f_0} h_3 \end{bmatrix} = \begin{bmatrix} X_1 & 0_3 & -{}^C_V R & {}^C_V R & 0_{3\times 1} & 0_3 \\ 2^I_V q^T & 0_{1\times 3} & 0_{1\times 3} & 0_{1\times 3} & 0 & 0_{1\times 3} \\ 0_{3\times 4} & 0_3 & {}^E_V R & 0_3 & H_\psi{}^V p_I & I_3 \\ X_2 & -{}^C_V R & 0_3 & 0_3 & 0_{3\times 1} & 0_3 \\ X_3 & 0_{9\times 3} & X_4 & -X_4 & 0_{9\times 1} & 0_{9\times 3} \\ 0_{3\times 4} & {}^E_V R & 0_3 & 0_3 & H_\psi{}^V v_I & 0_3 \end{bmatrix} \tag{4.22}$$

Adding the fourth column to the third column, $\mathcal{O}$ becomes

$$\mathcal{O} = \begin{bmatrix} X_1 & 0_3 & 0_3 & {}^{C}_{V}R & 0_{3\times1} & 0_3 \\ 2{}^{I}_{V}q^T & 0_{1\times3} & 0_{1\times3} & 0_{1\times3} & 0 & 0_{1\times3} \\ 0_{3\times4} & 0_3 & {}^{E}_{V}R & 0_3 & {H_\psi}^V p_I & I_3 \\ X_2 & -{}^{C}_{V}R & 0_3 & 0_3 & 0_{3\times1} & 0_3 \\ X_3 & 0_{9\times3} & 0_{9\times3} & -X_4 & 0_{9\times1} & 0_{9\times3} \\ 0_{3\times4} & {}^{E}_{V}R & 0_3 & 0_3 & {H_\psi}^V v_I & 0_3 \end{bmatrix} \tag{4.23}$$

${}^{E}_{V}R$ in the third column can be used to eliminate ${H_\psi}^V p_I$ in the fifth column and I_3 in the sixth column. Thus, $\mathcal{O}$ can be reduced to

$$\mathcal{O} = \begin{bmatrix} X_1 & 0_3 & 0_3 & {}^{C}_{V}R & 0_{3\times1} & 0_3 \\ 2{}^{I}_{V}q^T & 0_{1\times3} & 0_{1\times3} & 0_{1\times3} & 0 & 0_{1\times3} \\ 0_{3\times4} & 0_3 & {}^{E}_{V}R & 0_3 & 0_{3\times1} & 0_3 \\ X_2 & -{}^{C}_{V}R & 0_3 & 0_3 & 0_{3\times1} & 0_3 \\ X_3 & 0_{9\times3} & 0_{9\times3} & -X_4 & 0_{9\times1} & 0_{9\times3} \\ 0_{3\times4} & {}^{E}_{V}R & 0_3 & 0_3 & {H_\psi}^V v_I & 0_3 \end{bmatrix} \tag{4.24}$$

The sixth column corresponds to the translation part of the extrinsic parameter between frame $\{V\}$ and frame $\{E\}$. This column is not full rank, so the translation part is unobservable. Finally, we analyze the rotation part of the extrinsic parameter. Let us focus on ${H_\psi}^V v_I$ in the fifth column. The fifth column cannot be eliminated by other columns and is full rank in general case. So the rotational extrinsic parameter is observable. It is worth noting that the rank of fifth column can become zero if zero velocity motion incurs, more specifically, x and y direction. Therefore, horizontal velocity excitation affects the observability of the rotational extrinsic parameter.

4.7 Results

We develop the proposed algorithm based on Open-VINS [16], which is a state-of-the-art VIO framework. When GPS information is usable, the system state including the rotational extrinsic parameter between the GPS reference frame and the VIO reference frame is updated via Section 4.5.3. Since [48] is not open-sourced, we have implemented our own version by following their paper. In the results presented here, our implementation [48] is referred as **"GPS-VIO-fixed"**. The prefix "fixed" comes from the fact that the spatial transformation is marginalized after the initialization of the system. While our algorithm continues to calibrate the rotational extrinsic parameter online after initialization.

First, we design a simulation environment to verify the observability conclusion. Then, the proposed algorithm is evaluated on two public datasets. One is the small-scale EuRoC dataset [60], which has seen extensive use in the VIO research community. Noisy GPS measurements are simulated by adding Gaussian noise to the groundtruth position. It is featured by UAV flying. Another one is the large-scale KAIST dataset with real GPS measurements in challenging urban scenes [14]. It is featured by vehicular driving. The path length and GPS noise of each selected KAIST sequence is longer than 7km and larger than 6m, respectively.

4.7.1 Validation of the Observability Analysis

To verify our observability proof, we build a simulation environment based on Open-VINS [16]. The groundtruth trajectory of MH_01_easy in EuRoC dataset is used to generate simulated multi-sensor data, including 400Hz IMU, 10Hz image and 10Hz GPS. The noise of the GPS sensor is approximated by applying multivariate Gaussian noises with a standard deviation of 0.2m on the positions.

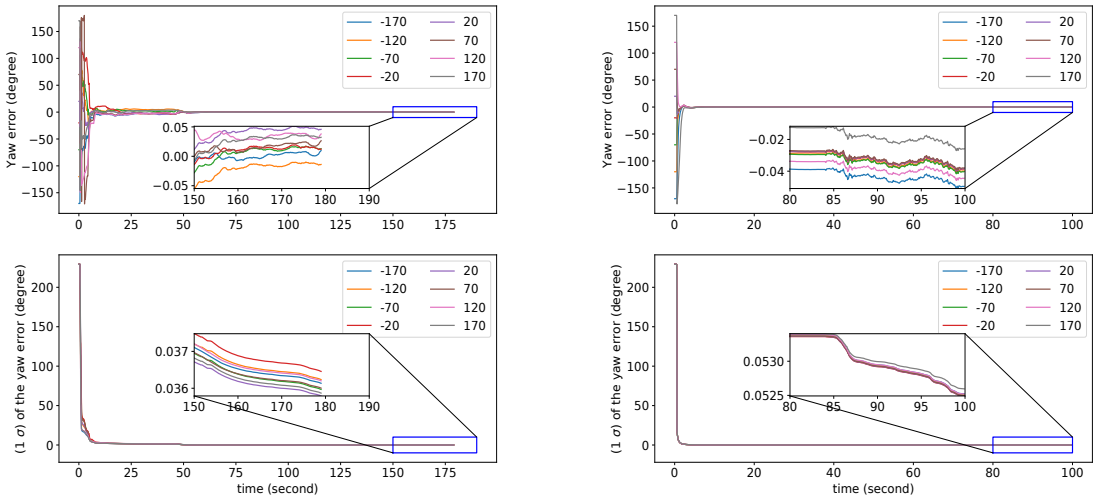
To verify the convergence capability of discovered observable quantity, the calibration of the rotation extrinsic parameter is performed with different initial guesses. We start with an error of 20 degrees and add 50 degrees increment until we reach 170 degrees, we then reiterate with negative angles from -20 to -170 degrees. Fig. 4.2a shows the convergence

of the yaw error. Between 21s to 45s, the convergence of yaw error reaches to steady state because of the stationary motion status. As mentioned in Section 4.6.2, zero velocity motion leads to the unobservability of rotation extrinsic parameter. At other times, there exist velocity excitation. Before 21s, the motion space near the starting point is relatively small compared to GPS noise. After 45s, the moving distance exceeds 10m, which is far greater than GPS noise.

Fig. 4.2a shows one standard deviation (1 σ) of the yaw error. Initial one standard deviation of the yaw error is set to 4 rad, considering the largest initial yaw error is close to π rad. The estimation of the yaw error consistently converges to near zero with small uncertainty, and the convergence process is robust to the relatively large initial error.

Apart from UAV trajectory, we also repeat the above steps with the planer vehicular trajectory of Urban39 in KAIST dataset. Larger GPS noise and practical GPS noise characteristic are considered. The vertical noise is set to twice the horizontal noise. The GPS noise is defined as

$$n_{gps} \sim \mathcal{N}\left(0_{3\times 1}, diag(1,1,4)\right) \tag{4.25}$$



(a) MH_01_easy (b) Urban39

Figure 4.2: Top: ψ convergence over time respect to different initial guesses. Bottom: One standard deviation (1 σ) of ψ.

Fig. 4.2b shows the convergence results with vehicular trajectory. Both yaw error and its corresponding uncertainty consistently converges to near zero, even for the near π rad initial error. All these results from Fig. 4.2 support that the rotational extrinsic parameter is observable.

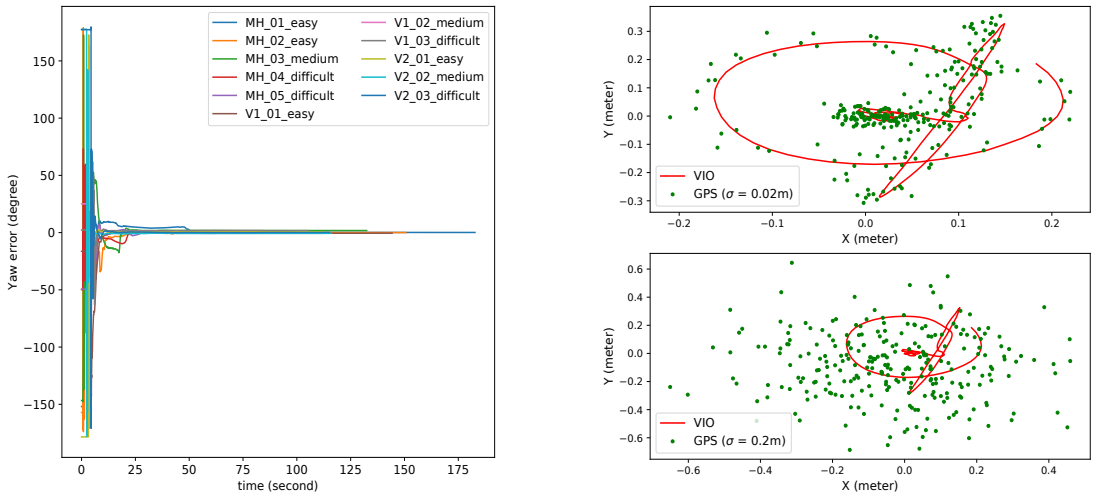
4.7.2 EuRoC dataset

There are 11 sequences in EuRoC dataset. Each sequence is classified into easy, medium or hard according to the level of difficulty for the VIO algorithms. Image and IMU data are available at 20Hz and 200Hz respectively, and the groundtruth position and orientation are provided at 200Hz. We test all sequences to verify the convergence of the rotational extrinsic parameter between the GPS reference frame and the VIO reference frame. Similarly to the previous experiment, the simulated GPS measurements are obtained by adding Gaussian noise ($\sigma = 0.2m$) to the groundtruth position. The GPS frequency is sampled to be 20Hz.

For the initialization of our GPS-VIO system, it is assumed that we do not have any accurate initial estimation for ψ. And the initial value of ψ is naively set as $\hat{\psi} = 0$. ${}^{E}\hat{p}_{V}$ is set as the first GPS measurement received after successful VIO initialization. The groundtruth of ψ is acquired by querying the groundtruth orientation value at the initialization time.

Fig. 4.3a shows the convergence of ψ over time. The range of the initial yaw error is $[-178.47°, 177.67°]$. Estimation error $\left(\hat{\psi} - \psi\right)$ of each sequence approaches to near zero quickly and perfectly. Results verify the observability of ψ.

Table 4.1 shows the Absolute Trajectory Error (ATE) of the different algorithms on all the sequences. We include the results for GPS positioning, optimization-based VIO (SVO2.0 [81]), filter-based VIO (Open-VINS [16]), two variants of the tightly-coupled optimization-based GPS-VIO approach [47], GPS-VIO-fixed [48] and our proposed algorithm. As [47] relies on manually setting initial rotational extrinsic parameter, we provide two variants: initializing ψ as zero as ours, or initializing ψ as groundtruth. Our approach outperforms other state-of-the-art competitors on most sequences because of online rotational calibration. Regarding the first three sequences, we achieve less but close accuracy compared to the second variant of [47]. The possible reason is that the VIO subsystem of [47], SVO2.0 [81],



(a) (b)

Figure 4.3: (a) ψ convergence over time. (b) Horizontal view of aligned trajectory with different level of GPS noise.

performs better than our VIO subsystem, Open-VINS [16], in the first three sequences. However, SVO2.0 suffers from relatively naive VIO initialization strategy [82] for most sequences.

4.7.3 KAIST dataset

The KAIST datasets are collected in highly complex urban environments. It is very challenging to achieve high-precision localization in these environments using consumer grade sensors. Because many moving objects exist in the streets and dense high-rise buildings corrupt GPS signals. Image, IMU and GPS of KAIST datasets are received at 10Hz, 100Hz and 5Hz respectively.

We refer to the initialization algorithm of [48] to obtain the initial ψ and ${}^{E}p_{V}$. The initialization distance is set as 20m. ${}^{E}p_{V}$ is fixed after initialization and only ψ is estimated online. As the groundtruth orientation in the GPS reference frame is unavailable (see Section 4.3 in [14]), $(\psi - \psi_0)$ is plotted in Fig. 4.4a to show the convergence trend over time. ψ_0 is the initial value of ψ. The deviation from the initial value is less than 2.5° for each sequence.

Although the calibration of time offset between GPS and IMU, ${}^{I}t_{G}$, is not the focus of this paper, we still need to deal with it carefully. It has a negative impact on the local-

Table 4.1: ATE (meter) Comparison with the SOTA on the EuRoC Dataset. The ATE of GPS trajectory is 0.347m.

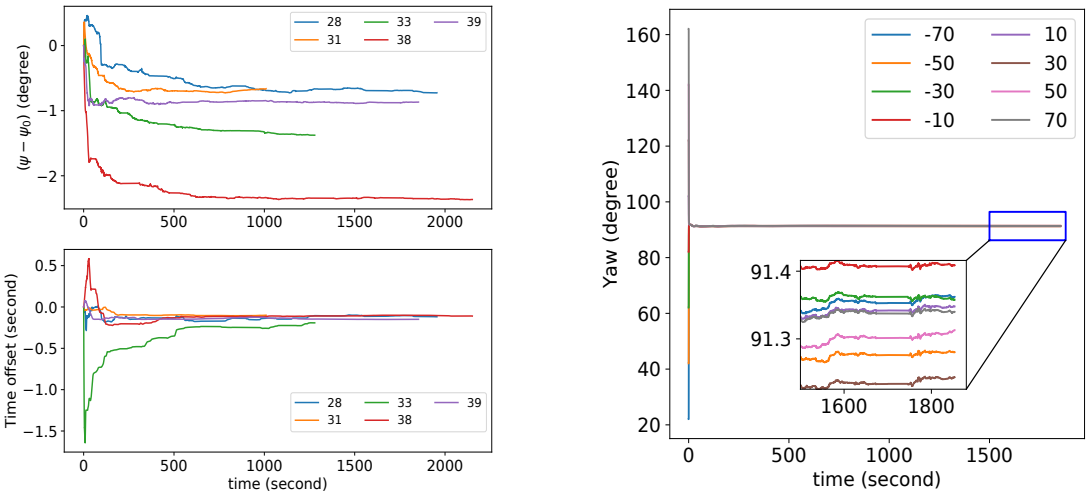
ID	VIO [81]	VIO [16]	A	B	C	Ours
MH01	0.064	0.084	0.137	**0.031**	0.114	0.036
MH02	0.052	0.086	0.110	**0.036**	0.126	0.040
MH03	0.118	0.124	0.119	**0.048**	0.174	0.062
MH04	0.203	0.169	0.292	0.068	0.080	**0.061**
MH05	0.240	0.200	0.312	0.056	0.176	**0.049**
V101	0.064	0.054	0.0	0.041	0.039	**0.037**
V102	0.082	0.046	0.312	0.048	0.050	**0.037**
V103	0.066	0.048	0.365	0.068	0.091	**0.041**
V201	0.085	0.041	0.106	0.038	0.098	**0.035**
V202	0.111	0.040	0.123	0.046	0.042	**0.033**
V203	0.156	0.067	0.154	0.098	0.073	**0.044**

[1] A: results of [47] by initializing ψ as zero.
[2] B: results of [47] by initializing ψ as groundtruth.
[3] C: results of GPS-VIO-fixed [48] by initializing ψ through Section IV in [48], which suffers from relatively large GPS noise (see Fig. 4.3b). The initialization distance is set as 2m.
[4] Ours: results of proposed method by initializing ψ as zero.

ization accuracy without proper handling, especially when different sensor clocks are not hardware-synchronized [48]. Fig. 4.4a also shows the time offset calibration results, which are initialized from 0s. The average final converged values is -0.13 ± 0.03s.

Fig. 4.4b shows the repeatability of yaw calibration for different initial values, with Urban39 dataset. These initial values are obtained by adding different perturbation to ψ_0. The range of perturbation is $[-70.0°, 70.0°]$.

We evaluate the ATE of GPS positioning, VIO (Open-VINS [16]), GPS-VIO-fixed and our proposed algorithm. Results are summarized in TABLE 4.2. Our algorithm provides the highest localization accuracy. VIO suffers from drift issue from long trajectory. Moreover, the scale information of VIO system is unobservable when the vehicle undergoes constant acceleration motion [83, 5]. These issues can be solved by fusing GPS measurements once GPS-VIO system is successfully initialized (see Urban33 in TABLE 4.2).



(a) (b)

Figure 4.4: (a) Top: $(\psi - \psi_0)$ convergence over time. Bottom: Calibration results of the time offset between GPS and IMU. (b) ψ convergence over time respect to different initial values. The labels of legend represent different perturbation values.

4.8 Conclusion

This paper presents a novel tightly-coupled filter-based GPS-VIO algorithm which can benefit from the online estimation of the rotational extrinsic parameter between the GPS and the VIO reference frame. The proposed algorithm is able to refine the rotational calibration, thus improve the localization performance. The novel study on the observability of extrinsic parameter demonstrates that nonlinear observability analysis is more comprehensive and profound than linear observability analysis, for a nonlinear system. It is advised to validate the unobservable property derived from linear observability analysis in simulations. In future, we will investigate if we can obtain better localization results by formulating the estimation algorithm directly with the GNSS raw observations [53, 84].

Table 4.2: ATE (meter / degree) Comparison with the SOTA on the KAIST Dataset. – means trajectory divergence.

ID	Path len(km)	GPS	VIO [16]	GPS-VIO-fixed [48]	Ours
28	11.5	8.66	10.78 / 1.44	7.71 / 1.75	**4.67** / **1.42**
31	11.4	7.26	76.87 / 1.58	6.85 / 1.62	**5.56** / **1.55**
33	7.6	8.95	–	7.77 / 2.90	**4.94** / **1.27**
38	11.4	7.09	7.53 / 1.26	5.53 / 1.25	**3.86** / **1.22**
39	11.1	6.43	8.73 / 1.93	5.50 / 1.48	**2.63** / **1.24**

Chapter 5

Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion

Online extrinsic calibration is crucial for building ”power-on-and-go” moving platforms, like robots and AR devices. However, blindly performing online calibration for unobservable parameter may lead to unpredictable results. In the literature, extensive studies have been conducted on the extrinsic calibration between IMU and camera, from theory to practice. It is well-known that the observability of extrinsic parameter can be guaranteed under sufficient motion excitation. Furthermore, the impacts of degenerate motions are also investigated. Despite these successful analyses, we identify an issue with respect to the existing observability conclusion. This paper focuses on the observability investigation for straight line motion, which is a common-seen and fundamental degenerate motion in applications. We analytically prove that pure translational straight line motion can lead to the unobservability of the rotational extrinsic parameter between IMU and camera (at least one degree of freedom). By correcting the existing observability conclusion, our novel theoretical finding

disseminates more precise principle to the research community and provides explainable calibration guideline for practitioners. Our analysis is validated by rigorous theory and experiments.

5.1 Related paper

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion." (Accepted by IROS 2025).

5.2 Relationships to other chapters

Similarly to Chapter 3 and Chapter 4, the online target-less extrinsic calibration between IMU and camera is presented in this chapter. The offline method developed in Chapter 6 could provide nominal calibration parameters for the research of this chapter. The observability for online calibration, particularly rotational calibration, is investigated in this chapter.

5.3 Introduction

In the last two decades, visual-inertial navigation systems (VINS) have gained great popularity thanks to their ability to provide real-time and precise 6 degree-of-freedom (DoF) motion tracking in unknown GPS-denied or GPS-degraded environments, through the usage of low-cost, low-power, and complementary visual-inertial sensor rigs [9, 6, 3]. An inertial sensor, IMU, provides high-frequency linear acceleration and local angular velocity measurements of the moving platform, with bias and noise. Therefore, integrating only the IMU measurements to obtain motion prediction inevitably suffers from drift. While visual sensors can estimate IMU bias and reduce the drift of pose estimation by perceiving static visual features from the surrounding environment.

To improve the accuracy, efficiency, robustness or consistency of pose estimation, numerous tightly-coupled visual-inertial odometry (VIO) algorithms have been proposed in the literature. These algorithms can be broadly divided into two categories: optimization-based methods and filter-based methods. Optimization-based methods include OKVIS [85], VINS-Mono [38], and ORB-SLAM3 [40]. Filter-based methods include ROVIO [86], Multi-State Constraint Kalman Filter (MSCKF) [41, 16], and SchurVINS [4].

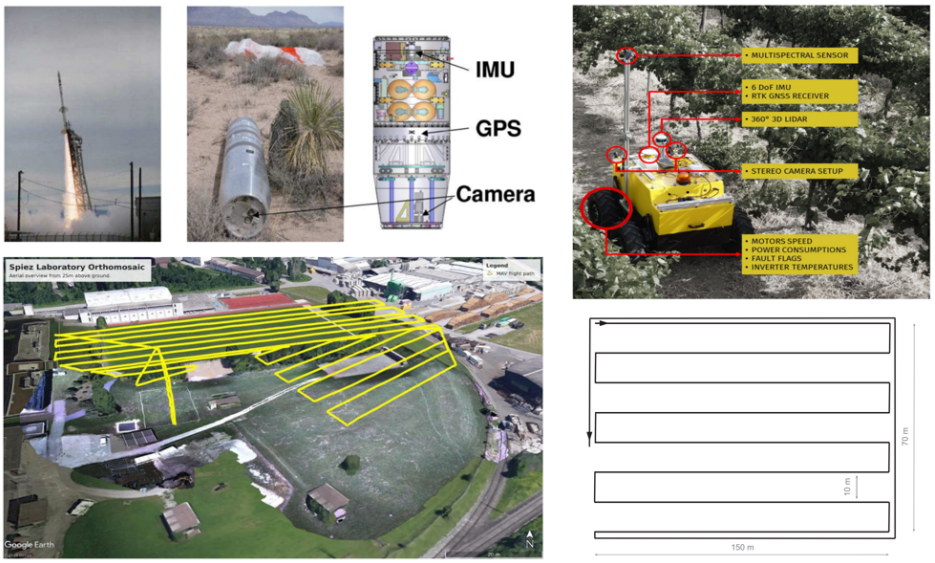


Figure 5.1: Various straight line movements. Top left: Spacecraft entry, descent, and landing [9]. Bottom left: MAV flight path [87]. Top right: Agrobot movement in a vineyard field [88]. Bottom right: Survey followed by Girona 1000 AUV [89].

Before running, VIO algorithm needs to know the extrinsic parameter between IMU and camera, including 3DoF translational part and 3DoF rotational part, which is a bridge to link measurements from different sensors. This extrinsic parameter is also critical for other visual perception applications, for example loop correction [38], dense map [90], and tracking [91, 92]. These visual perception results are represented in camera frame. To transfer these results to the body frame (IMU frame) of the robot or vehicle, accurate extrinsic parameter is desired. A small misalignment in the extrinsic parameter could generate a large drift and error.

The extrinsic parameter is usually assumed to be rigid and constant, however, this may

be not the case in practice. Considering that replacement and maintenance of sensors, and non-rigid deformation caused by mechanical vibration and varying temperature may lead to the alternation of extrinsic parameter, some researchers propose to add extrinsic parameter to state vector to perform *online calibration* [22, 23, 19]. If the extrinsic parameter is an observable state variable, online calibration can be resilient with poor prior calibration and converge to true value, which means robustness to the initial value. This feature helps to build ”power-on-and-go” moving platforms without the need for repetitive, tedious, manual *offline calibration*.

The success of online extrinsic parameter calibration depends on the observability. Remarkable works have studied the observability of extrinsic parameter between IMU and camera. With the help of artificial visual features on the calibration target board, [22] conclude that extrinsic parameter is observable if the moving platform undergoes at least 2DoF rotational excitation. An interesting corollary from [22] is that the observability of extrinsic parameter is independent of translational excitation. However, the conclusion of [22] is limited by the usage of calibration board, and cannot be applied to real operating environments without calibration board. [23] further extend the calibration of extrinsic parameter with target-less approach, and the conclusion is updated. The moving platform should undergo at least 2DoF motion excitation for both rotation and translation, to ensure the observability of extrinsic parameter.

The above-mentioned observability studies miss the analysis of degenerate motion profiles, which could be occurred and unavoidable in practice. As a supplement, [19] thoroughly explore the possible degenerate motion primitives and analyze the impact of degenerate motion on the observability of calibration parameters. We note that the rotational extrinsic parameter is summarized as observable for all identified degenerate motions (see Table I in [19]), except for no motion. However, by observing the top subplot of Fig. 2a in [19], we found that the rotational calibration results exhibit unexpected large RMSE (greater than 1 degree) for the case of pure translational straight line motion, which is clearly different from other motion cases. Actually, this distinct curve is an indicator for unobservability.

The inconsistency between the observability conclusion and the calibration results moti-

Table 5.1: Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion.

Motion	Pure VIO		Global-pose aided VIO	
	[19, 20]	Our novel finding	[19, 20]	Our novel finding
Pure translational straight line motion	observable	at least one unobservable DoF	observable	at least one unobservable DoF
Pure translational straight line motion with constant velocity	observable	fully unobservable	observable	at least one unobservable DoF

vates the following research question as the main purpose of this work:

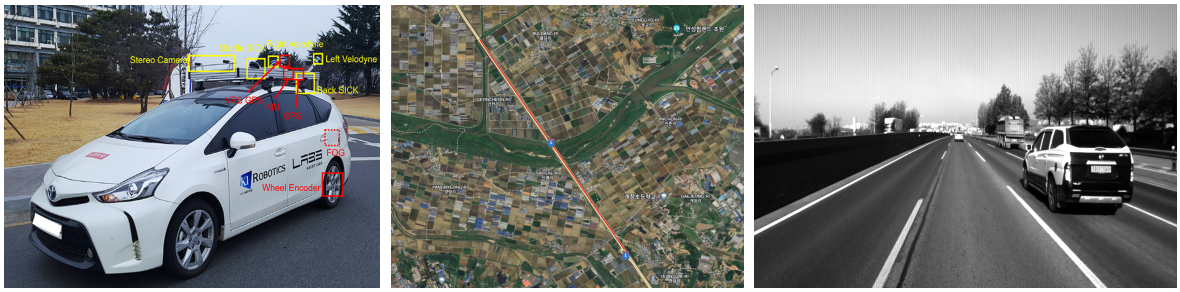
Is the rotational extrinsic parameter of (global-pose aided) VIO observable under pure translational straight line motion?

Straight line motions are quite common and fundamental in vehicle driving [14], agriculture [88], coverage survey [87, 89], and planetary exploration [11] (see Fig. 5.1 and Fig. 5.2). According to [19], if the rotational extrinsic parameter is observable, it is expected that practitioners would straightforward add this parameter to the state vector to perform online calibration, which has been integrated in numerous open-sourced VIO frameworks, like OKVIS [85], VINS-Mono [38], ROVIO [86], and Open-VINS [16].

However, according to our novel finding (see Tab. 5.1), the rotational extrinsic parameter has at least one unobservable DoF when the moving platform undergoes pure translational straight line motion. This implies that performing online rotational calibration is risky, as unobservability can lead to unpredictable and incorrect calibration results. Meanwhile, the misleading observability conclusion in [19] may have adverse effect on future research. For example, Table III in [20] is directly inherited from [19]. Therefore, it is vital to convey more precise principle to the community, otherwise incorrect conclusion would continue to mislead researchers. Next, we will verify our observability investigation through rigorous theory and solid experiments.

5.4 Notation

The main purpose of this paper is to investigate the observability of rotational extrinsic parameter between IMU and camera presented in [19]. When the moving platform follows a



(a) Dataset collection vehicle. (b) GPS trajectory. (c) Image from left camera.

Figure 5.2: Representative pure translational straight line motion from Urban22 sequence in KAIST dataset [14].

pure translational straight line motion (no rotation), our observability conclusion regarding this rotational extrinsic parameter is different from [19]. Like [19], we consider online calibration of rotational extrinsic parameter (rotational calibration) with two configurations, one is **pure VIO** and the other is **global-pose aided VIO**. In the following sections, we will directly analyze the observability matrix in [19]. As for the construction details of system model, measurement model, and observability matrix, interested readers are advised to refer to [19, 76].

The state vector considered in this paper is

$$x = \begin{bmatrix} {}^{I}_{G}q^T & b_g^T & {}^{G}v_I^T & b_a^T & {}^{G}p_I^T & {}^{C}_{I}q^T & {}^{G}p_f^T \end{bmatrix}^T \tag{5.1}$$

where ${}^{I}_{G}q$ represents the orientation of IMU frame $\{I\}$ with respect to global frame $\{G\}$, and its corresponding rotation matrix is given by ${}^{I}_{G}R$. ${}^{G}v_I$ and ${}^{G}p_I$ refer to the velocity and position of IMU in frame $\{G\}$. b_g and b_a represent the gyroscope and accelerometer biases. ${}^{G}p_f$ is augmented feature, or SLAM feature [16].

${}^{C}_{I}q$ is rotational calibration parameter, and its corresponding rotation matrix is ${}^{C}_{I}R$. Compared to equation (1) in [19], x does not include ${}^{C}p_I$ and t_d, as the online calibration of translational extrinsic parameter, as well as the time offset between IMU and camera, are not the focus of this paper. Our analysis is independent of ${}^{C}p_I$ and t_d.

In following sections, $[\bullet]_\times$ is denoted as the skew symmetric matrix corresponding to a

three-dimensional vector. To simplify the description, the hat symbol $(\hat{\bullet})$ is omitted, which does not affect observability analysis. Other notations are consistent with [19]. By assuming that the direction of straight line is denoted as d in the IMU frame $\{I\}$, we are ready for observability investigation now.

5.5 Observability Investigation for Pure VIO

Referring to equation (21) of [19], in the configuration of pure VIO, the observability matrix is

$$\begin{aligned} M_k &= \Xi_k \Xi_{\Gamma_k} \\ \Xi_{\Gamma_k} &= \begin{bmatrix} \Gamma_1 & \Gamma_2 & -I_3 \delta t_k & \Gamma_3 & -I_3 & \Gamma_4 & I_3 \end{bmatrix} \end{aligned} \tag{5.2}$$

Compared to equation (21) of [19], the element ${}^G_{I_k}R^I_C R$ corresponding to ${}^C p_I$, and the element Γ_5 corresponding to t_d, have been removed in Ξ_{Γ_k}. The expressions of $\Gamma_1 \sim \Gamma_4$ in Ξ_{Γ_k} are

$$\begin{aligned} \Gamma_1 &= \left[{}^G p_f - {}^G p_{I_1} - {}^G v_{I_1} \delta t_k + \tfrac{1}{2} {}^G g \delta t_k^2 \right]_\times {}^G_{I_1}R \\ \Gamma_2 &= \left[{}^G p_f - {}^G p_{I_k} \right]_\times {}^G_{I_k}R \Phi_{I12} - \Phi_{I52} \\ \Gamma_3 &= -\Phi_{I54} \\ \Gamma_4 &= \left[{}^G p_f - {}^G p_{I_k} \right]_\times {}^G_{I_k}R^I_C R \end{aligned} \tag{5.3}$$

The expression of Γ_1 in [19], equation (22), has small typos. We have corrected it by referring to equation (53) in [76].

In the context of pure translational motion, i.e. no rotation, the orientation of the moving platform does not change at any time. Therefore, ${}^G_{I_{(\bullet)}}R$ can be directly represented by ${}^G_I R$ (constant). Referring to equation (114) in [76]

$$\begin{aligned} \Gamma_3 &= -\Phi_{I54} = \int_{t_1}^{t_k} \int_{t_1}^{s} {}^G_{I_\tau}R \, d\tau ds \\ &= \left({}^G_I R \right) \int_{t_1}^{t_k} \int_{t_1}^{s} (1) d\tau ds = \tfrac{1}{2} {}^G_I R \delta t_k^2 \end{aligned} \tag{5.4}$$

The expressions of $\Gamma_1 \sim \Gamma_4$ in Ξ_{Γ_k} become

$$\begin{aligned}
\Gamma_1 &= \left[{}^G p_f - {}^G p_{I_1} - {}^G v_{I_1} \delta t_k + \tfrac{1}{2} {}^G g \delta t_k^2\right]_\times {}^G_I R \\
\Gamma_2 &= \left[{}^G p_f - {}^G p_{I_k}\right]_\times {}^G_I R \Phi_{I12} - \Phi_{I52} \\
\Gamma_3 &= \tfrac{1}{2} {}^G_I R \delta t_k^2 \\
\Gamma_4 &= \left[{}^G p_f - {}^G p_{I_k}\right]_\times {}^G_I R {}^I_C R
\end{aligned} \tag{5.5}$$

Lemma 5.5.1. *If pure VIO system undergoes pure translational straight line motion, the unobservable directions of ${}^C_I R$ depend on the projection of d[1] in the camera frame $\{C\}$. The corresponding right null space of M_k is*

$$N_1 = \begin{bmatrix} 0_{15\times 1} \\ {}^C_I R d \\ -\left[{}^G p_f - {}^G p_{I_1}\right]_\times {}^G_I R d \end{bmatrix} \tag{5.6}$$

Proof. Straight line motion indicates the following geometric constraint

$$\left[{}^{I_1} p_{I_k}\right]_\times d = 0 \tag{5.7}$$

Given the above constraint, we first verify that N_1 belongs to the right null space of Ξ_{Γ_k}.

$$\begin{aligned}
\Xi_{\Gamma_k} N_1 &= \Gamma_4 {}^C_I R d - \left[{}^G p_f - {}^G p_{I_1}\right]_\times {}^G_I R d \\
&= \left[{}^G p_f - {}^G p_{I_k}\right]_\times {}^G_I R d - \left[{}^G p_f - {}^G p_{I_1}\right]_\times {}^G_I R d \\
&= -\left[{}^G p_{I_k} - {}^G p_{I_1}\right]_\times {}^G_I R d
\end{aligned} \tag{5.8}$$

One geometric relationship can be utilized

$${}^G p_{I_k} = {}^G p_{I_1} + {}^G_{I_1} R {}^{I_1} p_{I_k} = {}^G p_{I_1} + {}^G_I R {}^{I_1} p_{I_k} \tag{5.9}$$

[1]The definition of d is described in the last paragraph of Sec. 5.4.

Subsequently

$$\begin{aligned}\Xi_{\Gamma_k} N_1 &= -\left[{}^G p_{I_k} - {}^G p_{I_1}\right]_\times {}^G_I R d \\ &= -\left[{}^G_I R\, {}^{I_1} p_{I_k}\right]_\times {}^G_I R d \\ &= -{}^G_I R \left[{}^{I_1} p_{I_k}\right]_\times {}^G_I R^T {}^G_I R d \\ &= -{}^G_I R \left[{}^{I_1} p_{I_k}\right]_\times d = 0\end{aligned} \tag{5.10}$$

Finally

$$\Rightarrow M_k N_1 = \Xi_k \Xi_{\Gamma_k} N_1 = 0 \tag{5.11}$$

Hence, N_1 belongs to the right null space of M_k. N_1 indicates that the unobservable directions of ${}^C_I R$ are dependent on the non-zero components of ${}^C_I R d$. □

Lemma 5.5.2. *If pure VIO system undergoes pure translational straight line motion with constant velocity, the 3DoF of ${}^C_I R$ are all unobservable. The corresponding right null space of M_k is*

$$N_2 = \begin{bmatrix} {}^I_G R \\ 0_3 \\ 0_3 \\ -{}^I_G R \left[{}^G g\right]_\times \\ 0_3 \\ -{}^C_I R {}^I_G R \\ 0_3 \end{bmatrix} \tag{5.12}$$

Proof. Straight line motion with constant velocity indicates the following geometric constraint

$${}^G p_{I_k} = {}^G p_{I_1} + {}^G v_{I_1} \delta t_k \tag{5.13}$$

Given the above constraint, we first verify that N_2 belongs to the right null space of Ξ_{Γ_k}.

$$\begin{aligned}\Xi_{\Gamma_k} N_2 &= \Gamma_1 {}^I_G R - \Gamma_3 {}^I_G R \left[{}^G g\right]_\times - \Gamma_4 {}^C_I R {}^I_G R \\ &= \left[{}^G p_f - {}^G p_{I_1} - {}^G v_{I_1} \delta t_k + \tfrac{1}{2} {}^G g \delta t_k^2\right]_\times \\ &\quad - \tfrac{1}{2}\left[{}^G g\right]_\times \delta t_k^2 - \left[{}^G p_f - {}^G p_{I_k}\right]_\times \\ &= \left[{}^G p_{I_k} - {}^G p_{I_1} - {}^G v_{I_1} \delta t_k\right]_\times = 0\end{aligned} \tag{5.14}$$

Finally

$$\Rightarrow M_k N_2 = \Xi_k \Xi_{\Gamma_k} N_2 = 0 \tag{5.15}$$

Hence, N_2 belongs to the right null space of M_k. N_2 indicates that the 3DoF of ${}^C_I R$ are all unobservable. □

Remark. *We note that the rotational extrinsic parameter ${}^C_I R$ has at least one degree of freedom that is unobservable when the platform undergoes pure translational straight line motion. More specifically, when moving with constant velocity, the 3 degrees of freedom of ${}^C_I R$ are completely unobservable. When moving with variable velocity, at least one degree of freedom is unobservable as $\left\| {}^C_I R d \right\| \neq 0$.*

5.6 Observability Investigation for Global-pose aided VIO

Like [19], the observability of rotational extrinsic parameter is also discussed in the configuration of global-pose aided VIO. Our conclusion is different from [19]. Referring to equation (40) of [19], the observability matrix is

$$\begin{aligned} M_k^{(g)} &= \Xi_k^{(g)} \Xi_{\Gamma_k}^{(g)} \\ \Xi_{\Gamma_k}^{(g)} &= \begin{bmatrix} \Gamma_1 & \Gamma_2 & -I_3 \delta t_k & \Gamma_3 & -I_3 & \Gamma_4 & I_3 \\ \Phi_{I11} & \Phi_{I12} & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ \Phi_{I51} & \Phi_{I52} & \Phi_{I53} & \Phi_{I54} & I_3 & 0_3 & 0_3 \end{bmatrix} \end{aligned} \tag{5.16}$$

The last two rows of $\Xi_{\Gamma_k}^{(g)}$ in [19] is incorrect. We have corrected it by multiplying the measurement Jacobian matrix with the state transition matrix. Detailed derivations are provided in Sec. 5.9 of supplementary material [21].

Lemma 5.6.1. *If global-pose aided VIO system undergoes pure translational straight line motion, the unobservable directions of ${}^C_I R$ depend on the projection of d in the camera frame $\{C\}$. The corresponding right null space of $M_k^{(g)}$ is N_1.*

Proof. A naive way of finding the corresponding right null space for $M_k^{(g)}$ is to test the product

of $\Xi_{\Gamma_k}^{(g)}$ and N_1

$$\Xi_{\Gamma_k}^{(g)} N_1 = \begin{bmatrix} \Gamma_4 {}_I^C Rd - \left[{}^G p_f - {}^G p_{I_1}\right]_\times {}_I^G Rd \\ 0_{3\times 1} \\ 0_{3\times 1} \end{bmatrix} \tag{5.17}$$

According to Theorem 5.5.1

$$\Xi_{\Gamma_k}^{(g)} N_1 = 0 \tag{5.18}$$

Finally

$$\Rightarrow M_k^{(g)} N_1 = \Xi_k^{(g)} \Xi_{\Gamma_k}^{(g)} N_1 = 0 \tag{5.19}$$

Hence, N_1 belongs to the right null space of $M_k^{(g)}$. N_1 indicates that the unobservable directions of ${}_I^C R$ are dependent on the non-zero components of ${}_I^C Rd$. □

Lemma 5.6.2. *If global-pose aided VIO system undergoes pure translational straight line motion with constant velocity, the unobservable directions of ${}_I^C R$ depend on the projection of d in the camera frame $\{C\}$. The corresponding right null space of $M_k^{(g)}$ is still N_1.*

Proof. A naive way of finding the corresponding right null space for $M_k^{(g)}$ is to test the product of $\Xi_{\Gamma_k}^{(g)}$ and N_2

$$\Xi_{\Gamma_k}^{(g)} N_2 = \begin{bmatrix} \Gamma_1 {}_G^I R - \Gamma_3 {}_G^I R \left[{}^G g\right]_\times - \Gamma_4 {}_I^C R {}_G^I R \\ \Phi_{I11} {}_G^I R \\ \Phi_{I51} {}_G^I R - \Phi_{I54} {}_G^I R \left[{}^G g\right]_\times \end{bmatrix} \tag{5.20}$$

According to Theorem 5.5.2

$$\Xi_{\Gamma_k}^{(g)} N_2 = \begin{bmatrix} 0_3 \\ \Phi_{I11} {}_G^I R \\ \Phi_{I51} {}_G^I R - \Phi_{I54} {}_G^I R \left[{}^G g\right]_\times \end{bmatrix} \tag{5.21}$$

Referring to equation (46) in [76], $\Phi_{I11} \neq 0$, it is clear that $\Xi_{\Gamma_k}^{(g)} N_2 \neq 0$. Therefore, the unobservable direction N_2 is no longer hold due to the inclusion of global pose measurement.

It is worth noting that Theorem 5.6.2 is a special case of Theorem 5.6.1. Hence, N_1 still belongs to the right null space of $M_k^{(g)}$. N_1 indicates that the unobservable directions of ${}_I^C R$

are dependent on the non-zero components of ${}^{C}_{I}Rd$. □

Remark. *We note that the rotational extrinsic parameter ${}^{C}_{I}R$ has at least one degree of freedom that is unobservable when the platform undergoes pure translational straight line motion, regardless of variable velocity or constant velocity. In the case of constant velocity, the unobservable directions can be decreased with the aides of global pose measurement, compared to the pure VIO configuration. More specifically, in the global-pose aided VIO configuration, the worst case is three degrees of freedom are unobservable, while the best case is only one degree of freedom is unobservable. The difference between our conclusion and [19] is marked in Tab. 5.1.*

5.7 Results

We conduct verification experiments based on Open-VINS [16]. As this paper focuses on the observability investigation of the rotational extrinsic parameter, we only perform online calibration for the rotational extrinsic parameter and set the translational extrinsic parameter and time offset as true values, referring to our state vector (Eq. (5.1)).

5.7.1 Comments on results in [19]

Table I from [19] show that the rotational extrinsic parameter is observable for pure translational motion. However, we find that [19] actually did not perform theoretical analysis on the rotational calibration (${}^{C}_{I}R$). Besides that, it can be seen from the top subplot of Fig. 2a in [19], if the simulation trajectory is a pure translational straight line motion with constant velocity, the calibration result of the rotational extrinsic parameter shows large RMSE (greater than 1 degree). Regarding the inconsistency between observability assertion and simulation result, no ablation experiments were conducted, by calibrating the rotational extrinsic parameter only and turning off the calibration of the translational extrinsic parameter and time offset. Moreover, Section VI of [19] did not validate the convergence consistency with different initial ${}^{C}_{I}R$. Section VII of [19] missed the verification of pure translational straight

Table 5.2: Final calibration results of the rotational extrinsic parameter for pure VIO system undergoes pure translational straight line motion with variable velocity. The absolute errors of roll, pitch, and yaw at 60s, are recorded with different perturbations.

Perturbations of (roll, pitch, yaw)	Case-1			Case-2			Case-3		
	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw
(2, -4, -5)	11.31	0.04	0.03	5.28	2.17	0.05	4.67	1.83	2.21
(-4, 3, 3)	0.37	0.05	0.02	2.16	0.91	0.03	4.04	1.69	1.69
(5, -2, -1)	13.38	0.04	0.02	7.02	2.89	0.09	1.01	0.41	0.53
(-1, -5, -3)	8.96	0.03	0.02	2.67	1.10	0.01	4.99	1.94	2.23
(3, 0, 1)	9.55	0.04	0.01	4.60	1.89	0.06	0.40	0.19	0.20
(1, 2, -4)	7.34	0.06	0.04	3.52	1.44	0.05	5.98	2.44	2.63
(0, 5, 2)	2.51	0.05	0.02	1.60	0.64	0.07	1.22	0.57	0.52
(-3, 4, 0)	1.58	0.05	0.03	0.63	0.28	0.03	4.82	1.98	2.08
(-5, 1, 4)	0.44	0.04	0.01	3.20	1.35	0.02	6.09	2.50	2.53
(4, -1, 5)	9.20	0.02	0.01	4.49	1.84	0.05	1.12	0.39	0.51
(-2, -3, -2)	7.36	0.04	0.02	0.77	0.30	0.01	5.89	2.34	2.56
Avg	6.55	0.04	0.02	3.27	1.35	0.04	3.66	1.48	1.61

line motion in real-world experiments.

5.7.2 Numerical Study

Employing the Open-VINS simulator and importing the desired 6DoF trajectory, realistic multi-sensor data are generated for experiments under two different configurations. For the pure VIO configuration, we generate IMU measurements at 400 Hz and image measurements at 10 Hz. For the global-pose aided VIO configuration, additional 10 Hz global-pose measurements are generated. The global-pose measurement noises are defined as

$$\begin{aligned} n_p &\sim \mathcal{N}\left(0_{3\times1}, \sigma_p^2 I_3\right), \sigma_p = 0.1m \\ n_\theta &\sim \mathcal{N}\left(0_{3\times1}, \sigma_\theta^2 I_3\right), \sigma_\theta = 0.1rad \end{aligned} \tag{5.22}$$

where n_p and n_θ represent Gaussian noises for global position and orientation measurement, respectively.

This paper focuses on pure translational straight line motion, therefore the orientation of the input trajectory, ${}^I_G R$, is set as I_3. To validate the observability assertion summarized in

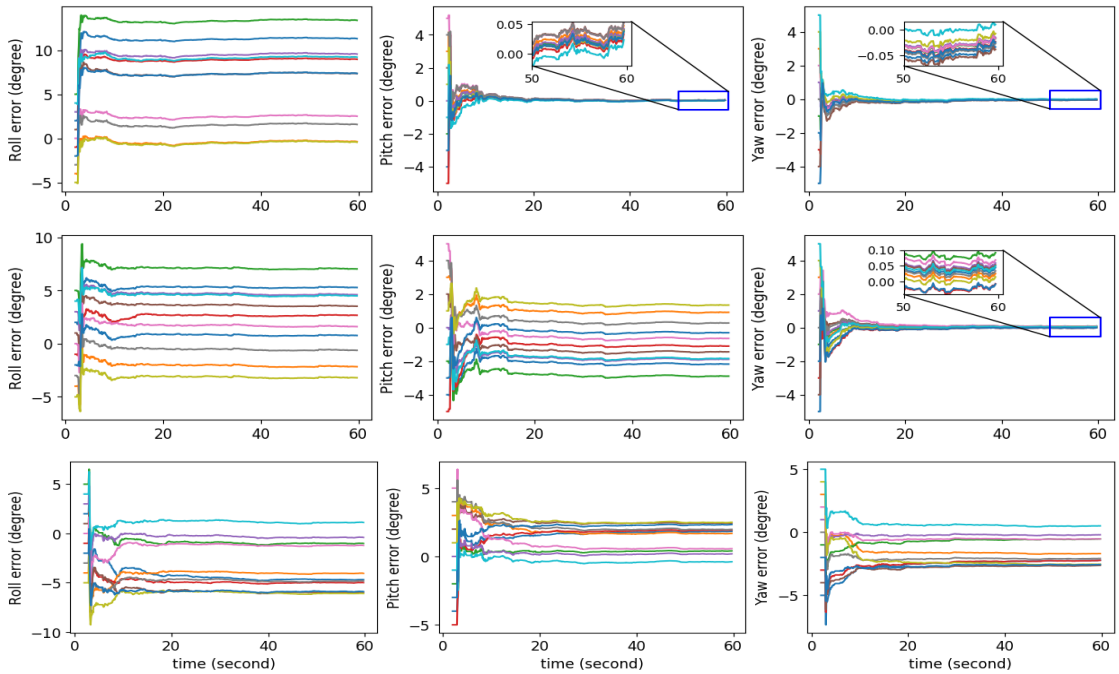


Figure 5.3: Calibration results for pure VIO system undergoes pure translational straight line motion with variable velocity. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds. Top to bottom corresponds to Case-1 to Case-3 in Sec. 5.7.2.

Tab. 5.1, two types of straight line motion with different velocity profiles are designed as

- Trajectory-1: ${}^{G}p_I = \begin{bmatrix} 2\cos\left(\frac{\pi}{5}t\right) & 0 & 0 \end{bmatrix}^T$.
- Trajectory-2: ${}^{G}p_I = \begin{bmatrix} 0.5t & 0 & 0 \end{bmatrix}^T$.

Trajectory-1 corresponds to variable velocity motion, while Trajectory-2 corresponds to constant velocity motion. The direction vector corresponding to both these two trajectories is $d = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$. As the unobservable directions of ${}^{C}_{I}R$ may depend on the non-zero components of ${}^{C}_{I}Rd$, three types of groundtruth ${}^{C}_{I}R$ are designed as

- Case-1:
$${}^{C}_{I}R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, {}^{C}_{I}Rd = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$
- Case-2:
$${}^{C}_{I}R = \begin{bmatrix} 0.707 & 0.707 & 0 \\ -0.707 & 0.707 & 0 \\ 0 & 0 & 1 \end{bmatrix}, {}^{C}_{I}Rd = \begin{bmatrix} 0.707 \\ -0.707 \\ 0 \end{bmatrix}.$$
- Case-3:
$${}^{C}_{I}R = \begin{bmatrix} 0.5 & 0.707 & -0.5 \\ -0.5 & 0.707 & 0.5 \\ 0.707 & 0 & 0.707 \end{bmatrix}, {}^{C}_{I}Rd = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.707 \end{bmatrix}.$$

For each case, we initialize ${}^{C}_{I}R$ by adding different perturbations to the three degrees of freedom of ${}^{C}_{I}R$ (roll, pitch, and yaw), and collect calibration error with respect to groundtruth ${}^{C}_{I}R$. The range of perturbation is $[-5.0°, 5.0°]$. If a certain degree of freedom is observable, it should be robust to different perturbations, namely, the calibration error should consistently converge to 0. On the contrary, if it is unobservable, the calibration error can not converge to 0 and is expected to be sensitive to the initial value.

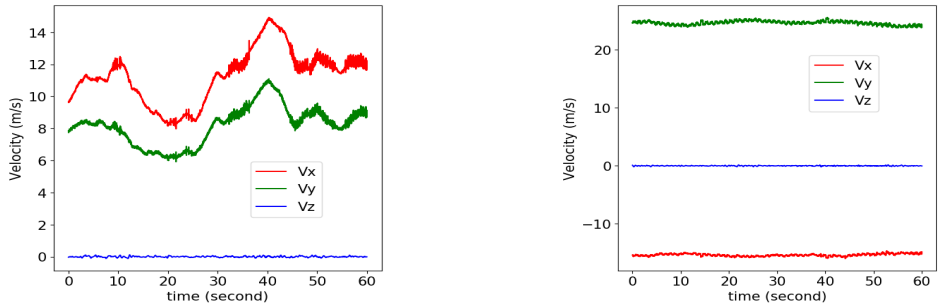
Firstly, we analyze the calibration results for Case-1 of Trajectory-1 in the pure VIO configuration, as shown in the Tab. 5.2. Pitch and yaw exhibit observable characteristic, while roll not. This is because the non-zero component of ${}^{C}_{I}Rd$ corresponds to roll. For Case-2,

Table 5.3: Numerical Study Results for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion.

Motion	Pure VIO		Global-pose aided VIO	
	calibration results	conclusion	calibration results	conclusion
Trajectory-1 in Sec. 5.7.2	Tab. 5.2 and Fig. 5.3	at least one unobservable DoF	Tab. 5.6 and Fig. 5.7 in supplementary material [21]	at least one unobservable DoF
Trajectory-2 in Sec. 5.7.2	Tab. 5.7 and Fig. 5.8 in supplementary material [21]	fully unobservable	Tab. 5.8 and Fig. 5.9 in supplementary material [21]	at least one unobservable DoF

yaw exhibits observable characteristic, while roll and pitch not. This is because non-zero components of ${}^{C}_{I}Rd$ correspond to roll and pitch. For Case-3, roll, pitch, and yaw all exhibit unobservable characteristic. This is because none of the three components of ${}^{C}_{I}Rd$ are zero. The calibration results over time are shown in the Fig. 5.3. Similar analysis also applies to different combinations of configurations and trajectories, please refer to Tab. 5.3 and Sec. 5.10 of supplementary material [21] for other results. These results successfully validate that our novel observability conclusions are correct. Overall, observable degree of freedom shows deterministic behavior, i.e. converging to groundtruth over time, while unobservable degree of freedom exhibits unpredictable behavior.

5.7.3 Real-world Dataset



(a) (b)

Figure 5.4: Velocity profiles of Urban34 (a) and Urban22 (b).

Straight line motions are quite common in real-world scenarios. On one hand, straight line cruise is the most efficient and energy-saving trajectory for most robot applications. On

the other hand, substantial artificial scenarios have specific constraints on motion, such as applications in agriculture, warehousing, logistics, and transportation.

The KAIST urban dataset [14] contains the driving scenario on the highway, as shown in the Fig. 5.2. Urban34 and Urban22 from this dataset are leveraged to confirm our observability finding, as these two sequences represent variable velocity motion and constant velocity motion, respectively. The vehicle used to collect data follows the same lane during driving, so its trajectory can be regarded as a pure translational straight line. Corresponding ${}^{C}_{I}Rd$ is

$$ {}^{C}_{I}Rd = \begin{bmatrix} -0.00413 \\ -0.01966 \\ 0.99980 \end{bmatrix} $$

The velocity curve of Urban34 sequence (see Fig. 5.4a) is variable over time. Fig. 5.5 shows the calibration results with the pure VIO configuration and the global-pose aided VIO configuration. Roll and pitch exhibit observable characteristic, while yaw not. This is because the non-zero component of ${}^{C}_{I}Rd$ is dominated by the yaw component (0.99980). The velocity curve of Urban22 sequence (see Fig. 5.4b) is approximately constant. Fig. 5.6 shows the calibration results of Urban22. In the pure VIO configuration, roll, pitch, and yaw all exhibit unobservable characteristic due to constant velocity motion. In the global-pose aided VIO configuration, unobservable degrees of freedom are reduced from 3 to 1 (yaw). Interestingly, the convergence error of pitch is larger than that of roll, which can be attributed to the fact that the absolute value of pitch component (0.01966) is larger than that of roll (0.00413).

Furthermore, we evaluate the localization accuracy with calibration (w. calib) and without calibration (wo. calib), under different perturbations on the rotational extrinsic parameter. Since real-world data is more sensitive than simulation data, the perturbation amplitude is reduced to half of its value listed in the Tab. 5.2. The Absolute Trajectory Error (ATE) results are reported in Tab. 5.4 and Tab. 5.5.

In the pure VIO configuration (Tab. 5.4), calibration significantly improves the localization accuracy for Urban34, as model error from two degrees of freedom (roll and pitch) of the rotational calibration parameter can be corrected to near 0, thanks to online calibration (see top of Fig. 5.5). Urban22 exhibits large localization error as scale becomes unobservable

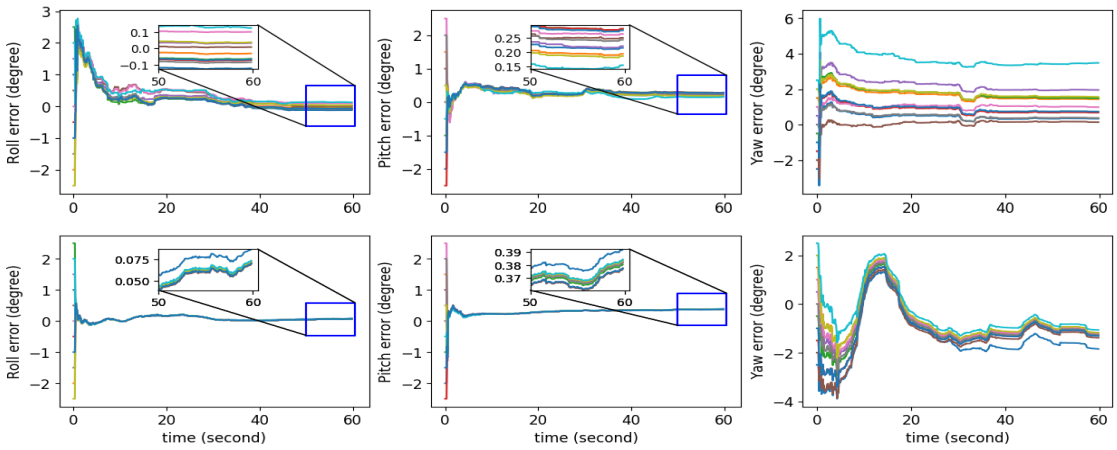


Figure 5.5: Calibration results for Urban34. Top: Results for pure VIO system. Bottom: Results for global-pose aided VIO system. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds.

under constant velocity motion [11]. And it is observed that performing calibration further degrades the localization due to the fully unobservable property of the rotational calibration parameter (see top of Fig. 5.6). In the global-pose aided VIO configuration (Tab. 5.5), the localization accuracy is mainly dominated by global pose measurements, thus the calibration of rotational extrinsic parameter has negligible impact on the accuracy.

Remark. *If the calibration parameter is observable, online calibration typically brings positive benefits to localization [8]. However, if it is unobservable, the impact of calibration on localization is unpredictable (negative, no impact or positive). In other words, we cannot determine the observability of the calibration parameter from localization accuracy.*

5.8 Conclusion

We investigate the observability from [19, 20], and prove that the common-seen pure translational straight line motion can lead to the unobservability of the rotational extrinsic parameter between IMU and camera (at least one degree of freedom). Our novel finding is carefully verified through rigorous theory, numerical study, and real-world experiment. This finding makes up for the shortcomings of the existing research conclusions. When the observability

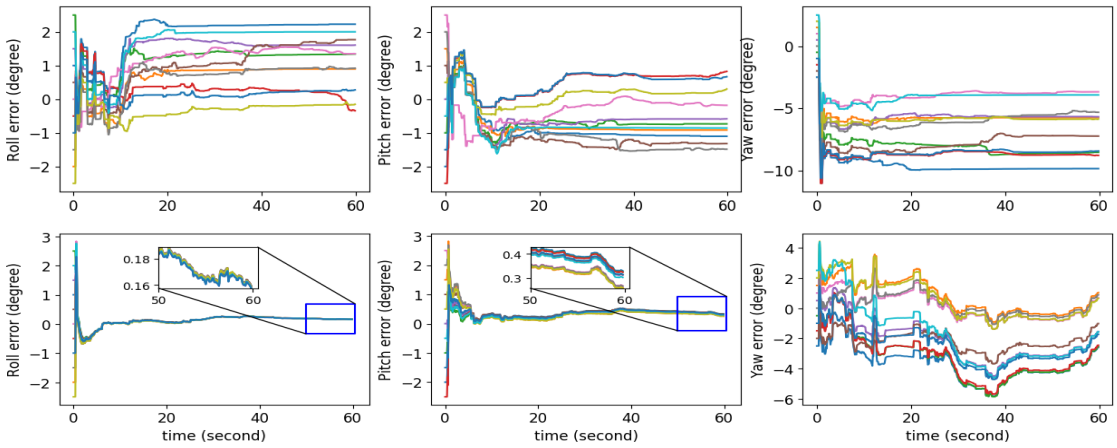


Figure 5.6: Calibration results for Urban22. Top: Results for pure VIO system. Bottom: Results for global-pose aided VIO system. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds.

Table 5.4: ATE (meter) Comparison for Pure VIO.

Perturbations	Urban34		Urban22	
	w. calib	wo. calib	w. calib	wo. calib
(1.0, -2.0, -2.5)	2.49	13.27	257.52	212.16
(-2.0, 1.5, 1.5)	4.24	108.88	258.47	88.39
(2.5, -1.0, -0.5)	2.90	106.31	268.02	308.66
(-0.5, -2.5, -1.5)	2.38	53.36	125.40	90.07
(1.5, 0.0, 0.5)	10.23	36.76	282.57	119.22
(0.5, 1.0, -2.0)	9.78	5.31	238.12	102.42
(0.0, 2.5, 1.0)	7.65	49.58	266.51	128.76
(-1.5, 2.0, 0.0)	2.30	80.54	198.29	54.08
(-2.5, 0.5, 2.0)	5.54	112.60	78.57	97.87
(2.0, -0.5, 2.5)	39.81	127.98	248.98	335.19
(-1.0, -1.5, -1.0)	2.37	45.99	94.43	56.92
Avg	8.15	67.33	210.63	144.89

Table 5.5: ATE (meter) Comparison for Global-pose aided VIO.

Perturbations	Urban34		Urban22	
	w. calib	wo. calib	w. calib	wo. calib
Avg	0.41	0.42	0.19	0.19

conclusion is inconsistent with the numerical study results (see our comments in Sec. 5.7.1), we recommend:

- Perform ablation experiments to eliminate the influence of other calibration parameters.
- Try different initial values to test the convergence consistency of the interested calibration parameter.

Mathematical derivations of this paper and [19] require delicate search for the null space of the observability matrix. And this process is case by case, which prompts us a research question for future work. Is there an automatic and natural way to find degenerate motion and corresponding unobservable degrees of freedom, thus avoiding potential manual missing or mistake?

Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion

Supplementary Material

5.9 Correction of the observability matrix for global-pose aided VIO

The observability matrix plays a key role for the observability analysis of a linear or nonlinear state estimator. According to the Section II.E of [19], the measurement Jacobian matrix and state transition matrix need to be calculated in advance to construct the observability matrix. For ease of description, recall our state vector (Eq. (5.1))

$$x = \begin{bmatrix} {}^{I}_{G}q^T & b_g^T & {}^{G}v_I^T & b_a^T & {}^{G}p_I^T & {}^{C}_{I}q^T & {}^{G}p_f^T \end{bmatrix}^T \tag{5.23}$$

According to the Section II.D of [19], the measurement Jacobian matrix corresponding to global pose measurement can be calculated as

$$H_{V_k} = \begin{bmatrix} I_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & I_3 & 0_3 & 0_3 \end{bmatrix} \tag{5.24}$$

Combining the measurement Jacobian matrix corresponding to visual measurement, H_{C_k}, the overall measurement Jacobian matrix can be denoted as

$$H_k = \begin{bmatrix} H_{C_k} \\ H_{V_k} \end{bmatrix} \tag{5.25}$$

Referring to equation (5) of [19], the expression of our state transition matrix is

$$\Phi(k,1)=\begin{bmatrix} \Phi_{I11} & \Phi_{I12} & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ \Phi_{I31} & \Phi_{I32} & I_3 & \Phi_{I34} & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & I_3 & 0_3 & 0_3 & 0_3 \\ \Phi_{I51} & \Phi_{I52} & \Phi_{I53} & \Phi_{I54} & I_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & I_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & I_3 \end{bmatrix} \tag{5.26}$$

Finally, the observability matrix for global-pose aided VIO can be constructed by multiplying the measurement Jacobian matrix with the state transition matrix

$$\begin{aligned} M_k^{(g)} &= H_k\Phi(k,1) \\ &= \begin{bmatrix} H_{C_k}\Phi(k,1) \\ H_{V_k}\Phi(k,1) \end{bmatrix} \\ &= \begin{bmatrix} \Xi_k\Xi_{\Gamma_k} \\ H_{V_k}\Phi(k,1) \end{bmatrix} \\ &= \begin{bmatrix} \Xi_k & 0 \\ 0 & I_6 \end{bmatrix} \\ &\quad \times \begin{bmatrix} \Gamma_1 & \Gamma_2 & -I_3\delta t_k & \Gamma_3 & -I_3 & \Gamma_4 & I_3 \\ \Phi_{I11} & \Phi_{I12} & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ \Phi_{I51} & \Phi_{I52} & \Phi_{I53} & \Phi_{I54} & I_3 & 0_3 & 0_3 \end{bmatrix} \end{aligned} \tag{5.27}$$

This completes the correction for the equation (40) of [19].

5.10 Additional results on numerical study

In this section, we will analyze additional calibration results described in Tab. 5.3 to complete the validation of our observable conclusions.

Tab. 5.6 shows the final calibration results of Trajectory-1 in the global-pose aided VIO

configuration. For Case-1, pitch and yaw exhibit observable characteristic, while roll not. This is because the non-zero component of ${}^{C}_{I}Rd$ corresponds to roll. For Case-2, yaw exhibits observable characteristic, while roll and pitch not. This is because non-zero components of ${}^{C}_{I}Rd$ correspond to roll and pitch. For Case-3, roll, pitch, and yaw all exhibit unobservable characteristic. This is because none of the three components of ${}^{C}_{I}Rd$ are zero. The calibration results over time are shown in the Fig. 5.7.

Tab. 5.7 shows the final calibration results of Trajectory-2 in the pure VIO configuration. For Case-1, Case-2, and Case-3, roll, pitch, and yaw all exhibit unobservable characteristic. This can be explained by Theorem 5.5.2, which indicates constant velocity motion lead to the fully unobservable property of the rotational extrinsic parameter. The calibration results over time are shown in the Fig. 5.8.

Tab. 5.8 shows the final calibration results of Trajectory-2 in the global-pose aided VIO configuration. We can still observe that the convergence of the rotational extrinsic parameter, depends on which components of ${}^{C}_{I}Rd$ are 0. The calibration results over time are shown in the Fig. 5.9.

These calibration results, and the corresponding observability conclusion they supported, are summarized in Tab. 5.3. Extensive experimental results demonstrate the correctness of our novel theoretical finding (see Tab. 5.1).

Table 5.6: Final calibration results of the rotational extrinsic parameter for global-pose aided VIO system undergoes pure translational straight line motion with variable velocity. The absolute errors of roll, pitch, and yaw at 60s, are recorded with different perturbations.

Perturbations of (roll, pitch, yaw)	Case-1			Case-2			Case-3		
	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw
(2, -4, -5)	0.89	0.07	0.01	4.00	1.73	0.02	4.98	1.84	2.15
(-4, 3, 3)	12.49	0.08	0.01	9.44	3.98	0.12	0.89	0.28	0.37
(5, -2, -1)	1.22	0.08	0.01	1.05	0.50	0.01	7.34	2.71	3.21
(-1, -5, -3)	5.82	0.08	0.01	6.82	2.89	0.06	3.62	1.33	1.55
(3, 0, 1)	4.31	0.08	0.01	2.87	1.26	0.02	6.01	2.22	2.62
(1, 2, -4)	5.33	0.08	0.01	4.83	2.07	0.03	2.92	1.06	1.25
(0, 5, 2)	8.68	0.08	0.01	5.59	2.39	0.05	2.98	1.08	1.29
(-3, 4, 0)	10.75	0.08	0.01	8.53	3.60	0.10	1.28	0.43	0.53
(-5, 1, 4)	13.39	0.08	0.01	10.38	4.37	0.14	0.92	0.29	0.38
(4, -1, 5)	4.49	0.08	0.01	1.88	0.85	0.01	8.02	2.96	3.52
(-2, -3, -2)	7.50	0.08	0.01	7.80	3.30	0.08	3.23	1.18	1.38
Avg	6.81	0.08	0.01	5.74	2.45	0.06	3.84	1.40	1.66

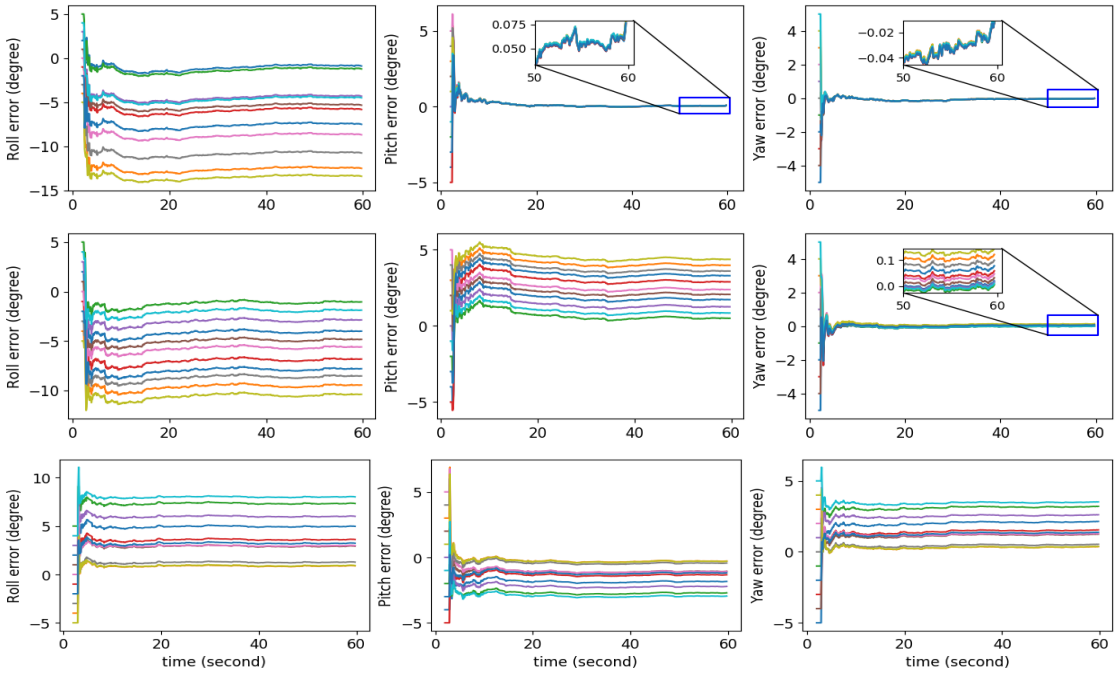


Figure 5.7: Calibration results for global-pose aided VIO system undergoes pure translational straight line motion with variable velocity. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds. Top to bottom corresponds to Case-1 to Case-3 in Sec. 5.7.2.

Table 5.7: Final calibration results of the rotational extrinsic parameter for pure VIO system undergoes pure translational straight line motion with constant velocity. The absolute errors of roll, pitch, and yaw at 60s, are recorded with different perturbations.

Perturbations of (roll, pitch, yaw)	Case-1			Case-2			Case-3		
	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw
(2, -4, -5)	0.46	5.65	4.81	4.63	3.70	2.06	1.29	10.38	4.34
(-4, 3, 3)	2.77	1.82	3.77	7.74	2.02	2.03	5.97	2.39	6.96
(5, -2, -1)	5.40	6.66	1.41	7.18	3.41	0.20	21.48	10.81	1.41
(-1, -5, -3)	2.40	6.27	2.12	0.61	2.73	0.27	0.72	2.23	0.53
(3, 0, 1)	5.02	6.24	0.77	3.21	0.99	0.94	7.66	16.23	4.27
(1, 2, -4)	0.13	4.23	4.48	3.18	0.37	2.77	6.71	10.27	3.55
(0, 5, 2)	0.86	0.99	2.97	2.20	1.87	1.18	2.39	11.00	6.96
(-3, 4, 0)	1.63	3.45	0.41	3.57	2.04	0.25	2.33	3.93	1.67
(-5, 1, 4)	3.36	2.93	4.19	8.86	2.06	3.06	5.93	0.16	9.01
(4, -1, 5)	5.83	3.71	4.64	1.12	1.61	4.09	10.51	8.93	11.47
(-2, -3, -2)	2.57	6.09	1.25	0.77	1.56	0.17	2.92	2.72	1.42
Avg	2.77	4.37	2.80	3.91	2.03	1.55	6.17	7.19	4.69

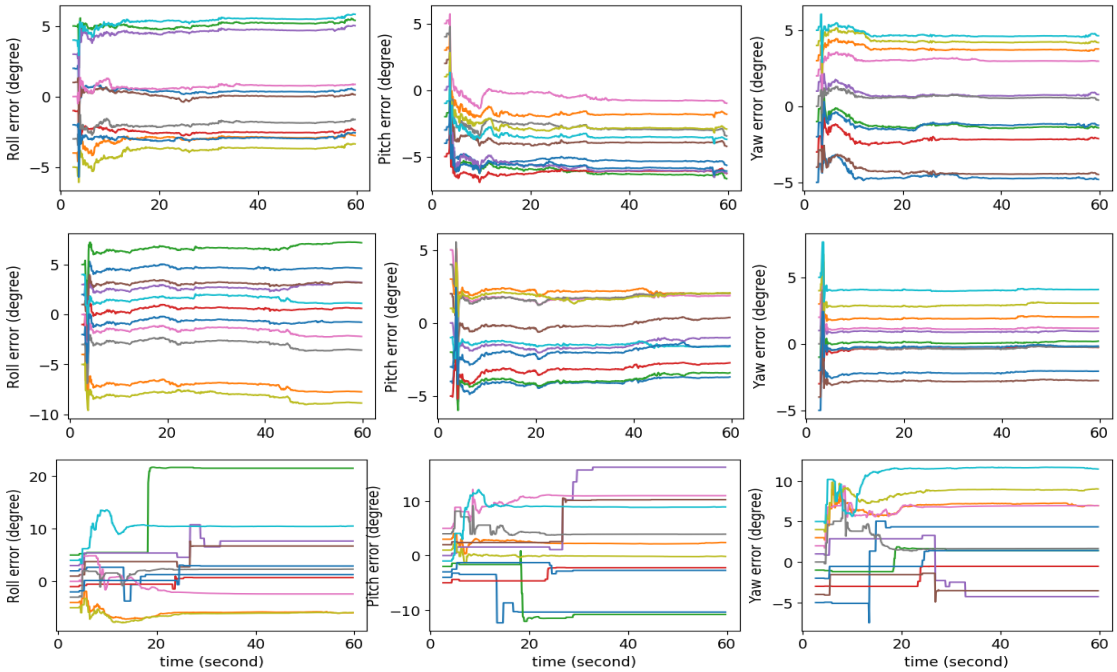


Figure 5.8: Calibration results for pure VIO system undergoes pure translational straight line motion with constant velocity. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds. Top to bottom corresponds to Case-1 to Case-3 in Sec. 5.7.2.

Table 5.8: Final calibration results of the rotational extrinsic parameter for global-pose aided VIO system undergoes pure translational straight line motion with constant velocity. The absolute errors of roll, pitch, and yaw at 60s, are recorded with different perturbations.

Perturbations of (roll, pitch, yaw)	Case-1			Case-2			Case-3		
	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw
(2, -4, -5)	1.90	0.01	0.01	4.38	1.83	0.01	0.30	0.13	0.11
(-4, 3, 3)	4.56	0.00	0.02	2.39	1.00	0.00	8.82	3.22	3.90
(5, -2, -1)	1.88	0.00	0.01	6.66	2.76	0.03	9.69	3.57	4.22
(-1, -5, -3)	4.39	0.01	0.00	2.25	0.95	0.02	4.42	1.67	1.84
(3, 0, 1)	0.70	0.00	0.01	4.33	1.79	0.01	9.39	3.48	4.11
(1, 2, -4)	1.07	0.00	0.00	2.45	1.02	0.03	4.05	1.52	1.76
(0, 5, 2)	0.09	0.01	0.01	0.56	0.22	0.01	6.75	2.53	2.95
(-3, 4, 0)	3.47	0.01	0.01	1.46	0.61	0.01	4.06	1.52	1.76
(-5, 1, 4)	6.06	0.00	0.02	2.67	1.12	0.01	9.84	3.58	4.33
(4, -1, 5)	1.60	0.00	0.02	4.82	1.98	0.03	14.31	5.16	6.35
(-2, -3, -2)	4.78	0.01	0.01	0.96	0.41	0.02	2.50	0.93	1.04
Avg	2.77	0.01	0.01	2.99	1.24	0.02	6.74	2.48	2.94

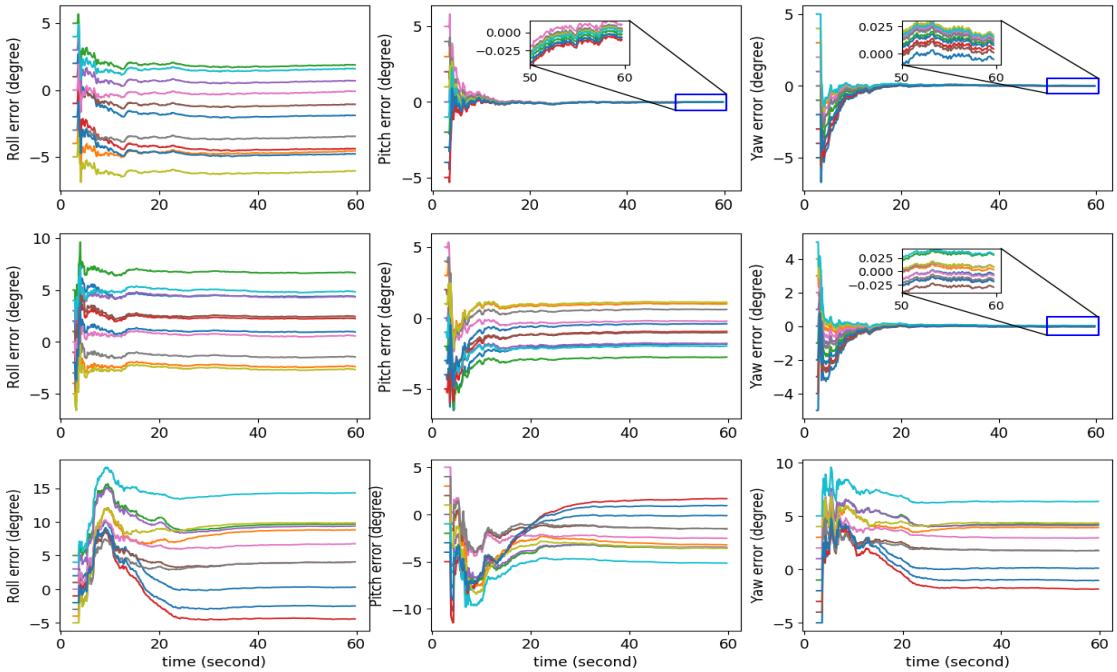


Figure 5.9: Calibration results for global-pose aided VIO system undergoes pure translational straight line motion with constant velocity. y-axis represents errors of the rotational calibration parameter over time respect to different initial guesses. x-axis represents time in seconds. Top to bottom corresponds to Case-1 to Case-3 in Sec. 5.7.2.

Chapter 6

Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Spatial-Temporal Calibration

Accurate state estimation is crucial for various intelligent and autonomous applications, such as robot navigation and augmented reality. Among different multi-sensor fusion options, visual-inertial fusion is widely deployed thanks to its lightweight and low-power characteristics. To bootstrap and achieve optimal state estimation, the spatial-temporal displacements between IMU and cameras must be calibrated in advance. Most existing calibration methods adopt continuous-time state representation, more specifically the B-spline. Despite these methods achieve precise spatial-temporal calibration, they suffer from high computational cost caused by continuous-time state representation. To this end, we propose a novel and extremely efficient calibration method leveraging discrete-time state representation. Experimental results demonstrate that the calibration accuracy of our method is comparable to that of the most popular calibration toolbox, Kalibr. More surprisingly, in terms of optimization time, our method can even accelerate up to 1000x compared to Kalibr. Exten-

sive evaluations show that our calibration method has no adverse effect on the accuracy of visual-inertial odometry (VIO). With the increasing production of cellphones, drones and other visual-inertial platforms, if one million devices need calibration around the world, saving one minute for the calibration of each device means saving 2083 work days in total. To benefit both the research and industry communities, our code will be open-source.

6.1 Related paper

- Song, Junlin, Antoine Richard, and Miguel Olivares-Mendez. "Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Spatial-Temporal Calibration." arXiv preprint arXiv:2509.12846 (2025).

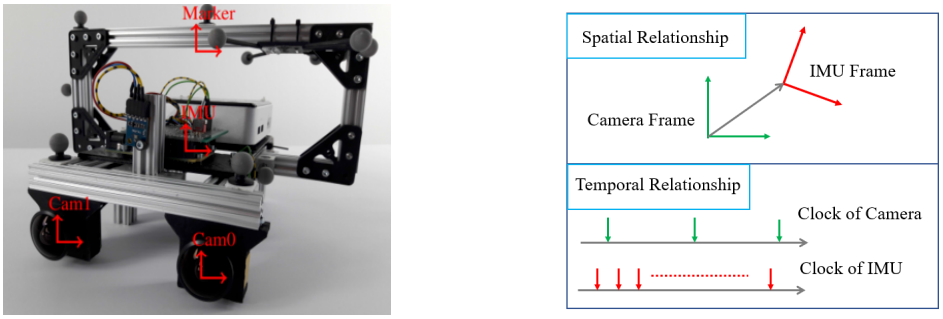
6.2 Relationships to other chapters

Similarly to the principle of target-based method presented in Chapter 3, an efficient calibration approach is developed in this chapter to estimate the spatial-temporal displacements between IMU and cameras. Offline calibration for IMU-Camera system should be done in advance to support Chapter 4 and Chapter 5, as these calibration parameters are indispensable for VIO system.

6.3 Introduction

State estimation is a fundamental research topic in the robotics and computer vision communities. There have been tremendous advances over the past few decades, from single sensor to multi-sensor fusion. In practice, the use of a single sensor may be limited by inherent flaws, such as scale ambiguity in monocular simultaneous localization and mapping (SLAM). Different sensors can complement each other, thus significantly improving the

overall localization and perception capability. Among different multi-sensor fusion schemes, visual-inertial fusion has attracted great attention, as visual-inertial sensor suite has several advantages: small size, low power consumption, and low cost. Nowadays, visual-inertial odometry (VIO) is widely used in AR/VR [12, 4], robotics [5, 6, 7, 93], and planetary exploration [10, 11, 94].



(a) (b)

Figure 6.1: (a) Stereo visual-inertial sensor prototype of the TUM-VI dataset [61]. (b) The spatial-temporal relationship between IMU and camera.

The successful running of a VIO system relies on the good quality of initialization, and the first prerequisite for visual-inertial initialization is the spatial-temporal calibration for IMU and cameras (see Fig. 6.1). The spatial calibration parameter plays the role of aligning the coordinate frames for different sensor measurements. The temporal calibration parameter aligns different clocks, which timestamp the measurements. Temporal calibration is especially critical when strict hardware synchronization is unavailable.

To address the calibration problem for IMU and cameras, extensive studies have been conducted in the literature, from theory to practice. Currently, almost all IMU-Camera calibration methods employ a continuous-time state representation based on the B-spline. This type of method can obtain accurate and consistent calibration results with the aid of a calibration board, and the representative work is termed **Kalibr**, developed by [25]. However, Kalibr suffers from high computational cost due to its B-spline based state representation. To reduce computational complexity, [27] further derive a novel and efficient derivative calculation method for the B-spline on Lie groups [55]. As IMU-Camera calibration method proposed by [27] has been integrated into Basalt project [66, 39, 27], hereafter, we also use

Table 6.1: State dimensions comparison of different calibration methods on the EuRoC [60] and TUM-VI [61] calibration sequences. For TUM-VI dataset, $imu1$ is used as an example sequence here. Image frequency is decreased from 20hz to 10hz, and 5hz.

Dataset	Duration (s)	Methods	Image frequency (hz)		
			20	10	5
EuRoC	71.9	Kalibr [25]	64888	64888	64888
		Basalt [27]	43256	43256	43256
		Ours	12747	6375	3198
TUM-VI	51.9	Kalibr [25]	46846	46846	46846
		Basalt [27]	31232	31232	31232
		Ours	9345	4683	2352

Basalt to refer to the calibration method presented in [27].

Except for continuous-time state representation, discrete-time state representation can also be applied to the spatial-temporal IMU-Camera calibration task. Surprisingly, there has been rare exploration in this direction in the decade after the release of the Kalibr toolbox. Many researchers believe that discrete-time state representation is difficult or inferior for temporal calibration [25, 56, 57]. For example, authors of Kalibr, [25] are concerned that discrete-time state representation requires a new state at each measurement time, which could be challenging for the utilization of high-frequency IMU measurements, and subsequent estimator design for temporal calibration.

In fact, this concern can be addressed by aggregating IMU measurements over a short period of time. Inspired by IMU preintegration [58, 59], we propose a novel optimization-based IMU-Camera calibration method with discrete-time state representation. Several IMU measurements between two consecutive images are aggregated as one pseudo-measurement, thus greatly reducing the state dimensions that need to be optimized (see TABLE 6.1).

MVIS [29] is another discrete-time calibration method based on IMU preintegration, with appealing full calibration capability. However, due to the use of a gravity-aligned reference frame, MVIS sacrifices efficiency by introducing 3D feature positions in the state vector. Instead, our method eliminates features from the state vector by adopting a reference frame similar to Kalibr and Basalt, thus fully unleashing the efficiency power of discrete-time calibration (see TABLE 6.2). More differences between MVIS and our method are further illus-

trated in Section 6.4.

Key contributions of this work can be summarized as:

- We propose a novel IMU-Camera calibration method based on discrete-time state representation. Our method jointly determines the spatial-temporal calibration parameters between IMU and cameras, IMU motion states (poses, velocities, and biases), and gravity. To the best of our knowledge, this is the first method that performs the gravity estimation with IMU preintegration model.
- The significance of Midpoint-based IMU preintegration for time offset estimation is highlighted, which has not been revealed in previous work. Our derivations with vivid graphs greatly mitigate the difficulty of understanding the preintegration process, which could be easily extended to other high-frequency sensors.
- Extensive experimental results demonstrate that our method offers unparalleled efficiency compared to the currently available implementations adopted continuous-time state representation, while maintaining competitive calibration accuracy. Moreover, our method does not cause accuracy loss for VIO.
- We will open-source code to facilitate the benchmark between discrete-time state representation and continuous-time state representation in other state estimation tasks, and to benefit the research and industry communities.

6.4 Related work

Over the past two decades, there have been numerous studies on VIO techniques dedicated to provide 6 degree-of-freedom (DoF) motion tracking in unknown GPS-denied or GPS-degraded environments. Stereo VIO is commonly used in practice, as it offers a more stable scale metric than monocular VIO [5, 11]. Existing Stereo VIO methods can be broadly divided into two categories, optimization-based methods and filter-based methods. Optimization-based stereo VIO include OKVIS [85], Basalt [39] and ORB-SLAM3 [40].

Table 6.2: Comparisons between representative offline IMU-Camera calibration methods and our method.

Method	Year	Continuous?	Target-based?	Constant IMU biases?	Open-Source?	Optimization efficiency
Kalibr[25]	2013	Continuous	Target-based	No	Yes	baseline
Basalt [27]	2020	Continuous	Target-based	Yes	Yes	approximately 10x faster than Kalibr
MVIS [29]	2024	Discrete	Target-based	No	No [a]	approximately 3x faster than Kalibr
iKalibr [30]	2025	Continuous	Target-less	Yes	Yes	Slower than Kalibr due to heavy SfM [95]
Ours	2025	Discrete	Target-based	Yes	Yes [b]	(300-1000)x faster than Kalibr

[a] The relevant code of MVIS [29] (`https://github.com/yangyulin/mvis.git`) is unavailable.
[b] Our code will be available after the paper publication.

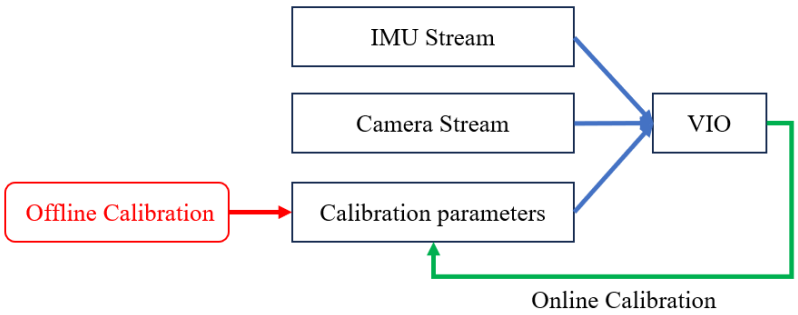


Figure 6.2: Offline calibration and online calibration for VIO applications. The focus of this paper is highlighted in red. Please note, some VIO estimators do not have online calibration functionality, for example Basalt [39], ORB-SLAM3 [40] and SchurVINS [4].

Filter-based stereo VIO include ROVIO [86], Multi-State Constraint Kalman Filter (MSCKF) [41, 16], and SchurVINS [4]. To avoid deteriorating the state estimation performance of VIO, it is vital to feed accurate spatial-temporal calibration parameters for the estimator, to align IMU and camera measurements.

The spatial-temporal calibration between multiple sensors is essentially a state estimation problem. Observability determines whether an interested state can be recovered through available measurements. Profound theoretical research has been conducted on the observability of spatial parameters between IMU and cameras. With the help of artificial visual features on a calibration board, [22], and [23], conclude that spatial parameters are observable when the moving platform undergoes at least 2DoF rotational excitation. An interesting corollary from [22] is that the observability of spatial parameters is independent of translational excitation. These two theoretical works indicate that practitioners need to apply sufficient rotational excitation to the visual-inertial sensor suite. However, [22] and [23] cannot handle the case of temporal misalignment, as IMU and cameras are assumed to be precisely synchronized. To address temporal misalignment issue, [24], and [35], propose online temporal calibration for IMU and monocular camera, based on MSCKF and fixed-lag sliding window optimization, respectively. Online target-less calibration [24, 35, 16] typically requires good initial guess, otherwise the state estimator could be easily converged to local minima or even diverged, especially for spatial calibration parameters. Moreover, online calibration may suffer from motion degradation [19].

Compared to the online target-less calibration, offline taget-based calibration has the potential to offer more accurate and consistent results, benefitting from full-batch optimization and geometric prior of the calibration board. Hence, offline target-based calibration is indispensable for many applications (see Fig. 6.2), especially the commercial product, such as cellphones, drones, autonomous vehicles, and AR glasses. Currently, the most popular offline taget-based calibration toolbox is Kalibr [25]. This work represents the estimated state as a continuous function of time, more specifically the B-spline. One disadvantage of Kalibr is its high computational cost. In order to reduce computational complexity and accelerate optimization, [27] propose a novel formulation to speed up time derivatives and Jacobian

computation for B-spline on Lie groups. In addition, some studies [96, 97] have investigated whether it is better to represent state as one spline in $SE(3)$, or rather use a split representation of two splines in $\mathbb{R}^3$ and $SO(3)$. In general, split representation is recognized better, in terms of trajectory representation and computational efficiency. Motivated by [96, 97], [27] employ split representation in their implementation for further acceleration. Experimental results in [27] prove that this split representation indeed improves computational efficiency.

For more applications of the B-spline and other continuous-time state representation, readers are referred to [57]. On the one hand, continuous-time state representation has been increasingly employed in robotic state estimation over the last years, following the efficient B-spline developed by [27]. On another hand, compared to continuous-time state representation, discrete-time state representation showcases competitive or even better state estimation performance on some public benchmarks, for example, the odometry leaderboard[1] of KITTI [98], and the SLAM challenge leaderboard[2] of HILTI [99]. Those excellent works in leaderboards demonstrate that discrete-time state representation is still promising for many applications, even though it does not have the generality as continuous-time state representation. Meanwhile, we notice in the supplementary material of [100] that the IMU-LiDAR calibration using the discrete-time state representation exhibits superior efficiency performance than the method based on continuous-time state representation [77].

The above observations on different state representations prompt us to consider if it is possible to leverage discrete-time state representation to implement a more efficient calibration toolbox for IMU and cameras. This novel toolbox is supposed to be matchable in accuracy compared to its competitors [25, 27], but superior in efficiency. The efficiency advantage of discrete-time state representation is likely to be achieved with the advance of IMU preintegration on Lie groups [101, 58, 59, 102, 103]. By aggregating IMU measurements between two consecutive images, state dimensions become linearly dependent on the number of images used for optimization, instead of the number of IMU measurements. As a counterpart, state dimensions of the continuous-time state representation based methods,

[1] https://www.cvlibs.net/datasets/kitti/eval_odometry.php
[2] https://hilti-challenge.com/leader-board-2023.html

Kalibr [25] and Basalt [27], are linearly dependent on the number of IMU measurements (much more than images), as IMU raw measurements are utilized to construct residuals for optimization. [27] make some effort to reduce the state dimensions. In practice, a data sequence used for the offline IMU-Camera calibration typically takes about 1 minute. By assuming IMU biases remain unchanged in a short period of time, IMU biases are modeled from time-varying state variables in Kalibr [25], to time-invariant state variables in Basalt [27]. Unfortunately, limited by mathematical formulation, both Kalibr [25] and Basalt [27] are unable to reduce the state dimensions to the extent of our method (see TABLE 6.1).

The closest to our work is MVIS [29], which is a discrete-time calibration method based on IMU preintegration. Like many existing IMU preintegration approaches [59, 39, 102, 103], [29] adopts Euler integration scheme. According to a recent benchmark study [56] and our extensive experiments in Section 6.7, merely using Euler integration can not make discrete-time method comparable to continuous-time method in terms of time offset estimation. To address this issue, we further improve IMU preintegration with Midpoint integration. To the best of our knowledge, this is the first work to reveal the significance of Midpoint integration for time offset estimation. Moreover, our method repeats the integration for each iterative step, which is different from all previous preintegration methods. This strategy improves the accuracy of IMU constraints by decreasing the linearization error due to the change of IMU biases, with little computing cost. Benefitting from these schemes, our method could overcome a larger time offset (150 ms) than what was reported in MVIS (5 ms).

The second difference between [29] and our method is the initialization of the IMU-Camera extrinsics. In [29], these parameters need to be decided manually. In contrast, all calibration parameters are automatically initialized in our method, without extra manual effort.

The third difference is the choice of reference coordinate frame for the state vector. [29] uses gravity-aligned reference frame, like many VIO estimators. This choice eliminates the necessity of estimating the gravity direction. However, 3D feature positions become unknown with respect to this frame. Inserting them into the state vector increases the dimensions for optimization. There are two main disadvantages:

- The efficiency of optimization is decreased due to larger state dimensions. According to [29], the speedup of [29] over Kalibr is only around 3x. In contrast, our method can even accelerate up to 1000x.

- The estimator of [29] leads to 4 unobservable directions, i.e., the global yaw rotation and the global translation [76]. However, our method does not have these unobservable directions, as our reference frame is attached to the calibration board, and all 3D feature positions are known in this frame. For detailed observability analysis, we refer interested readers to [22].

In contrast to [22, 29], we develop a novel discrete-time IMU constraint, which allows gravity estimation with the IMU preintegration model, thus relaxing the strictly gravity-aligned setup for the calibration board [22], and eliminating all 3D feature positions from the state vector [29].

As Basalt [27] is about 10x faster than Kalibr, it can be inferred that [29] is not faster than the fastest continuous-time method. While, our efficiency outperforms Basalt by a large margin. To the best of our knowledge, our method is the first discrete-time spatial-temporal calibration method that can truly surpass existing continuous-time methods in efficiency, unleashing the efficiency power of discrete-time state representation. Comparisons between representative offline IMU-Camera calibration methods and our method are provided in TABLE 6.2.

One more difference is the camera measurement model. [29] uses interpolation model that requires two bounding IMU poses. This choice may not be applicable for low-frequency cameras, such as 5Hz, with a gap of up to 200 ms between two images. Instead, we introduce a different model approximated by constant velocity motion. With the convergence of time offset, this approximation becomes more and more accurate. For experiments, our model can perfectly handle low-frequency camera measurements, which was not thoroughly tested in [29].

The aim of this work is to explore the efficiency boundary of offline target-based approach. Therefore, the joint calibration for all parameters [29] is not our current focus. Intrin-

sic parameters of IMU and cameras are assumed to be precalibrated individually, like [30]. Observability analysis also beyonds our scope, as sufficient motion excitations are assumed to be guaranteed for target-based approach. Differentiate from the online calibration during the running of VIO, the trajectory during offline calibration is typically pre-designed with full excitations on all degree of freedoms, especially for factory calibration. Thus, degenerate motion profiles could be avoided with under-controlled trajectories. Instead, for online calibration, degenerate motions could occur or even dominate, such as pure translational motion for ground vehicles and drones. To avoid potential risks, online calibration is not integrated for some VIO estimators, for example Basalt [39], ORB-SLAM3 [40] and SchurVINS [4].

6.5 Notation

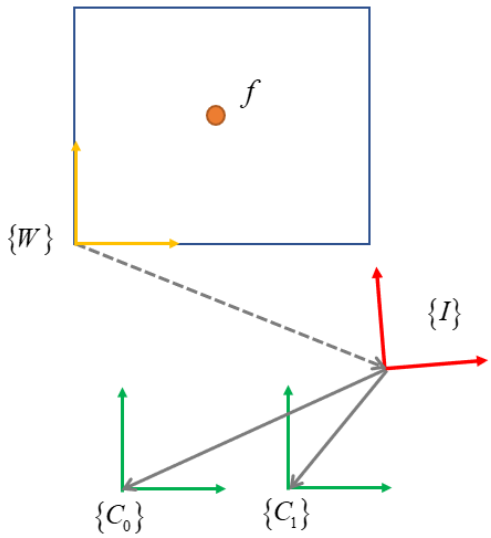


Figure 6.3: Coordinate frames for the IMU-Camera calibration with a calibration board.

Our calibration method is illustrated with an example of spatial-temporal calibration between IMU and stereo camera. Fig. 6.3 shows all coordinate frames involved in the calibration. $\{W\}$ represents the world reference frame attached to the calibration board, which is

static during calibration. $\{I\}$ represents IMU coordinate frame, $\{C_0\}$ and $\{C_1\}$ denote left and right camera coordinate frames, respectively. $\{I\}$, $\{C_0\}$ and $\{C_1\}$ are assumed to be rigidly linked.

We use ${}^{W}(\bullet)$ to represent a physical quantity in the frame $\{W\}$. The position of a point I in the frame $\{W\}$ is expressed as ${}^{W}p_I$. The velocity of a point I in the frame $\{W\}$ is expressed as ${}^{W}v_I$. The local angular velocity of $\{I\}$ is denoted as ω. A rotation matrix is employed to represent the rotation of a rigid body. ${}^{W}_{I}R$ represents rotation from frame $\{I\}$ to frame $\{W\}$. ${}^{W}_{I}T$ represents 6DoF rigid body transformation from frame $\{I\}$ to frame $\{W\}$

$$ {}^{W}_{I}T = \begin{bmatrix} {}^{W}_{I}R & {}^{W}p_I \\ 0 & 1 \end{bmatrix}, {}^{W}_{I}R \in SO(3), {}^{W}_{I}T \in SE(3) \tag{6.1} $$

The 6DoF transformations between two camera frames and IMU frame are noted as spatial calibration parameters $\left\{{}^{I}_{C_0}T, {}^{I}_{C_1}T\right\}$. In our formulation, IMU time clock is treated as time reference in the estimator. The stereo camera is assumed to be already time-synchronized. The time offset only exists for camera clock and IMU clock, which is the temporal calibration parameter t_d. If the image timestamp at camera clock is t_C, then the corresponding timestamp at IMU clock should be shifted with time offset

$$ t_I = t_C + t_d \tag{6.2} $$

The interested spatial-temporal calibration parameter set involved in this problem setting is $\left\{{}^{I}_{C_0}T, {}^{I}_{C_1}T, t_d\right\}$. $[\bullet]_\times$ is denoted as the skew symmetric matrix corresponding to a three-dimensional vector. The transpose of a matrix is $[\bullet]^T$.

6.6 Methodology

Traditional IMU-Camera calibration methods typically uses continuous-time state representation (B-spline) and constructs IMU measurement model using raw measurements. Due to the high frequency of IMU measurements, continuous-time state representation leads to

high-dimensional state and expensive computational cost. The key idea and main contribution of this paper is to leverage discrete-time state representation to explore how far we can accelerate the spatial-temporal calibration between IMU and cameras, without compromising accuracy. This exploration is of great significance with the vast usage of visual-inertial sensors. Inspired by on-manifold IMU preintegration [58, 59], multiple IMU raw measurements between two adjacent images are aggregated into a single pseudo-measurement, significantly reducing the state dimensions of discrete-time state representation. The only concern is the impact on calibration accuracy, which could be addressed by a novel design presented in Section 6.6.1.

Like most target-based calibration methods (Kalibr [25] and Basalt [27]), we use a grid of AprilTag [64] as the calibration board, as shown in Fig. 6.7. The coordinate frames involved in calibration are depicted in Fig. 6.3. The timestamp of the ith image is t_i. The image coordinate of the lth AprilTag corner f_l detected in the ith image of the nth camera is ${}^{n}u_{il}$. Its associated 3D coordinates in $\{W\}$, ${}^{W}p_{f_l}$, is known as geometric prior from the calibration board. When performing calibration, it is required to wave the sensor rig in front of a calibration board and apply sufficient motion excitation, especially in rotation [22, 23]. The optimization variable set χ of our calibration method can be defined as a vector of several discrete-time state variables

$$
\begin{aligned}
\chi &= \begin{bmatrix} x_I^T & x_{calib}^T \end{bmatrix}^T \\
x_I &= \begin{bmatrix} x_{I_0}^T & \cdots & x_{I_i}^T & \cdots & x_{I_M}^T \end{bmatrix}^T \\
x_{I_i} &= \begin{bmatrix} {}^{W}_{I_i}R^T & {}^{W}v_{I_i}^T & {}^{W}p_{I_i}^T \end{bmatrix}^T \\
x_{calib} &= \begin{bmatrix} {}^{I}_{C_0}T^T & {}^{I}_{C_1}T^T & t_d & b_\omega^T & b_a^T & \theta & \phi \end{bmatrix}^T
\end{aligned}
\tag{6.3}
$$

Where M is the index of the last image. χ includes the IMU motion states at different image timestamps, x_I, as well as the calibration state, x_{calib}. IMU motion state corresponding to the ith image x_{I_i} includes both pose and velocity $\{{}^{W}_{I_i}R, {}^{W}v_{I_i}, {}^{W}p_{I_i}\}$, which are expressed in world frame $\{W\}$. Calibration state x_{calib} contains both a set of spatial-temporal calibration parameters $\{{}^{I}_{C_0}T, {}^{I}_{C_1}T, t_d\}$, and IMU biases $\{b_\omega, b_a\}$. b_ω and b_a represent the gyroscope

bias and the accelerometer bias, respectively. As the duration of a calibration sequence is typically short (about 1 minute, see TABLE 6.1), IMU biases can be assumed to be time-invariant, like Basalt [27] and iKalibr [30], as shown in TABLE 6.2.

As gravity is included in accelerometer measurement model, and unknown in our case, it is necessary to estimate gravity ${}^{W}g$, which is also expressed in world frame $\{W\}$. The norm of gravity can be assumed to be known and remain constant[3], thus ${}^{W}g \in S^2$. We use straightforward and simple spherical coordinate to parameterize gravity

$$ {}^{W}g \triangleq {}^{W}g\left(\rho,\theta,\phi\right) = \rho \begin{bmatrix} \cos\left(\theta\right)\sin\left(\phi\right) \\ \sin\left(\theta\right)\sin\left(\phi\right) \\ \cos\left(\phi\right) \end{bmatrix} \tag{6.4} $$

By utilizing prior knowledge of gravity norm, $\rho = \left\|{}^{W}g\right\| = 9.81 m/s^2$, the dimension of gravity to be optimized can be reduced from 3 to 2. Therefore, we only include $\{\theta, \phi\}$ in x_{calib}. For other parameterization methods of element on S^2, interested readers are referred to [104, 86, 105].

6.6.1 IMU pseudo-measurement model

To achieve the goal of reducing the state dimension, multiple IMU measurements between two adjacent images are aggregated into a single pseudo-measurement, which is accomplished by IMU preintegration. On-manifold IMU preintegration is originally designed for a VIO estimator [58, 59]. In this section, we implement a novel on-manifold IMU preintegration to cater for the calibration task. Our pseudo-measurement model has several differences from many existing IMU preintegration approaches [59, 39, 102, 103]:

1. These existing preintegration models adopt Euler integration. We empirically find it is not sufficient to ensure that our calibration accuracy is comparable to Kalibr [25] and

[3] With the increasing interest and investment on the robotic exploration of extraterrestrial space [10, 11, 94], IMU-Camera calibration maybe executed on other celestial body in the future, like the Moon or Mars. The gravity norm should be adapted accordingly. If gravity norm is unknown, full 3DoF gravity can be easily included in Eq. 6.3.

Basalt [27]. Therefore, we derive Midpoint integration to improve integration accuracy, producing more accurate IMU constraints.

2. These existing preintegration models do not need repeated IMU integration during the iteration process of optimization, because IMU and camera are assumed to be time-synchronized. In contrast, our IMU pseudo-measurement model designed for joint spatial-temporal calibration inevitably requires integration from scratch, as temporal calibration can lead to the shift of image timestamps, thus changing the IMU integration interval, as shown in Fig. 6.4.

3. These existing preintegration models do not optimize gravity with the residual model (see Eq. 6.18), because gravity is assumed to be known at the VIO initialization stage. While, our IMU pseudo-measurement model supports gravity optimization, addressing the unknown gravity direction with respect to the world frame built on the calibration board.

4. Lastly, these existing preintegration models require different IMU biases for different IMU factors. While, our IMU pseudo-measurement model supports the same IMU biases for all IMU factors, further reducing the state dimensions that need to be optimized.

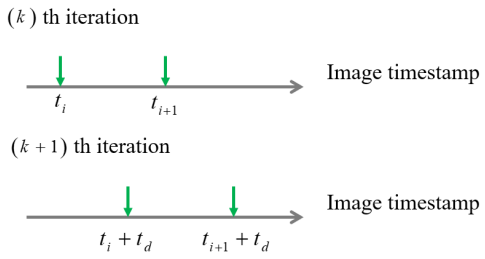


Figure 6.4: Time shift of images due to the time offset t_d between camera and IMU.

Next, we will detail how to build our IMU pseudo-measurement model, as well as the formulation difference with existing preintegration models. [59] is referred as an example. All IMU measurements from frame i to frame $i+1$ are collected for integration, with an

integration interval of $[t_i, t_{i+1}]$. The IMU measurements at two image timestamps, t_i and t_{i+1}, can be obtained by linear interpolation if necessary. The preintegration items [59] are connected with two IMU motion states. And this connection is used to build IMU constraints. Therefore, preliminary knowledge about preintegration items should be introduced at the beginning. The IMU preintegration items of frame i are denoted as

$$\Delta_{i,i+1} = \begin{bmatrix} \Delta R_{i,i+1} \\ \Delta v_{i,i+1} \\ \Delta p_{i,i+1} \end{bmatrix} \tag{6.5}$$

By equivalently modifying the equation (33) from [59], the calculation of preinegration items over $[t_i, t_{i+1}]$ can be expressed as an iterative formulation

$$\begin{aligned} \Delta R_{i,j+1} &= \Delta R_{i,j} Exp\left(\omega_j \Delta t\right) \\ \Delta v_{i,j+1} &= \Delta v_{i,j} + \Delta R_{i,j} a_j \Delta t \\ \Delta p_{i,j+1} &= \Delta p_{i,j} + \Delta v_{i,j} \Delta t + \tfrac{1}{2} \Delta R_{i,j} a_j \Delta t^2 \end{aligned} \tag{6.6}$$

Where Δt is the time interval between two consecutive IMU measurements, $\Delta t = t_{j+1} - t_j$. And t_j represents IMU sampling timestamp, starting from $t_j = t_i$. During iteration, t_j subjects to

$$t_i \leq t_j < t_{j+1} \leq t_{i+1} \tag{6.7}$$

ω_j and a_j in Eq. 6.6 represent de-biased IMU angular velocity and linear acceleration measurements

$$\begin{aligned} \omega_j &= \tilde{\omega}_j - b_\omega \\ a_j &= \tilde{a}_j - b_a \end{aligned} \tag{6.8}$$

$\tilde{\omega}_j$ and $\tilde{a}_j$ are raw IMU angular velocity and linear acceleration measurements, respec-

tively. The initial value of $\Delta_{i,i+1}$ is set as

$$\Delta_{i,i} = \begin{bmatrix} \Delta R_{i,i} \\ \Delta v_{i,i} \\ \Delta p_{i,i} \end{bmatrix} = \begin{bmatrix} I_{3\times3} \\ 0_{3\times1} \\ 0_{3\times1} \end{bmatrix} \tag{6.9}$$

Eq. 6.6 is used for iterative calculation until $t_{j+1} = t_{i+1}$. At this time $\Delta_{i,i+1}$ is determinate. Please note, this iterative process is essentially Euler integration [59]. During the integration interval $[t_j, t_{j+1}]$, only the IMU measurement at t_j is used. The formulation adopted in [39] is different from Eq. 6.6. For example, the rotation integration is calculated as

$$\Delta R_{i,j+1} = \Delta R_{i,j} Exp\left(\omega_{j+1}\Delta t\right) \tag{6.10}$$

But in this case, only the IMU measurement at t_{j+1} is used. In order to improve the integration accuracy, an effective idea is to adopt a higher order integration method, such as Midpoint integration. Specifically, the average value of two IMU measurements is utilized to better approximate the integration. The corresponding iterative equation is improved from Eq. 6.6 to

$$\begin{aligned} \Delta R_{i,j+1} &= \Delta R_{i,j} Exp\left(\bar{\omega}_{j,j+1}\Delta t\right) \\ \Delta v_{i,j+1} &= \Delta v_{i,j} + \bar{a}_{j,j+1}\Delta t \\ \Delta p_{i,j+1} &= \Delta p_{i,j} + \Delta v_{i,j}\Delta t + \tfrac{1}{2}\bar{a}_{j,j+1}\Delta t^2 \end{aligned} \tag{6.11}$$

Where $\bar{\omega}_{j,j+1}$ and $\bar{a}_{j,j+1}$ denote the average IMU angular velocity and linear acceleration measurements

$$\begin{aligned} \bar{\omega}_{j,j+1} &= \tfrac{1}{2}\left(\omega_j + \omega_{j+1}\right) \\ \bar{a}_{j,j+1} &= \tfrac{1}{2}\left(\Delta R_{i,j} a_j + \Delta R_{i,j+1} a_{j+1}\right) \end{aligned} \tag{6.12}$$

For a clear mathematical model, Eq. 6.11 and Eq. 6.12 are abstracted as a function

$$\Delta_{i,j+1} = f\left(\Delta_{i,j}, \tilde{\omega}_j, \tilde{\omega}_{j+1}, \tilde{a}_j, \tilde{a}_{j+1}, b_\omega, b_a\right) \tag{6.13}$$

Using the simpler form above, the Jacobian of $\Delta_{i,j+1}$ with respect to IMU biases can be

clearly expressed with a recursive formulation

$$\begin{aligned}\frac{\partial \Delta_{i,j+1}}{\partial b_\omega} &= \frac{\partial f}{\partial \Delta_{i,j}} \frac{\partial \Delta_{i,j}}{\partial b_\omega} + \frac{\partial f}{\partial b_\omega} \\ \frac{\partial \Delta_{i,j+1}}{\partial b_a} &= \frac{\partial f}{\partial \Delta_{i,j}} \frac{\partial \Delta_{i,j}}{\partial b_a} + \frac{\partial f}{\partial b_a}\end{aligned} \tag{6.14}$$

Please note, the index in the above equation runs from j to $j+1$. The corresponding initial Jacobian is set as

$$\begin{aligned}\frac{\partial \Delta_{i,i}}{\partial b_\omega} &= 0_{9\times 3} \\ \frac{\partial \Delta_{i,i}}{\partial b_a} &= 0_{9\times 3}\end{aligned} \tag{6.15}$$

We adopt left perturbation to calculate Jacobian for the element on Lie group [55]. With the iteration of Eq. 6.11, IMU noise generated from $\{\Delta_{i,j}, \tilde{\omega}_j, \tilde{\omega}_{j+1}, \tilde{a}_j, \tilde{a}_{j+1}\}$ (see Eq. 6.13) is propagated as

$$\begin{aligned}\Sigma_{i,j+1} &= \left(\frac{\partial f}{\partial \Delta_{i,j}}\right) \Sigma_{i,j} \left(\frac{\partial f}{\partial \Delta_{i,j}}\right)^T \\ &+ \left(\frac{\partial f}{\partial \tilde{\omega}_j}\right) \Sigma_\omega \left(\frac{\partial f}{\partial \tilde{\omega}_j}\right)^T + \left(\frac{\partial f}{\partial \tilde{\omega}_{j+1}}\right) \Sigma_\omega \left(\frac{\partial f}{\partial \tilde{\omega}_{j+1}}\right)^T \\ &+ \left(\frac{\partial f}{\partial \tilde{a}_j}\right) \Sigma_a \left(\frac{\partial f}{\partial \tilde{a}_j}\right)^T + \left(\frac{\partial f}{\partial \tilde{a}_{j+1}}\right) \Sigma_a \left(\frac{\partial f}{\partial \tilde{a}_{j+1}}\right)^T\end{aligned} \tag{6.16}$$

Where Σ_ω and Σ_a represent noise covariance for IMU angular velocity and linear acceleration measurements. Given this iterative formulation, the covariance of $\Delta_{i,i+1}$, $\Sigma_{i,i+1}$, is obtained together when the integration of $\Delta_{i,i+1}$ is finished. The initial value of $\Sigma_{i,i+1}$ is set as

$$\Sigma_{i,i} = 0_{9\times 9} \tag{6.17}$$

To avoid distracting from the model, the analytical on-manifold Jacobian of f (Eq. 6.13) with respect to all involved variables, $\{\Delta_{i,j}, \tilde{\omega}_j, \tilde{\omega}_{j+1}, \tilde{a}_j, \tilde{a}_{j+1}, b_\omega, b_a\}$, are provided in Appendix 6.10. Given these derivatives, Eq. 6.14 and Eq. 6.16 become computable now. The whole integration process is summarized in Algorithm 1.

The relationship between $\Delta_{i,i+1}$ and two consecutive IMU states (x_{I_i} and $x_{I_{i+1}}$, see Eq. 6.3) models motion constraint via IMU integration. The corresponding IMU residuals[4] are

[4]The formulation of our residual is different from the equation (45) in [59], as we do not incorporate IMU biases update described in Section VI.C of [59]. Instead, we perform reintegration for each iteration (see Fig. 6.4).

Algorithm 1: Midpoint-based IMU preintegration

Input: Two image timestamps (t_i and t_{i+1}), current estimate of the time offset t_d between camera and IMU

Output: Preintegrated terms ($\Delta_{i,i+1}$), preintegrated Jacobians ($\frac{\partial \Delta_{i,i+1}}{\partial b_\omega}$, $\frac{\partial \Delta_{i,i+1}}{\partial b_a}$), and preintegrated covariance ($\Sigma_{i,i+1}$)

`/* Time shift of image timestamps due to the time offset` t_d `*/`

$t_i \leftarrow t_i + t_d$;

$t_{i+1} \leftarrow t_{i+1} + t_d$;

Obtain a set of $\{\omega_j, a_j\}$, subjects to Eq. 6.7;

Initialize relevant variables, according to Eq. 6.9, Eq. 6.15, and Eq. 6.17;

while $t_{j+1} \leq t_{i+1}$ **do**

  update preintegrated items for residual evaluation, according to Eq. 6.12 and Eq. 6.11

  update preintegrated Jacobians for IMU biases estimation, according to Eq. 6.14

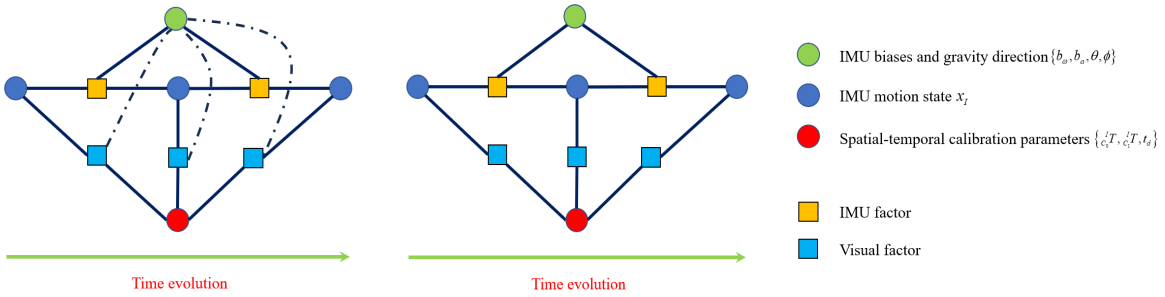
  update preintegrated covariance for residual weight, according to Eq. 6.16

end

constructed as

$$
\begin{aligned}
r_{\Delta R_{i,i+1}} &= Log\left(\Delta R_{i,i+1} {}^{W}_{I_{i+1}}R^T {}^{W}_{I_i}R\right) \\
r_{\Delta v_{i,i+1}} &= {}^{W}_{I_i}R^T \left({}^{W}v_{I_{i+1}} - {}^{W}v_{I_i} - {}^{W}g dt\right) - \Delta v_{i,i+1} \\
r_{\Delta p_{i,i+1}} &= {}^{W}_{I_i}R^T \left({}^{W}p_{I_{i+1}} - {}^{W}p_{I_i} - {}^{W}v_{I_i} dt - \tfrac{1}{2}{}^{W}g dt^2\right) \\
&\quad -\Delta p_{i,i+1} \\
r_{I_{i,i+1}} &\triangleq r_{I_{i,i+1}}\left(x_{I_i}, x_{I_{i+1}}, b_\omega, b_a, \theta, \phi\right) = \begin{bmatrix} r_{\Delta R_{i,i+1}} \\ r_{\Delta v_{i,i+1}} \\ r_{\Delta p_{i,i+1}} \end{bmatrix}
\end{aligned}
\tag{6.18}
$$

Where $Log\left(\bullet\right)$ maps the element on a Lie group to the tangent space vector [55]. dt is the time gap between two images, $dt = t_{i+1} - t_i$. For the Jacobian of residual $r_{I_{i,i+1}}$ with respect to x_{I_i} and $x_{I_{i+1}}$, we adopt left perturbation for the element on Lie group, as previous derivatives of f (Eq. 6.13). The Jacobian of $r_{I_{i,i+1}}$ with respect to IMU biases, $\{b_\omega, b_a\}$, can be obtained from Eq. 6.14. Finally, the Jacobian of $r_{I_{i,i+1}}$ with respect to the gravity direction



(a) (b) (c)

Figure 6.5: (a) Factor graph of our calibration method, if b_ω is included in the state variable set of camera measurement model. The dashed dot line illustrates the influence of b_ω on the factor graph. (b) Factor graph of our calibration method, if b_ω is excluded from the state variable set of camera measurement model. (c) State variables are denoted as circles. Measurement factors are denoted as squares. IMU factor and visual factor are detailed in Section 6.6.1 and Section 6.6.2, respectively.

$\{\theta, \phi\}$ can be computed according to the chain rule

$$\begin{aligned} g_s &= \begin{bmatrix} \theta & \phi \end{bmatrix}^T \\ \frac{\partial r_{I_{i,i+1}}}{\partial g_s} &= \frac{\partial r_{I_{i,i+1}}}{\partial {}^W g} \frac{\partial {}^W g}{\partial g_s} \end{aligned} \tag{6.19}$$

Up until now, we complete the mathematical details of the IMU pseudo-measurement model between two adjacent frames. By using vivid graphs in Appendix 6.10, we greatly mitigate the difficulty of understanding the preintegration process, enhancing the readability of the paper.

6.6.2 Camera measurement model with time offset

The IMU pseudo-measurement model described in the previous section can only constrain IMU states, IMU biases, and gravity direction (see Eq. 6.3 and Eq. 6.18). And it is unable to link to interested spatial-temporal calibration parameters. To this end, we introduce a camera measurement model with time offset, to provide additional constraints on both the spatial and temporal calibration parameters

$${}^n r_{il} = \pi \left({}^I_{C_n} T^{-1} {}^W_{I_i} T (t_i + t_d)^{-1} {}^W p_{f_l} \right) - {}^n u_{il} \tag{6.20}$$

$\pi(\bullet)$ is a fixed camera projection function [66, 67]. The camera intrinsic parameters are assumed to be pre-calibrated. ${}^{n}r_{il}$ represents the pixel residual (reprojection error) obtained by the nth camera, observing the lth AprilTag corner at the ith frame. The corresponding measurement covariance is typically set by engineering experience, such as fixed at 1 pixel.

Eq. 6.20 models the effect of time offset t_d on the IMU state. Due to temporal misalignment, the IMU pose corresponding to the ith frame is shifted from ${}^{W}_{I_i}T(t_i)$ to ${}^{W}_{I_i}T(t_i + t_d)$. According to Eq. 6.1, the notation of ${}^{W}_{I_i}T(t_i)$ is

$$
{}^{W}_{I_i}T(t_i) = {}^{W}_{I_i}T = \begin{bmatrix} {}^{W}_{I_i}R & {}^{W}p_{I_i} \\ 0 & 1 \end{bmatrix} \tag{6.21}
$$

Similarly, the expression of ${}^{W}_{I_i}T(t_i + t_d)$ is

$$
\begin{aligned}
{}^{W}_{I_i}T(t_i + t_d) &= \begin{bmatrix} {}^{W}_{I_i}R(t_i + t_d) & {}^{W}p_{I_i}(t_i + t_d) \\ 0 & 1 \end{bmatrix} \\
{}^{W}_{I_i}R(t_i + t_d) &= {}^{W}_{I_i}R Exp(\omega_i t_d) \\
{}^{W}p_{I_i}(t_i + t_d) &= {}^{W}p_{I_i} + {}^{W}v_{I_i} t_d
\end{aligned} \tag{6.22}
$$

ω_i is the IMU angular velocity at t_i, and can be directly obtained by the de-biased gyroscope measurement (see Eq. 6.8). ${}^{W}v_{I_i}$ is the IMU linear velocity at t_i, including in the state vector (Eq. 6.3). ${}^{W}_{I_i}T(t_i + t_d)$ is approximated by a constant velocity motion model.

By analyzing Eq. 6.20 and Eq. 6.22, state variables linked to the camera measurement model can be obtained. These state variables are denoted with a set

$$
x_s = \{ {}^{I}_{C_n}T, {}^{W}_{I_i}R, {}^{W}p_{I_i}, b_\omega, {}^{W}v_{I_i}, t_d \} \tag{6.23}
$$

We empirically find that excluding b_ω and ${}^{W}v_{I_i}$ from x_s does not influence the convergence of b_ω (see Fig. 6.8, Fig. 6.9 and Fig. 6.10), but greatly simplifies the factor graph model for calibration, from Fig. 6.5a to Fig. 6.5b. Therefore, new x_s becomes

$$
x_s = \{ {}^{I}_{C_n}T, {}^{W}_{I_i}R, {}^{W}p_{I_i}, t_d \} \tag{6.24}
$$

The analytical on-manifold Jacobian of ${}^{n}r_{il}$ with respect to all involved variables in x_s are provided in Appendix 6.11.

6.6.3 Full-batch nonlinear least squares optimization

By integrating all raw image pixel measurements and IMU pseudo-measurements, we formulate the full-batch nonlinear least squares optimization as

$$\begin{aligned}\chi = \arg\min \Bigg\{ & \sum_{n=0}^{1} \sum_{i=0}^{M} \sum_{l \in K_{ni}} \rho \left(\left\| {}^{n}r_{il} \right\|_{\Sigma_C}^2 \right) \\ & + \sum_{i=0}^{M-1} \left\| r_{I_{i,i+1}} \right\|_{\Sigma_{i,i+1}}^2 \Bigg\} \\ {}^{n}r_{il} = \pi & \left({}^{I}_{C_n}T^{-1} {}^{W}_{I_i}T (t_i + t_d)^{-1} {}^{W}p_{f_l} \right) - {}^{n}u_{il} \\ r_{I_{i,i+1}} & \triangleq r_{I_{i,i+1}} \left(x_{I_i}, x_{I_{i+1}}, b_\omega, b_a, \theta, \phi \right)\end{aligned} \tag{6.25}$$

Where M is the index of the last image. ${}^{n}r_{il}$ is the pixel residual from the nth camera. $r_{I_{i,i+1}}$ is the IMU pseudo-measurement residual. Detailed definitions of these two types of residuals are provided in Section 6.6.1 and Section 6.6.2. K_{ni} represent the set of corner points observed by the nth camera at the ith frame. $\rho(\bullet)$ is a robust kernel function [65]. Typically, robust Huber kernel function is used to mitigate the impact of pixel observation outliers.

The factor graph for this nonlinear least squares optimization is shown in Fig. 6.5b. Levenberg-Marquardt algorithm is adopted to minimize Eq. 6.25 and update the optimal estimation iteratively.

At the end of each iteration, the timestamps of all images are shifted with the current estimate of time offset

$$t_i \leftarrow t_i + t_d \tag{6.26}$$

As shown in Fig. 6.4, the integration interval for each IMU pseudo-measurement at the next iteration should be updated accordingly

$$[t_i, t_{i+1}] \leftarrow [(t_i + t_d), (t_{i+1} + t_d)] \tag{6.27}$$

Since IMU motion states (x_I in Eq. 6.3) are dependent on image time, they are shifted together in time domain with the update of time offset, as depicted in Fig. 6.6.

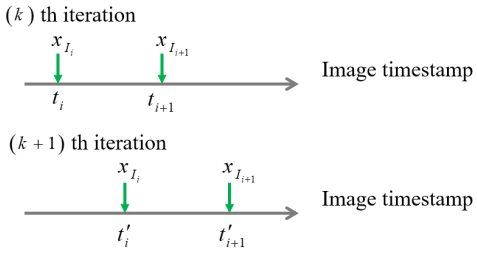


Figure 6.6: Time shift of each IMU motion state corresponding to image. After the time shift of images, t_i and t_{i+1} become t'_i and t'_{i+1}, respectively. $t'_i = t_i + t_d$, $t'_{i+1} = t_{i+1} + t_d$.

6.6.4 State initialization

Nonlinear least squares optimization requires setting reasonable initial guess for state variables to ensure convergence. Firstly, we compute initial 6DoF camera trajectory with PnP algorithm

$$\left\{ {}^{W}_{C_{ni}}T | n \in \{0,1\}, i \in \{0, \cdots, M\} \right\} \tag{6.28}$$

n is the camera index and i is the image index. Leveraging the initial estimation of camera rotation, the camera angular velocity can be obtained by differentiating two consecutive estimates. Furthermore, by aligning the camera angular velocity with the IMU angular velocity, the initial rotational extrinsic parameter is produced

$$\min_{{}^{I}_{C_n}R} \sum_{i} \left\| {}^{I}_{C_n}R\omega_{ni} - \tilde{\omega}_{I_i} \right\| \tag{6.29}$$

Where ω_{ni} denotes the angular velocity of the nth camera at the ith frame. $\tilde{\omega}_{I_i}$ is the corresponding IMU angular velocity measurement at t_i. This estimate is rough because it ignores the influence of gyroscope bias (see Eq. 6.8), and the differential error from ω_{ni}.

The initial translational extrinsic parameter is set as zero. Based on the initial 6DoF extrinsic parameter between IMU and camera, and camera trajectory, the initial pose of an

IMU motion state (x_{I_i}, see Eq. 6.3) is calculated as

$$ {}^{W}_{I_i}T = {}^{W}_{C_{ni}}T\left({}^{I}_{C_n}T\right)^{-1} \tag{6.30} $$

The initial velocity of an IMU motion state is inferred by differentiating two consecutive IMU poses. Like the initialization of translational extrinsic parameter, the initial IMU biases and time offset are set to zero.

According to the accelerometer measurement model

$$ \tilde{a} = {}^{W}_{I_i}R^{-1}\left({}^{W}a_{I_i} - {}^{W}g\right) + b_a \tag{6.31} $$

Where ${}^{W}a_{I_i}$ is the acceleration of IMU frame at t_i. By assuming ${}^{W}a_{I_i} \approx 0$ and using the initial guess of b_a, a rough estimate of gravity can be obtained via

$$ {}^{W}g = -{}^{W}_{I_i}R\left(\tilde{a} - b_a\right) = -{}^{W}_{I_i}R\tilde{a} \tag{6.32} $$

Then, ${}^{W}g$ is normalized with the prior knowledge about gravity norm ($9.81m/s^2$)

$$ {}^{W}g = \frac{9.81}{\|{}^{W}g\|}{}^{W}g \tag{6.33} $$

According to Eq. 6.4, $\{\theta, \phi\}$ can be initialized. At this point, we complete the initialization for all state variables.

As the focus of this work is on estimator design for calibration, we perform relatively coarse initialization. However, this simple scheme works effectively in all of our experiments. The only concerned point is the extremely large time offset, for example 1.0 second. Possible improvements on the initialization of temporal and rotational calibration parameters are discussed in Section 6.8.

6.7 Results

We benchmark the calibration accuracy and computational efficiency of the proposed method with two state-of-the-art (SOTA) baseline methods, **Kalibr** [25] and **Basalt** [27], both of which adopt continuous-time state representation. Compared to Kalibr, Basalt accelerates calibration through split B-spline representation and advanced derivative calculation method. Default optimized configurations for them are used for fair comparison, like the experiments in MVIS [29]. To demonstrate the performance of our method, experiments are designed to address the following questions

1. Is the spatial-temporal calibration accuracy of the proposed method comparable with Kalibr and Basalt?
2. Is the optimization time of the proposed method much less than that of Kalibr and Basalt?
3. As illustrated in Section 6.6.1, do we need to replace Euler integration with Midpoint integration?
4. Is the proposed method robust to large IMU biases?
5. Is the proposed method robust to large time offset?
6. Does VIO incur accuracy loss by using the proposed calibration method?

Question 1) and 2) are used to evaluate if our method achieves the key objective. Question 3) and 4) are designed to demonstrate the effectiveness of the IMU pseudo-measurement model (Section 6.6.1). While question 5) is designed to validate the effectiveness of the camera measurement model with time offset (Section 6.6.2). Question 6) can demonstrate whether our calibration method has an adverse effect on the accuracy of VIO.

We perform experiments with IMU-Camera calibration sequences from three popular VIO datasets, EuRoC [60], TUM-VI [61] and UZH-FPV [62]. EuRoC and TUM-VI datasets provide stereo image and IMU data at 20Hz and 200Hz, respectively. For the UZH-FPV

Table 6.3: Timestamp shifting for IMU data.

Perturbation	Value (ms)
δt_1	50
δt_2	40
δt_3	30
δt_4	20
δt_5	10
δt_6	0
δt_7	-10
δt_8	-20
δt_9	-30
δt_{10}	-40
δt_{11}	-50

dataset, the acquisition frequencies of the camera and the IMU are 30Hz and 500Hz, respectively. As shown in Fig. 6.7, both indoor and outdoor scenarios have been taken into account. To demonstrate the calibration performance at different camera frequencies, new calibration sequences are generated by reducing the original image frequency from 20Hz (or 30Hz) to 10Hz, and 5Hz.

The reference value of the temporal calibration parameter can be obtained from the dataset providers. It is convenient to reset the desired time offset by manually shifting the timestamp of IMU data with a certain value (see TABLE 6.3). The shifted time offset modifies the new reference value of the temporal calibration parameter. The reference value of the spatial calibration parameter is obtained by calibrating the original sequence (20Hz or 30Hz image) with Kalibr.

The proposed calibration method is evaluated with two variants with different IMU integration methods for the IMU pseudo-measurement model (Section 6.6.1). If the IMU pseudo-measurement model uses Euler integration, our method is referred to as **"Ours (Euler)"**. Specifically, the integration model in [39] is used. If Midpoint integration (see Eq. 6.12 and Eq. 6.11) is employed, our method is referred to as **"Ours (Midpoint)"**.

All the experiments are conducted on a laptop computer with an Intel(R) Xeon(R) W-10855M CPU @ 2.80GHz, and 16 GB of RAM.

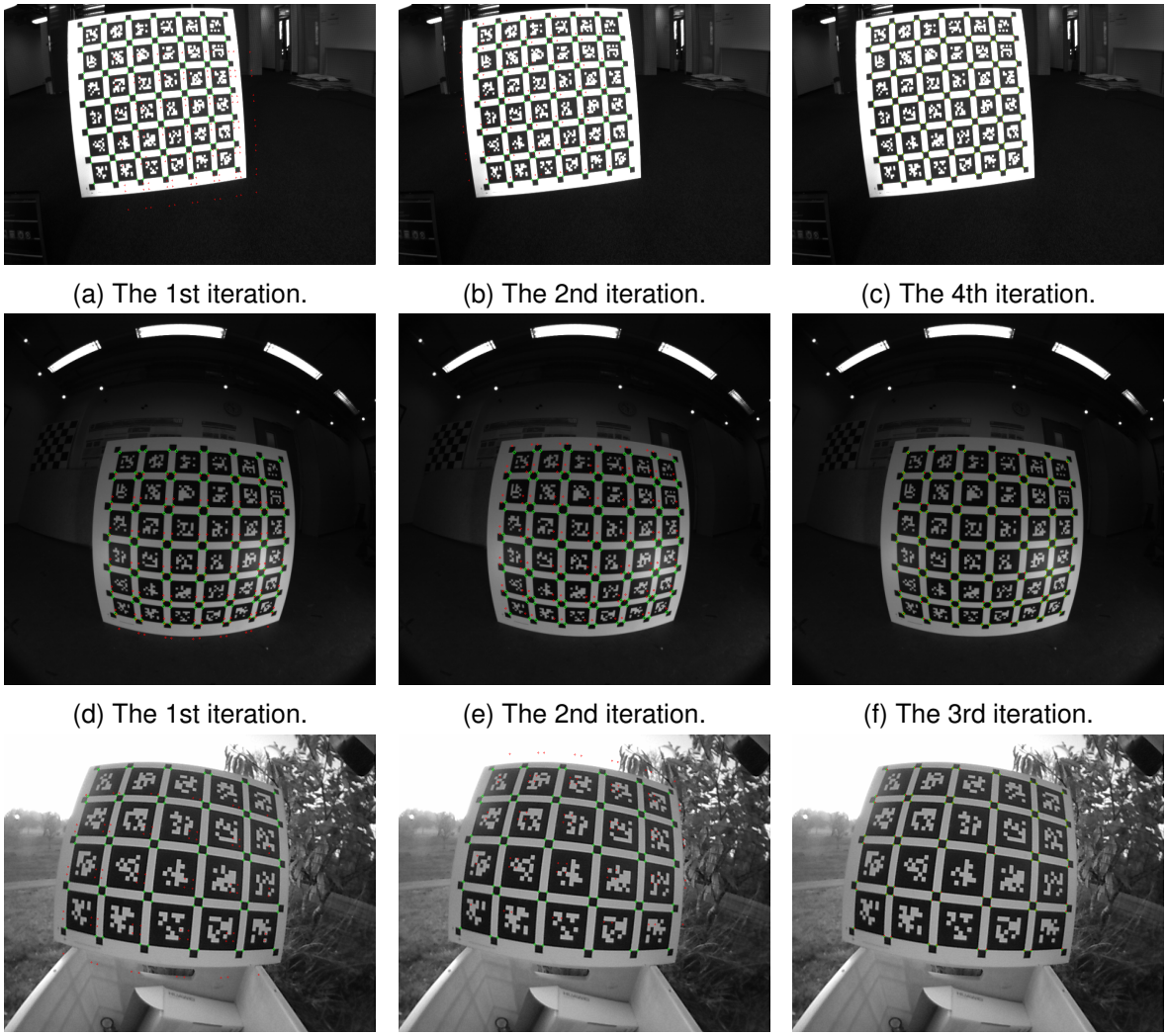
Table 6.4: Average metrics of different calibration methods on the EuRoC dataset. Evaluation metrics include the average RMSE results of spatial-temporal calibration (rotation, translation, time offset), reprojection error, optimization time and speed up of our method compared to SOTA baselines.

Metrics (unit)	Methods	20 Hz		10 Hz		5 Hz	
		Camera C_0	Camera C_1	Camera C_0	Camera C_1	Camera C_0	Camera C_1
Rotation (degree)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.009 ± 0.000	0.009 ± 0.000	0.024 ± 0.000	0.024 ± 0.000
	Basalt	0.497 ± 0.622	0.502 ± 0.621	0.486 ± 0.621	0.492 ± 0.620	0.437 ± 0.649	0.444 ± 0.648
	Ours (Euler)	0.012 ± 0.000	0.023 ± 0.000	0.014 ± 0.000	0.028 ± 0.000	0.090 ± 0.000	0.103 ± 0.000
	Ours (Midpoint)	0.015 ± 0.000	0.014 ± 0.000	0.009 ± 0.000	0.015 ± 0.000	0.041 ± 0.000	0.047 ± 0.000
Translation (cm)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.025 ± 0.000	0.025 ± 0.000	0.078 ± 0.000	0.078 ± 0.000
	Basalt	0.870 ± 1.117	0.870 ± 1.116	0.798 ± 1.054	0.798 ± 1.052	0.506 ± 0.604	0.506 ± 0.602
	Ours (Euler)	0.045 ± 0.000	0.049 ± 0.000	0.073 ± 0.000	0.075 ± 0.000	0.301 ± 0.000	0.296 ± 0.000
	Ours (Midpoint)	0.039 ± 0.000	0.048 ± 0.000	0.039 ± 0.000	0.050 ± 0.000	0.047 ± 0.000	0.058 ± 0.000
Time offset (ms)	Kalibr	0.035 ± 0.000		0.044 ± 0.000		0.066 ± 0.000	
	Basalt	12.971 ± 17.839		11.397 ± 15.590		10.529 ± 16.550	
	Ours (Euler)	2.456 ± 0.000		2.449 ± 0.000		2.483 ± 0.000	
	Ours (Midpoint)	0.043 ± 0.000		0.068 ± 0.000		0.158 ± 0.000	
Reprojection error (pixel)	Kalibr	0.373 ± 0.000		0.375 ± 0.000		0.374 ± 0.000	
	Basalt	0.263 ± 0.071		0.266 ± 0.072		0.246 ± 0.060	
	Ours (Euler)	0.208 ± 0.000		0.210 ± 0.000		0.217 ± 0.000	
	Ours (Midpoint)	0.209 ± 0.000		0.211 ± 0.000		0.213 ± 0.000	
Optimization time (s)	Kalibr	144.170 ± 0.837		108.134 ± 0.496		51.255 ± 0.419	
	Basalt	14.919 ± 2.495		15.064 ± 2.284		15.730 ± 3.027	
	Ours (Euler)	0.289 ± 0.004		0.136 ± 0.015		0.189 ± 0.338	
	Ours (Midpoint)	0.290 ± 0.004		0.140 ± 0.012		0.081 ± 0.001	
Speedup of Ours (Midpoint) compared to Kalibr		497.138x		772.386x		632.778x	
Speedup of Ours (Midpoint) compared to Basalt		51.445x		107.600x		194.198x	

6.7.1 EuRoC dataset

For each sequence with specific image frequency, we shift the timestamp of IMU data from -50ms to 50ms, with 10ms increment. In this way, the number of calibration sequences for each image frequency is increased to 11, as shown in TABLE 6.3. The average RMSE results (rotation, translation, time offset) of spatial-temporal calibration with different methods are summarized under different image frequencies in TABLE 6.4. When the image frequency is 20Hz, the rotation and translation RMSE of Kalibr is 0 degree and 0 cm, respectively. This is because the reference value of spatial calibration parameter is obtained via Kalibr itself. The time offset RMSE of Kalibr is only 0.035 ms. With the decrease of image frequency, Kalibr's spatial-temporal calibration results gradually deviate from the reference value. This deviation is very small. Specifically, the rotation deviation is less than 0.05 degree, the translation deviation is less than 0.1 cm, and the time offset deviation is less than 0.1 ms. These results demonstrate Kalibr's excellent and reliable calibration accuracy.

Another baseline, Basalt, exhibits large RMSE results for spatial-temporal calibration.



(a) The 1st iteration. (b) The 2nd iteration. (c) The 4th iteration.

(d) The 1st iteration. (e) The 2nd iteration. (f) The 3rd iteration.

(g) The 1st iteration. (h) The 2nd iteration. (i) The 4th iteration.

Figure 6.7: (a-c) Representative image from EuRoC dataset. (d-f) Representative image from TUM-VI dataset. (g-i) Representative image from UZH-FPV dataset. Expected corner positions (green) and predicted corner positions (red) in the image. Please zoom in 300% to view convergence details.

Table 6.5: Average metrics of different calibration methods on the TUM-VI dataset. Evaluation metrics include the average RMSE results of spatial-temporal calibration (rotation, translation, time offset), reprojection error, optimization time and speed up of our method compared to SOTA baselines.

Metrics (unit)	Methods	20 Hz		10 Hz		5 Hz	
		Camera C_0	Camera C_1	Camera C_0	Camera C_1	Camera C_0	Camera C_1
Rotation (degree)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
	Basalt	0.002 ± 0.000	0.003 ± 0.000	0.002 ± 0.000	0.003 ± 0.000	0.002 ± 0.000	0.003 ± 0.000
	Ours (Euler)	0.002 ± 0.000	0.003 ± 0.000	0.003 ± 0.000	0.002 ± 0.000	0.004 ± 0.000	0.002 ± 0.000
	Ours (Midpoint)	0.003 ± 0.000	0.002 ± 0.000	0.004 ± 0.000	0.001± 0.000	0.005 ± 0.000	0.001 ± 0.000
Translation (cm)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.007 ± 0.000	0.007 ± 0.000	0.013 ± 0.000	0.013 ± 0.000
	Basalt	0.069 ± 0.000	0.077 ± 0.000	0.069 ± 0.000	0.077 ± 0.000	0.069 ± 0.000	0.077 ± 0.000
	Ours (Euler)	0.011 ± 0.000	0.007 ± 0.000	0.015 ± 0.000	0.009 ± 0.000	0.020 ± 0.000	0.014 ± 0.000
	Ours (Midpoint)	0.017 ± 0.000	0.014 ± 0.000	0.021 ± 0.000	0.017 ± 0.000	0.027 ± 0.000	0.022 ± 0.000
Time offset (ms)	Kalibr	0.165 ± 0.000		0.163 ± 0.000		0.167 ± 0.000	
	Basalt	0.168 ± 0.000		0.168 ± 0.000		0.168 ± 0.000	
	Ours (Euler)	2.341 ± 0.000		2.341 ± 0.000		2.339 ± 0.000	
	Ours (Midpoint)	0.165 ± 0.000		0.164 ± 0.000		0.165 ± 0.000	
Reprojection error (pixel)	Kalibr	0.085 ± 0.000		0.085 ± 0.000		0.085 ± 0.000	
	Basalt	0.087 ± 0.000		0.087 ± 0.000		0.087 ± 0.000	
	Ours (Euler)	0.087 ± 0.000		0.087 ± 0.000		0.087 ± 0.000	
	Ours (Midpoint)	0.087 ± 0.000		0.088 ± 0.000		0.088 ± 0.000	
Optimization time (s)	Kalibr	100.420 ± 5.842		63.454 ± 4.071		44.640 ± 2.465	
	Basalt	5.890 ± 0.651		5.901 ± 0.686		5.959 ± 0.899	
	Ours (Euler)	0.201 ± 0.013		0.093 ± 0.011		0.049 ± 0.005	
	Ours (Midpoint)	0.196 ± 0.020		0.094 ± 0.011		0.050 ± 0.005	
Speedup of Ours (Midpoint) compared to Kalibr		512.347x		675.043x		892.800x	
Speedup of Ours (Midpoint) compared to Basalt		30.051x		62.777x		119.180x	

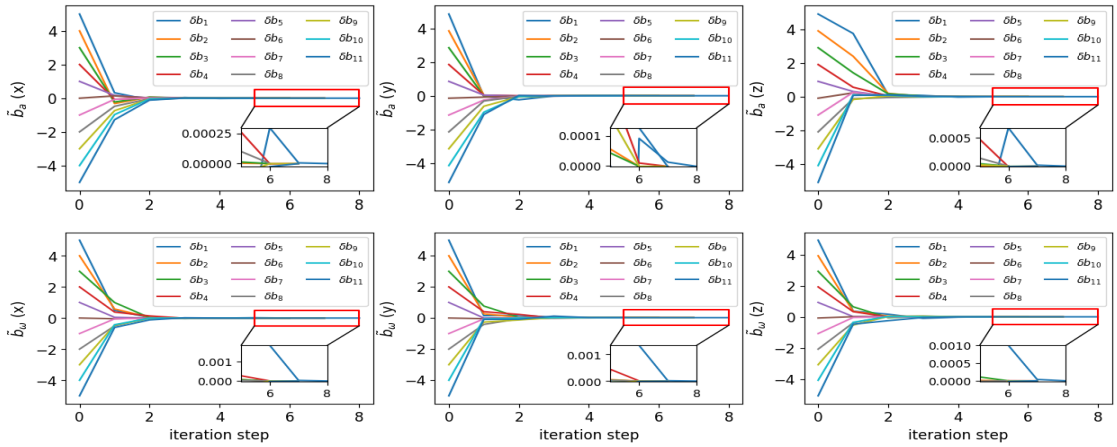


Figure 6.8: Convergence results of the IMU biases under different perturbations (TABLE 6.6) with EuRoC dataset. x-axis represents iteration steps. The units for b_a and b_ω are in m/s^2 and rad/s. The estimation error perfectly approach to zero for each component of IMU biases, with only 8 steps.

Particularly, the time offset RMSE of Basalt is greater than 10 ms for all image frequencies. The large calibration error indicates that Basalt performs unreliable calibration, although its reprojection error is smaller than Kalibr. By carefully examining the implementation of Kalibr and Basalt, we found that the poor performance of Basalt can be attributed to its coarse initialization for time offset. Basalt initializes the time offset to 0 ms. While Kalibr initializes the time offset with cross-correlation algorithm [106], which is able to reduce the initial error of time offset to less than 1 ms typically. This finer initialization strategy provides better initial guess for time offset, making Kalibr robust to different time offsets.

The RMSE results of Ours (Euler) are better than Basalt, and validate that our method is robust to different time offsets, although we initialize the time offset to 0 ms, like Basalt. This may be explained by the lower state dimensions and complexity of our method, as shown in TABLE 6.1. The higher complexity of the whole system may cause the optimization more vulnerable to poor initial guess. We note that the time offset RMSE (around 2.5 ms) of Ours (Euler) is greater than Kalibr, even though it is already less than the measurement period of IMU (5 ms). Time offset estimation issue for discrete-time state representation is also reported in a recent SLAM benchmark study [56], which adopts Euler integration for IMU preintegration [59]. Another variation of our method, Ours (Midpoint), successfully addresses this issue by reducing the time offset RMSE to less than 0.2 ms. In addition, the rotation RMSE of Ours (Midpoint) is less than 0.05 degree, and the translation RMSE of Ours (Midpoint) is less than 0.1 cm. Impressive numerical results demonstrate that the calibration accuracy of Ours (Midpoint) is comparable to Kalibr, and the necessity of upgrading IMU integration from Euler integration to Midpoint integration (Section 6.6.1).

Advantages on efficiency

Apart from the accuracy metric, TABLE 6.4 also shows the efficiency metric, more specifically, the optimization time of each calibration method. When the image frequency is 20Hz, the average optimization time of Kalibr is 144.17 s, which is approximately twice the duration of the calibration sequence (71.9 s). Kalibr decreases the optimization time by reducing the image frequency, because camera measurements become less. The optimization time of

Basalt can be reduced to around 15 s. While Ours (Euler) is much faster than both Kalibr and Basalt, reducing the optimization time to less than 0.3 s. Our method significantly benefits from the lower state dimension, as shown in TABLE 6.1. The optimization time of Ours (Midpoint) is further improved. When the image frequency is 5Hz, the optimization time is even reduced to 0.081 s. On average, Ours (Midpoint) is 634x faster than Kalibr.

Robustness to large time offset

To verify the robustness of Ours (Midpoint) to large time offset, we shift IMU timestamp with 150 ms for the original calibration sequence. The spatial-temporal calibration results are similar to TABLE 6.4. Fig. 6.7 shows the process of predicted corner points approaching expected corner points, visualizing how the reprojection error gradually converges to the sub-pixel level. If the cross-correlation of Kalibr is disabled, both Kalibr and Basalt cannot converge when the time offset exceeds 30 ms. These experiments demonstrate the effectiveness of the camera measurement model (Section 6.6.2), and the larger convergence radius of our estimator for time offset, compared to Kalibr and Basalt.

Robustness to large IMU biases

To verify the robustness of Ours (Midpoint) to large IMU biases, we manually add different bias perturbations to the raw IMU measurements of the original calibration sequence. IMU biases are modified with the given perturbations. Assuming the original IMU biases obtained through Basalt[5] are $\{b_\omega, b_a\}$, the given perturbations are $\{\delta b_\omega, \delta b_a\}$, and the IMU biases obtained from Ours (Midpoint) are $\left\{\hat{b}_\omega, \hat{b}_a\right\}$, then the estimation error of IMU biases is calculated as

$$\begin{aligned} \tilde{b}_\omega &= \hat{b}_\omega - b_\omega - \delta b_\omega \\ \tilde{b}_a &= \hat{b}_a - b_a - \delta b_a \end{aligned} \tag{6.34}$$

As shown in TABLE 6.6, for δb_ω, the perturbation range of each component is designed as $[-5, 5]\, rad/s$. Maximum absolute perturbation value (5 rad/s) is much larger than practical

[5]Both Basalt and our method adopt a time-invariant model for IMU biases.

Table 6.6: Perturbations on IMU biases.

Perturbation	$\delta b_a\ (m/s^2)$			$\delta b_\omega\ (rad/s)$		
	x	y	z	x	y	z
δb_1	5	5	5	5	5	5
δb_2	4	4	4	4	4	4
δb_3	3	3	3	3	3	3
δb_4	2	2	2	2	2	2
δb_5	1	1	1	1	1	1
δb_6	0	0	0	0	0	0
δb_7	-1	-1	-1	-1	-1	-1
δb_8	-2	-2	-2	-2	-2	-2
δb_9	-3	-3	-3	-3	-3	-3
δb_{10}	-4	-4	-4	-4	-4	-4
δb_{11}	-5	-5	-5	-5	-5	-5

value (typically in the level of 1e-2), as we want to challenge the convergence basin of our method. For δb_a, the perturbation range of each component is $[-5, 5]\, m/s^2$. The estimation error of IMU biases under different perturbations is shown in Fig. 6.8. After just 8 iterations, both $\tilde{b}_\omega$ and $\tilde{b}_a$ can converge to zero perfectly. The numerical results of spatial-temporal calibration are similar to TABLE 6.4. These results validate the effectiveness of the IMU pseudo-measurement model (Section 6.6.1), and the robustness of our method to large IMU biases.

Impact on VIO accuracy

To evaluate the influence of spatial-temporal parameters obtained by different calibration methods on the localization accuracy of VIO, a SOTA VIO estimator, Open-VINS [16], is used for the localization experiments. Open-VINS has the capability to perform online spatial-temporal calibration. If online calibration is enabled, Open-VINS is denoted as Open-VINS (w. calib). Otherwise, Open-VINS is denoted as Open-VINS (wo. calib). Absolute trajectory error (ATE) [107] is utilized as an accuracy metric of VIO, and obtained by aligning the estimated trajectory with the groundtruth trajectory in posyaw mode. Results are reported in TABLE 6.7. This table demonstrates that using our calibration method does not cause

Table 6.7: ATE (meter) Comparison on the EuRoC Dataset with different calibration parameters from Kalibr, Basalt and Ours (Midpoint), respectively.

Sequence	Open-VINS (w. calib)			Open-VINS (wo. calib)		
	Kalibr	Basalt	Ours	Kalibr	Basalt	Ours
MH_01_easy	0.056	0.053	0.053	0.035	0.036	0.042
MH_02_easy	0.058	0.057	0.062	0.046	0.046	0.052
MH_03_medium	0.067	0.056	0.052	0.055	0.056	0.050
MH_04_difficult	0.068	0.076	0.069	0.056	0.053	0.048
MH_05_difficult	0.051	0.047	0.055	0.045	0.045	0.047
V1_01_easy	0.106	0.088	0.124	0.133	0.130	0.097
V1_02_medium	0.105	0.071	0.072	0.072	0.079	0.065
V1_03_difficult	0.110	0.183	0.168	0.127	0.131	0.103
V2_01_easy	0.127	0.159	0.139	0.123	0.111	0.102
V2_02_medium	0.370	0.170	0.220	0.182	0.225	0.163
V2_03_difficult	0.364	0.271	0.271	0.226	0.195	0.202
Avg	0.135	0.112	0.117	0.100	0.101	0.088

accuracy loss for VIO, compared to Kalibr or Basalt.

6.7.2 TUM-VI dataset

The collection platform of the TUM-VI dataset is shown in Fig. 6.1. Like the experiments for EuRoC dataset, we perform augmentation operation for calibration sequence by shifting IMU timestamp manually. The average RMSE results of spatial-temporal calibration with different methods are presented in TABLE 6.5. The rotation and translation RMSE results of four methods are very small, and these results show that their accuracy is comparable to each other. Inspecting the temporal calibration results, Ours (Midpoint) is almost the same as Kalibr and Basalt. While Ours (Euler) exhibits a larger RMSE (around 2.3 ms), which once again demonstrates the necessity of Midpoint integration. The reprojection error is approximately 0.09 pixel for all methods.

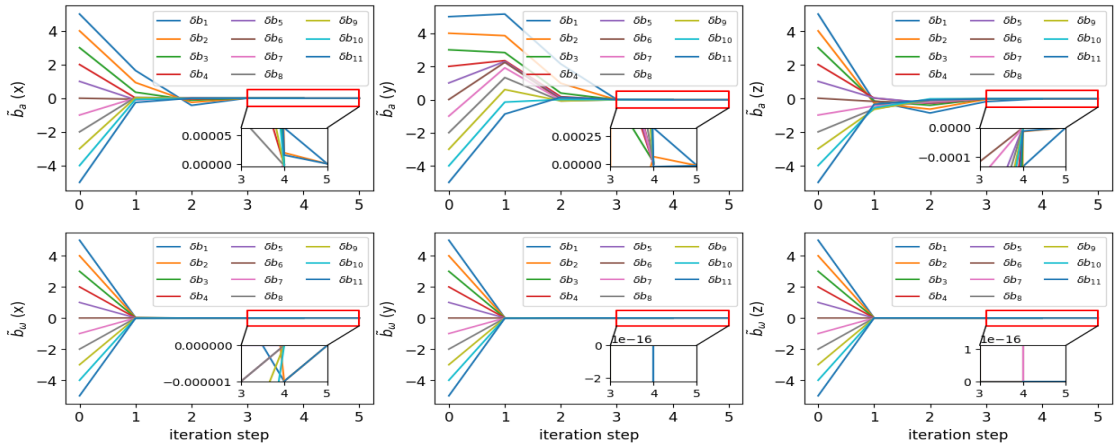


Figure 6.9: Convergence results of the IMU biases under different perturbations (TABLE 6.6) with TUM-VI dataset. x-axis represents iteration steps. The units for b_a and b_ω are in m/s^2 and rad/s. The estimation error perfectly approach to zero for each component of IMU biases, with only 5 steps.

Advantages on efficiency

Comparing the optimization time of different methods, on average, Ours (Midpoint) is 693x faster than Kalibr, and 71x faster than Basalt. TABLE 6.5 shows that our method is comparable to Kalibr and Basalt in accuracy, and has remarkable improvement in efficiency.

Robustness to large time offset and IMU biases

Similar to EuRoC dataset, we perform additional experiments to verify the robustness of our method. Fig. 6.7 visualizes the sub-pixel level convergence of the reprojection error under large time offset (150ms). Fig. 6.9 shows the robustness of Ours (Midpoint) to large IMU biases. All numerical results of spatial-temporal calibration are similar to TABLE 6.5. These results validate the robustness and effectiveness of Ours (Midpoint).

Impact on VIO accuracy

The impact of different calibration methods on VIO accuracy is presented in TABLE 6.8. This table again shows that our calibration method generates comparable VIO accuracy, as

Table 6.8: ATE (meter) Comparison on the TUM-VI Dataset with different calibration parameters from Kalibr, Basalt and Ours (Midpoint), respectively.

Sequence	Open-VINS (w. calib)			Open-VINS (wo. calib)		
	Kalibr	Basalt	Ours	Kalibr	Basalt	Ours
room1	0.058	0.064	0.054	0.052	0.066	0.058
room2	0.090	0.102	0.104	0.055	0.063	0.056
room3	0.083	0.083	0.076	0.092	0.071	0.077
room4	0.030	0.033	0.042	0.036	0.032	0.032
room5	0.089	0.088	0.095	0.109	0.092	0.090
Avg	0.070	0.074	0.074	0.069	0.065	0.063

Kalibr or Basalt.

6.7.3 UZH-FPV dataset

To test the versatility of different calibration methods, in this section, we choose a dataset collected in outdoor environments, UZH-FPV [62]. This dataset aims to benchmark the performance of different VIO algorithms in high-speed agile motion scenarios.

Like in the previous two sections, we perform an augmentation operation for the calibration sequence by shifting the IMU timestamp. The average RMSE results of spatial-temporal calibration with different methods are presented in TABLE 6.9. All four methods show very small RMSE. Basalt exhibits relatively larger RMSE results for spatial-temporal calibration, indicating that Basalt performs unreliable calibration, although its reprojection error is smaller than Kalibr. In terms of the temporal calibration results, Ours (Midpoint) is almost the same as Kalibr. In contrast, Ours (Euler) exhibits a larger RMSE (around 0.9 ms), which once again demonstrates the necessity of Midpoint integration.

Advantages on efficiency

Comparing the optimization time of different methods, on average, Ours (Midpoint) is 797x faster than Kalibr, and 63x faster than Basalt. TABLE 6.9 shows that our method has a remarkable improvement in efficiency. We noticed an unexpected increase in the optimization

Table 6.9: Average metrics of different calibration methods on the UZH-FPV dataset. Evaluation metrics include the average RMSE results of spatial-temporal calibration (rotation, translation, time offset), reprojection error, optimization time and speed up of our method compared to SOTA baselines.

Metrics (unit)	Methods	30 Hz		10 Hz		5 Hz	
		Camera C_0	Camera C_1	Camera C_0	Camera C_1	Camera C_0	Camera C_1
Rotation (degree)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.012 ± 0.000	0.012 ± 0.000	0.015 ± 0.000	0.015 ± 0.000
	Basalt	0.046 ± 0.042	0.050 ± 0.035	0.029 ± 0.012	0.035 ± 0.009	0.033 ± 0.024	0.037 ± 0.015
	Ours (Euler)	0.024 ± 0.000	0.032 ± 0.000	0.026 ± 0.000	0.035 ± 0.000	0.030 ± 0.000	0.031 ± 0.000
	Ours (Midpoint)	0.024 ± 0.000	0.034 ± 0.000	0.030 ± 0.000	0.034 ± 0.000	0.035 ± 0.000	0.028 ± 0.000
Translation (cm)	Kalibr	0.000 ± 0.000	0.000 ± 0.000	0.024 ± 0.000	0.024 ± 0.000	0.042 ± 0.000	0.042 ± 0.000
	Basalt	0.180 ± 0.274	0.198 ± 0.273	0.109 ± 0.061	0.127 ± 0.060	0.176 ± 0.278	0.194 ± 0.276
	Ours (Euler)	0.024 ± 0.000	0.040 ± 0.000	0.034 ± 0.000	0.043 ± 0.000	0.049 ± 0.000	0.054 ± 0.000
	Ours (Midpoint)	0.029 ± 0.000	0.040 ± 0.000	0.044 ± 0.000	0.049 ± 0.000	0.058 ± 0.000	0.060 ± 0.000
Time offset (ms)	Kalibr	0.042 ± 0.000		0.010 ± 0.000		0.007 ± 0.000	
	Basalt	0.404 ± 0.665		0.142 ± 0.174		0.143 ± 0.175	
	Ours (Euler)	0.917 ± 0.000		0.946 ± 0.000		0.954 ± 0.000	
	Ours (Midpoint)	0.068 ± 0.000		0.035 ± 0.000		0.023 ± 0.000	
Reprojection error (pixel)	Kalibr	0.336 ± 0.000		0.336 ± 0.000		0.336 ± 0.000	
	Basalt	0.147 ± 0.004		0.145 ± 0.001		0.146 ± 0.002	
	Ours (Euler)	0.145 ± 0.000		0.146 ± 0.000		0.148 ± 0.000	
	Ours (Midpoint)	0.146 ± 0.000		0.148 ± 0.000		0.150 ± 0.000	
Optimization time (s)	Kalibr	116.184 ± 0.383		143.110 ± 0.873		78.118 ± 0.275	
	Basalt	8.818 ± 2.398		7.833 ± 1.573		8.320 ± 2.260	
	Ours (Euler)	0.356 ± 0.029		0.126 ± 0.007		0.079 ± 0.001	
	Ours (Midpoint)	0.338 ± 0.019		0.132 ± 0.008		0.081 ± 0.005	
Speedup of Ours (Midpoint) compared to Kalibr		343.740x		1084.167x		964.420x	
Speedup of Ours (Midpoint) compared to Basalt		26.089x		59.341x		102.716x	

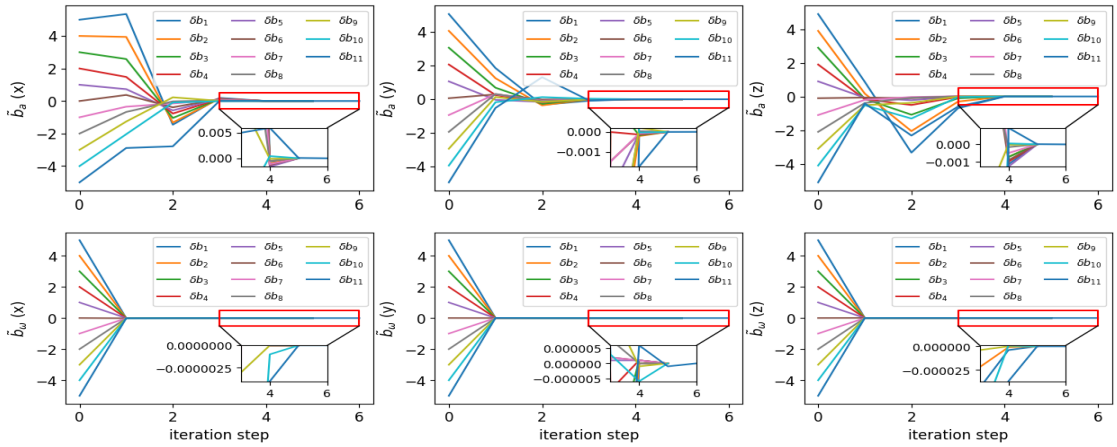


Figure 6.10: Convergence results of the IMU biases under different perturbations (TABLE 6.6) with UZH-FPV dataset. x-axis represents iteration steps. The units for b_a and b_ω are in m/s^2 and rad/s. The estimation error perfectly approach to zero for each component of IMU biases, with only 6 steps.

Table 6.10: ATE (meter) Comparison on the UZH-FPV Dataset with different calibration parameters from Kalibr, Basalt and Ours (Midpoint), respectively.

Outdoor Sequence	Open-VINS (w. calib)			Open-VINS (wo. calib)		
	Kalibr	Basalt	Ours	Kalibr	Basalt	Ours
forward_1	0.532	0.529	0.443	0.445	0.476	0.443
forward_3	1.177	1.000	1.003	1.109	1.018	1.034
forward_5	0.259	0.317	0.294	0.283	0.228	0.259
Avg	0.656	0.615	0.580	0.612	0.574	0.579

[a] The groundtruth trajectories of other outdoor forward facing sequences are unavailable.

time for Kailbr when the image frequency is decreased from 30Hz to 10Hz. This is due to the increased iteration steps of Kalibr. Similarly, Basalt can barely benefit from decreasing camera frequencies in terms of efficiency. More results can be observed from TABLE 6.4 and TABLE 6.5. In contrast, our method consistently improves its optimization time. These results demonstrate that dimension reduction by discrete-time state representation is more favorable in designing more stable nonlinear optimization.

Robustness to large time offset and IMU biases

Similar to EuRoC dataset, we perform additional experiments to verify the robustness of our method. Fig. 6.7 visualizes the sub-pixel level convergence of the reprojection error under large time offset (150ms). Fig. 6.10 shows the robustness of Ours (Midpoint) to large IMU biases. All numerical results of spatial-temporal calibration are similar to TABLE 6.9. These results validate the robustness and effectiveness of Ours (Midpoint).

Impact on VIO accuracy

The impact of different calibration methods on VIO accuracy is presented in TABLE 6.10. This table again shows that our calibration method generates comparable VIO accuracy, as Kalibr or Basalt.

6.8 Discussions

- Proposed calibration method may be further improved by feeding better initial guess to nonlinear least squares optimization. For example, the initialization of time offset can be obtained via one-dimensional cross-correlation [106, 108], like Kalibr [25]. Moreover, initial temporal and rotational calibration parameters could be jointly estimated through the extended three-dimensional cross-correlation presented by [34]. Sufficient rotational excitation must be guaranteed for the successful application of cross-correlation methods.

- IMU pseudo-measurement model may be further improved from the Midpoint integration to the 4th order Runge-Kutta integration. It would be interesting for practitioners by benchmarking these two integration methods in accuracy and efficiency.

6.9 Conclusion

In this work, we propose a novel IMU-Camera calibration method based on discrete-time state representation. Benefitting from this representation, we push the efficiency boundary for the spatial-temporal IMU-Camera calibration. Experimental results demonstrate that our method is extremely faster than the SOTA methods using continuous-time state representation, while maintaining comparable calibration accuracy. This novel method is more favorable for resource-constrained platforms and could accelerate the field test for robots that need VIO localization. Moreover, this method has the potential to empower higher productivity for commercial products that require massive IMU-Camera factory calibration, such as cellphones, drones, autonomous vehicles, and AR glasses.

On the other hand, our method paves the way for the transition from continuous-time state representation to discrete-time state representation. In the future, we plan to extend our work to IMU-LiDAR calibration [77]. We are also interested in benchmarking other state estimation tasks with continuous-time and discrete-time state representations. For example, we notice that the SLAM benchmark in [56] maybe unfair for discrete-time state represen-

tation, as Euler integration is used for IMU preintegration, which could be improved with Midpoint integration as demonstrated in Section 6.7.

Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Calibration

Supplementary Material

6.10 Jacobians of Eq. 6.13

According to Eq. 6.11 and Eq. 6.12, $\bar{\omega}_{j,j+1}$ and $\bar{a}_{j,j+1}$ can be computed as

$$\begin{aligned}
\bar{\omega}_{j,j+1} &= \tfrac{1}{2}\left(\omega_j + \omega_{j+1}\right) \\
\bar{a}_{j,j+1} &= \tfrac{1}{2}\left(\Delta R_{i,j} a_j + \Delta R_{i,j+1} a_{j+1}\right) \\
&= \tfrac{1}{2}\left(\Delta R_{i,j} a_j + \Delta R_{i,j+1} a_{j+1}\right) \\
&= \tfrac{1}{2}\left(\Delta R_{i,j} a_j + \Delta R_{i,j} Exp\left(\bar{\omega}_{j,j+1}\Delta t\right) a_{j+1}\right) \\
&= \tfrac{1}{2}\Delta R_{i,j}\left(a_j + Exp\left(\bar{\omega}_{j,j+1}\Delta t\right) a_{j+1}\right)
\end{aligned} \tag{6.35}$$

The Jacobians of $\bar{\omega}_{j,j+1}$ with respect to ω_j and ω_{j+1} are

$$\begin{aligned}
\tfrac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_j} &= \tfrac{1}{2} I_3 \\
\tfrac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_{j+1}} &= \tfrac{1}{2} I_3
\end{aligned} \tag{6.36}$$

To better understand how the chain rule is performed, we draw the relationship between f (Eq. 6.13) and $\{\Delta_{i,j}, \omega_j, \omega_{j+1}, a_j, a_{j+1}\}$ in Fig. 6.11, Fig. 6.12 and Fig. 6.13.

The Jacobians corresponding to $\bar{a}_{j,j+1}$ (see Fig. 6.12 or Fig. 6.13) are

$$\begin{aligned}
\tfrac{\partial \bar{a}_{j,j+1}}{\partial a_j} &= \tfrac{1}{2}\Delta R_{i,j} \\
\tfrac{\partial \bar{a}_{j,j+1}}{\partial a_{j+1}} &= \tfrac{1}{2}\Delta R_{i,j+1} \\
\tfrac{\partial \bar{a}_{j,j+1}}{\partial \Delta R_{i,j}} &= -\left[\bar{a}_{j,j+1}\right]_\times \\
\tfrac{\partial \bar{a}_{j,j+1}}{\partial \bar{\omega}_{j,j+1}} &= -\tfrac{1}{2}\Delta R_{i,j+1}\left[a_{j+1}\right]_\times J_r\left(\bar{\omega}_{j,j+1}\Delta t\right)\Delta t \\
\tfrac{\partial \bar{a}_{j,j+1}}{\partial \omega_j} &= \tfrac{\partial \bar{a}_{j,j+1}}{\partial \bar{\omega}_{j,j+1}} \tfrac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_j} \\
\tfrac{\partial \bar{a}_{j,j+1}}{\partial \omega_{j+1}} &= \tfrac{\partial \bar{a}_{j,j+1}}{\partial \bar{\omega}_{j,j+1}} \tfrac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_{j+1}}
\end{aligned} \tag{6.37}$$

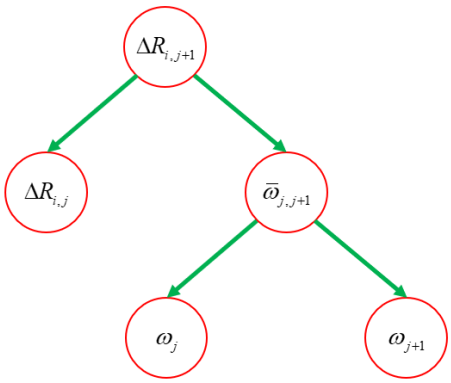


Figure 6.11: The relationship between $\Delta R_{i,j+1}$ and $\{\Delta_{i,j}, \omega_j, \omega_{j+1}, a_j, a_{j+1}\}$.

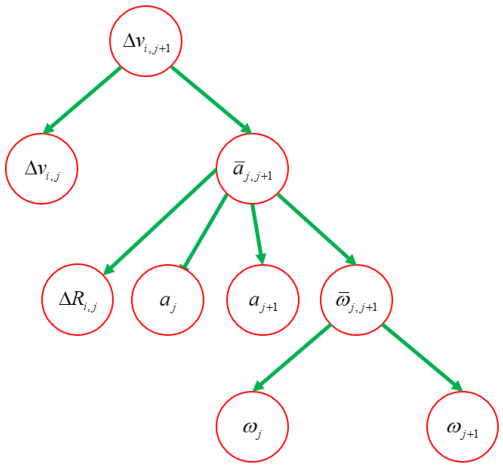


Figure 6.12: The relationship between $\Delta v_{i,j+1}$ and $\{\Delta_{i,j}, \omega_j, \omega_{j+1}, a_j, a_{j+1}\}$.

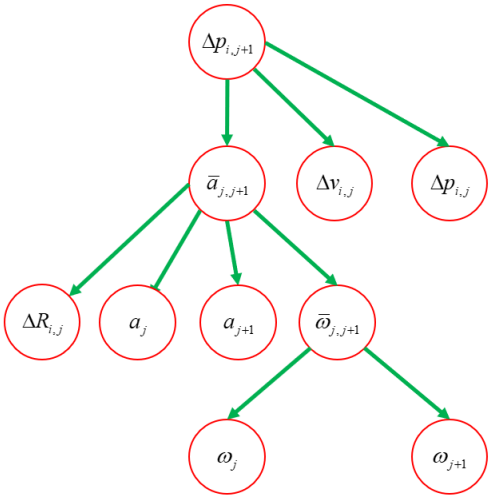


Figure 6.13: The relationship between $\Delta p_{i,j+1}$ and $\{\Delta_{i,j}, \omega_j, \omega_{j+1}, a_j, a_{j+1}\}$.

Where $J_r(\bullet)$ is the right Jacobian of SO(3) [73].

Recall Eq. 6.11,

$$
\begin{aligned}
&\frac{\partial \Delta R_{i,j+1}}{\partial \Delta R_{i,j}} = I_3 \\
&\frac{\partial \Delta R_{i,j+1}}{\partial \bar{\omega}_{j,j+1}} = \Delta R_{i,j+1} J_r\left(\bar{\omega}_{j,j+1}\Delta t\right)\Delta t \\
&\frac{\partial \Delta v_{i,j+1}}{\partial \Delta v_{i,j}} = I_3 \\
&\frac{\partial \Delta v_{i,j+1}}{\partial \Delta R_{i,j}} = \frac{\partial \Delta v_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \Delta R_{i,j}} = (\Delta t) \frac{\partial \bar{a}_{j,j+1}}{\partial \Delta R_{i,j}} \\
&\frac{\partial \Delta p_{i,j+1}}{\partial \Delta p_{i,j}} = I_3 \\
&\frac{\partial \Delta p_{i,j+1}}{\partial \Delta v_{i,j}} = I_3 \Delta t \\
&\frac{\partial \Delta p_{i,j+1}}{\partial \Delta R_{i,j}} = \frac{\partial \Delta p_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \Delta R_{i,j}} = \left(\tfrac{1}{2}\Delta t^2\right) \frac{\partial \bar{a}_{j,j+1}}{\partial \Delta R_{i,j}}
\end{aligned}
\tag{6.38}
$$

The Jacobians of f with respect to $\Delta_{i,j}$ can be obtained by the above equation. The remaining Jacobians are related to $\{a_j, a_{j+1}, \omega_j, \omega_{j+1}\}$.

The jacobians of f with respect to a_j are

$$
\begin{aligned}
&\frac{\partial \Delta v_{i,j+1}}{\partial a_j} = \frac{\partial \Delta v_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial a_j} = (\Delta t) \frac{\partial \bar{a}_{j,j+1}}{\partial a_j} \\
&\frac{\partial \Delta p_{i,j+1}}{\partial a_j} = \frac{\partial \Delta p_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial a_j} = \left(\tfrac{1}{2}\Delta t^2\right) \frac{\partial \bar{a}_{j,j+1}}{\partial a_j}
\end{aligned}
\tag{6.39}
$$

The Jacobians of f with respect to a_{j+1} are

$$\begin{aligned}\frac{\partial \Delta v_{i,j+1}}{\partial a_{j+1}} &= \frac{\partial \Delta v_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial a_{j+1}} = (\Delta t)\, \frac{\partial \bar{a}_{j,j+1}}{\partial a_{j+1}} \\ \frac{\partial \Delta p_{i,j+1}}{\partial a_{j+1}} &= \frac{\partial \Delta p_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial a_{j+1}} = \left(\tfrac{1}{2}\Delta t^2\right) \frac{\partial \bar{a}_{j,j+1}}{\partial a_{j+1}}\end{aligned} \tag{6.40}$$

The Jacobians of f with respect to ω_j are

$$\begin{aligned}\frac{\partial \Delta R_{i,j+1}}{\partial \omega_j} &= \frac{\partial \Delta R_{i,j+1}}{\partial \bar{\omega}_{j,j+1}} \frac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_j} \\ \frac{\partial \Delta v_{i,j+1}}{\partial \omega_j} &= \frac{\partial \Delta v_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_j} = (\Delta t)\, \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_j} \\ \frac{\partial \Delta p_{i,j+1}}{\partial \omega_j} &= \frac{\partial \Delta p_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_j} = \left(\tfrac{1}{2}\Delta t^2\right) \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_j}\end{aligned} \tag{6.41}$$

Where $\frac{\partial \Delta R_{i,j+1}}{\partial \bar{\omega}_{j,j+1}}$ is available in Eq. 6.38.

The Jacobians of f with respect to ω_{j+1} are

$$\begin{aligned}\frac{\partial \Delta R_{i,j+1}}{\partial \omega_{j+1}} &= \frac{\partial \Delta R_{i,j+1}}{\partial \bar{\omega}_{j,j+1}} \frac{\partial \bar{\omega}_{j,j+1}}{\partial \omega_{j+1}} \\ \frac{\partial \Delta v_{i,j+1}}{\partial \omega_{j+1}} &= \frac{\partial \Delta v_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_{j+1}} = (\Delta t)\, \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_{j+1}} \\ \frac{\partial \Delta p_{i,j+1}}{\partial \omega_{j+1}} &= \frac{\partial \Delta p_{i,j+1}}{\partial \bar{a}_{j,j+1}} \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_{j+1}} = \left(\tfrac{1}{2}\Delta t^2\right) \frac{\partial \bar{a}_{j,j+1}}{\partial \omega_{j+1}}\end{aligned} \tag{6.42}$$

According to Eq. 6.8

$$\begin{aligned}\frac{\partial f}{\partial \tilde{\omega}_j} &= \frac{\partial f}{\partial \omega_j} \\ \frac{\partial f}{\partial \tilde{\omega}_{j+1}} &= \frac{\partial f}{\partial \omega_{j+1}} \\ \frac{\partial f}{\partial \tilde{a}_j} &= \frac{\partial f}{\partial a_j} \\ \frac{\partial f}{\partial \tilde{a}_{j+1}} &= \frac{\partial f}{\partial a_{j+1}}\end{aligned} \tag{6.43}$$

Therefore, we have completed the Jacobian calculation of $\left\{\frac{\partial f}{\partial \Delta_{i,j}}, \frac{\partial f}{\partial \tilde{\omega}_j}, \frac{\partial f}{\partial \tilde{\omega}_{j+1}}, \frac{\partial f}{\partial \tilde{a}_j}, \frac{\partial f}{\partial \tilde{a}_{j+1}}\right\}$.

The Jacobians of f with respect to $\{b_\omega, b_a\}$ are

$$\begin{aligned}\frac{\partial f}{\partial b_\omega} &= \frac{\partial f}{\partial \omega_j} \frac{\partial \omega_j}{\partial b_\omega} + \frac{\partial f}{\partial \omega_{j+1}} \frac{\partial \omega_{j+1}}{\partial b_\omega} = -\left(\frac{\partial f}{\partial \omega_j} + \frac{\partial f}{\partial \omega_{j+1}}\right) \\ \frac{\partial f}{\partial b_a} &= \frac{\partial f}{\partial a_j} \frac{\partial a_j}{\partial b_a} + \frac{\partial f}{\partial a_{j+1}} \frac{\partial a_{j+1}}{\partial b_a} = -\left(\frac{\partial f}{\partial a_j} + \frac{\partial f}{\partial a_{j+1}}\right)\end{aligned} \tag{6.44}$$

Finally, the Jacobians of f (Eq. 6.13) with respect to all involved variables, $\{\Delta_{i,j}, \tilde{\omega}_j, \tilde{\omega}_{j+1}, \tilde{a}_j, \tilde{a}_{j+1}, b_\omega, b_a\}$, are analytical computed.

6.11 Jacobians of pixel measurement residual

For the ease of Jacobian derivation, Eq. 6.20 is rewritten as

$$
\begin{aligned}
{}^{n}r_{il} &= \pi\left({}^{C}p_f\right) - {}^{n}u_{il} \\
{}^{C}p_f &\triangleq {}^{C}_{W}T\,{}^{W}p_{f_l} \\
{}^{C}_{W}T &\triangleq {}^{I}_{C_n}T^{-1}\,{}^{W}_{I_i}T(t_i + t_d)^{-1}
\end{aligned}
\tag{6.45}
$$

The Jacobian of the pixel residual ${}^{n}r_{il}$ with respect to the 3D point in camera frame ${}^{C}p_f$ is $\frac{\partial {}^{n}r_{il}}{\partial {}^{C}p_f}$. This is determined by the camera projection model [66, 67]. The Jacobian of the pixel residual ${}^{n}r_{il}$ with respect to the camera pose ${}^{C}_{W}T$ is

$$
\frac{\partial {}^{n}r_{il}}{\partial {}^{C}_{W}T} = \frac{\partial {}^{n}r_{il}}{\partial {}^{C}p_f}\frac{\partial {}^{C}p_f}{\partial {}^{C}_{W}T}
\tag{6.46}
$$

The Jacobian of ${}^{C}p_f$ with respect to ${}^{C}_{W}T$ is

$$
\frac{\partial {}^{C}p_f}{\partial {}^{C}_{W}T} = \left({}^{C}_{W}T\,{}^{W}p_{f_l}\right)^{\odot}
\tag{6.47}
$$

Where $\odot$ is an operator for the homogeneous coordinate [73, Sec. 7.1.8].

Using the last equation from Eq. 6.45, and Eq. 6.22, the Jacobians of ${}^{C}_{W}T$ with respect to $\left\{{}^{I}_{C_n}T, {}^{W}p_{I_i}, {}^{W}_{I_i}R, t_d\right\}$ are

$$
\begin{aligned}
\frac{\partial {}^{C}_{W}T}{\partial {}^{I}_{C_n}T} &= -I_6 \\
\frac{\partial {}^{C}_{W}T}{\partial {}^{W}_{I_i}T} &= -Ad\left({}^{I}_{C_n}T^{-1}\right) \\
\frac{\partial {}^{W}_{I_i}T}{\partial {}^{W}p_{I_i}} &= \begin{bmatrix} {}^{W}_{I_i}R^T \\ 0_3 \end{bmatrix} \\
\frac{\partial {}^{W}_{I_i}T}{\partial {}^{W}_{I_i}R} &= \begin{bmatrix} 0_3 \\ {}^{W}_{I_i}R^T \end{bmatrix} \\
\frac{\partial {}^{W}_{I_i}T}{\partial t_d} &= \begin{bmatrix} {}^{W}_{I_i}R^T\,{}^{W}v_{I_i} \\ \omega_i \end{bmatrix}
\end{aligned}
\tag{6.48}
$$

Where $Ad\left(\bullet\right)$ is the adjoint of SE(3) [73]. Finally, the analytical on-manifold Jacobian of

${}^{n}r_{il}$ with respect to all involved variables in x_s (Eq. 6.24) can be calculated via the chain rule

$$\begin{aligned}
\frac{\partial {}^{n}r_{il}}{\partial {}^{I}_{C_n}T} &= \frac{\partial {}^{n}r_{il}}{\partial {}^{C}_{W}T}\frac{\partial {}^{C}_{W}T}{\partial {}^{I}_{C_n}T} \\
\frac{\partial {}^{n}r_{il}}{\partial {}^{W}p_{I_i}} &= \frac{\partial {}^{n}r_{il}}{\partial {}^{C}_{W}T}\frac{\partial {}^{C}_{W}T}{\partial {}^{W}_{I_i}T}\frac{\partial {}^{W}_{I_i}T}{\partial {}^{W}p_{I_i}} \\
\frac{\partial {}^{n}r_{il}}{\partial {}^{W}_{I_i}R} &= \frac{\partial {}^{n}r_{il}}{\partial {}^{C}_{W}T}\frac{\partial {}^{C}_{W}T}{\partial {}^{W}_{I_i}T}\frac{\partial {}^{W}_{I_i}T}{\partial {}^{W}_{I_i}R} \\
\frac{\partial {}^{n}r_{il}}{\partial t_d} &= \frac{\partial {}^{n}r_{il}}{\partial {}^{C}_{W}T}\frac{\partial {}^{C}_{W}T}{\partial {}^{W}_{I_i}T}\frac{\partial {}^{W}_{I_i}T}{\partial t_d}
\end{aligned} \tag{6.49}$$

Chapter 7

Conclusion and perspectives

This thesis focuses on the calibration problems involved in three types of visual localization applications. The first type is vision-based relative localization. The second type is multi-sensor localization with the participation of GPS. The third type is multi-sensor localization without the participation of GPS. These three types of applications almost cover all aspects of visual localization in the field of robotics.

For vision-based relative localization, a high-precision global pose sensor is desired for the evaluation and its spatial-temporal relationship with the camera should be calibrated. Two novel calibration algorithms, including target-based and target-less methods, are proposed to estimate the spatial-temporal parameters between the camera and the global pose sensor. The mathematical model can be easily extended to other global sensors, such as RTK-GPS [109]. In addition, since the measurement model considers the intrinsic parameters of the camera, the calibration algorithm can support joint optimization with intrinsic parameters for different camera models, improving ease of use.

For multi-sensor localization with the participation of GPS, specifically the GPS-VIO system, the observability of extrinsic parameters between GPS and VIO is studied. A misleading conclusion in the existing research is identified, namely that the rotational extrinsic parameter is unobservable. Experiments demonstrate that the rotational extrinsic parameter is observable in fact, therefore its online calibration can improve localization accuracy overall. Novel theoretical support for this conclusion is provided by nonlinear observability analysis.

This study indicates that, for the observability analysis of calibration parameters, the nonlinear observability analysis is more comprehensive and profound than the linear observability analysis. For similar future work, it is suggested conducting sufficient experiments to verify the conclusions obtained from the linear observability analysis to avoid potential mistakes.

For the VIO system, the observability from [19, 20] is revisited for online extrinsic calibration. A novel proof shows that the common-seen pure translational straight line motion can lead to the unobservability of the rotational extrinsic parameter between IMU and camera (at least one degree of freedom). The unobservability still holds for global-pose aided VIO. This finding makes up for the shortcomings of existing research conclusions.

Lastly, to empower productivity, this thesis pushes the boundaries of research from the perspective of calibration efficiency. Traditional offline calibration methods typically use continuous-time state representation, which is accurate but inefficient. To address this limitation, a novel calibration algorithm is proposed based on discrete-time state representation, which significantly reduces optimization time without sacrificing accuracy. This study can greatly improve the efficiency of factory calibration for industry products and is particularly suitable for resource-constrained platforms. Moreover, this research paves the way for the transition from continuous-time state representation to discrete-time state representation. In the future, this research idea could be expanded to accelerate the calibration of multi-visual-inertial systems [29], and generalized to other multi-sensor systems [30, 110], for example, event-related systems, which are qualified to deal with challenging scenarios that are inaccessible to traditional cameras, such as high-speed motion and/or high dynamic range (HDR) illumination[111, 112, 113]. It is also interesting to reevaluate and rethink whether continuous-time state representation has advantages over discrete-time state representation for various state estimation applications[114, 115, 57].

References

[1] Google. *ARCore*. https://developers.google.com/ar.

[2] Apple. *ARKit*. https://developer.apple.com/augmented-reality.

[3] Meta. *Oculus*. https://store.facebook.com/quest.

[4] Yunfei Fan, Tianyu Zhao, and Guidong Wang. "SchurVINS: Schur Complement-Based Lightweight Visual Inertial Navigation System". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17964–17973.

[5] Kejian J Wu, Chao X Guo, Georgios Georgiou, and Stergios I Roumeliotis. "Vins on wheels". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 5155–5162.

[6] Jeffrey Delmerico and Davide Scaramuzza. "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots". In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 2502–2509.

[7] Jeonguk Kang, Hyunbin Kim, and Kyung-Soo Kim. "VIEW: Visual-Inertial External Wrench Estimator for Legged Robot". In: *IEEE Robotics and Automation Letters* (2023).

[8] Junlin Song, Pedro J Sanchez-Cuevas, Antoine Richard, Raj Thilak Rajan, and Miguel Olivares-Mendez. "GPS-VIO Fusion with Online Rotational Calibration". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 11906–11912.

[9] Anastasios I Mourikis, Nikolas Trawny, Stergios I Roumeliotis, Andrew E Johnson, Adnan Ansar, and Larry Matthies. “Vision-aided inertial navigation for spacecraft entry, descent, and landing”. In: *IEEE Transactions on Robotics* 25.2 (2009), pp. 264–280.

[10] David S Bayard, Dylan T Conway, Roland Brockers, Jeff H Delaune, Larry H Matthies, Håvard F Grip, Gene B Merewether, Travis L Brown, and Alejandro M San Martin. “Vision-based navigation for the NASA mars helicopter”. In: *AIAA Scitech 2019 Forum*. 2019, p. 1411.

[11] Jeff Delaune, David S Bayard, and Roland Brockers. “Range-visual-inertial odometry: Scale observability without excitation”. In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2421–2428.

[12] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. “Project aria: A new tool for egocentric multi-modal ai research”. In: *arXiv preprint arXiv:2308.13561* (2023).

[13] Suyoung Kang, Ryan Soussan, Daekyeong Lee, Brian Coltin, Andres Mora Vargas, Marina Moreira, Katie Browne, Ruben Garcia, Maria Bualat, Trey Smith, et al. “Astrobee iss free-flyer datasets for space intra-vehicular robot navigation research”. In: *IEEE Robotics and Automation Letters* 9.4 (2024), pp. 3307–3314.

[14] Jinyong Jeong, Younggun Cho, Young-Sik Shin, Hyunchul Roh, and Ayoung Kim. “Complex urban dataset with multi-level sensors from highly diverse urban environments”. In: *The International Journal of Robotics Research* 38.6 (2019), pp. 642–657.

[15] Haotian Li, Yuying Zou, Nan Chen, Jiarong Lin, Xiyuan Liu, Wei Xu, Chunran Zheng, Rundong Li, Dongjiao He, Fanze Kong, et al. “MARS-LVIG dataset: A multi-sensor aerial robots SLAM dataset for LiDAR-visual-inertial-GNSS fusion”. In: *The International Journal of Robotics Research* 43.8 (2024), pp. 1114–1127.

[16] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. “Openvins: A research platform for visual-inertial estimation”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 4666–4672.

[17] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. “Fast-lio2: Fast direct lidar-inertial odometry”. In: *IEEE Transactions on Robotics* 38.4 (2022), pp. 2053–2073.

[18] Chunran Zheng, Wei Xu, Zuhao Zou, Tong Hua, Chongjian Yuan, Dongjiao He, Bingyang Zhou, Zheng Liu, Jiarong Lin, Fangcheng Zhu, et al. “Fast-livo2: Fast, direct lidar-inertial-visual odometry”. In: *IEEE Transactions on Robotics* (2024).

[19] Yulin Yang, Patrick Geneva, Kevin Eckenhoff, and Guoquan Huang. “Degenerate motion analysis for aided ins with online spatial and temporal sensor calibration”. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 2070–2077.

[20] Yulin Yang, Patrick Geneva, Xingxing Zuo, and Guoquan Huang. “Online self-calibration for visual-inertial navigation: Models, analysis, and degeneracy”. In: *IEEE Transactions on Robotics* 39.5 (2023), pp. 3479–3498.

[21] Junlin Song, Antoine Richard, and Miguel Olivares-Mendez. “Observability Investigation for Rotational Calibration of (Global-pose aided) VIO under Straight Line Motion”. In: *arXiv preprint arXiv:2503.00027* (2025).

[22] Faraz M Mirzaei and Stergios I Roumeliotis. “A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation”. In: *IEEE transactions on robotics* 24.5 (2008), pp. 1143–1156.

[23] Jonathan Kelly and Gaurav S Sukhatme. “Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration”. In: *The International Journal of Robotics Research* 30.1 (2011), pp. 56–79.

[24] Mingyang Li and Anastasios I Mourikis. “Online temporal calibration for camera–IMU systems: Theory and algorithms”. In: *The International Journal of Robotics Research* 33.7 (2014), pp. 947–964.

[25] Paul Furgale, Joern Rehder, and Roland Siegwart. “Unified temporal and spatial calibration for multi-sensor systems”. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 1280–1286.

[26] Joern Rehder, Janosch Nikolic, Thomas Schneider, Timo Hinzmann, and Roland Siegwart. “Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes”. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 4304–4311.

[27] Christiane Sommer, Vladyslav Usenko, David Schubert, Nikolaus Demmel, and Daniel Cremers. “Efficient derivative computation for cumulative b-splines on lie groups”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 11148–11156.

[28] Xiangyang Zhi, Jiawei Hou, Yiren Lu, Laurent Kneip, and Sören Schwertfeger. “Multical: Spatiotemporal calibration for multiple IMUs, cameras and LiDARs”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 2446–2453.

[29] Yulin Yang, Patrick Geneva, and Guoquan Huang. “Multi-visual-inertial system: Analysis, calibration, and estimation”. In: *The International Journal of Robotics Research* 43.13 (2024), pp. 1995–2026.

[30] Shuolong Chen, Xingxing Li, Shengyu Li, Yuxuan Zhou, and Xiaoteng Yang. “ikalibr: Unified targetless spatiotemporal calibration for resilient integrated inertial systems”. In: *IEEE Transactions on Robotics* (2025).

[31] Junlin Song, Antoine Richard, and Miguel Olivares-Mendez. “Unleashing the Power of Discrete-Time State Representation: Ultrafast Target-based IMU-Camera Spatial-Temporal Calibration”. In: *arXiv preprint arXiv:2509.12846* (2025).

[32] Joern Rehder, Roland Siegwart, and Paul Furgale. “A general approach to spatiotemporal calibration in multisensor systems”. In: *IEEE Transactions on Robotics* 32.2 (2016), pp. 383–398.

[33] Fadri Furrer, Marius Fehr, Tonci Novkovic, Hannes Sommer, Igor Gilitschenski, and Roland Siegwart. “Evaluation of combined time-offset estimation and hand-eye calibration on robotic datasets”. In: *Field and Service Robotics: Results of the 11th International Conference*. Springer. 2018, pp. 145–159.

[34] Kejie Qiu, Tong Qin, Jie Pan, Siqi Liu, and Shaojie Shen. “Real-time temporal and rotational calibration of heterogeneous sensors using motion correlation analysis”. In: *IEEE Transactions on Robotics* 37.2 (2020), pp. 587–602.

[35] Tong Qin and Shaojie Shen. “Online temporal calibration for monocular visual-inertial systems”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 3662–3669.

[36] Yulin Yang, Patrick Geneva, Xingxing Zuo, and Guoquan Huang. “Online imu intrinsic calibration: Is it necessary?” In: *Robotics: Science and Systems*. 2020.

[37] Guoquan Huang. “Visual-inertial navigation: A concise review”. In: *2019 international conference on robotics and automation (ICRA)*. IEEE. 2019, pp. 9572–9582.

[38] Tong Qin, Peiliang Li, and Shaojie Shen. “Vins-mono: A robust and versatile monocular visual-inertial state estimator”. In: *IEEE Transactions on Robotics* 34.4 (2018), pp. 1004–1020.

[39] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jörg Stückler, and Daniel Cremers. “Visual-inertial mapping with non-linear factor recovery”. In: *IEEE Robotics and Automation Letters* 5.2 (2019), pp. 422–429.

[40] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam”. In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 1874–1890.

[41] Anastasios I Mourikis and Stergios I Roumeliotis. “A multi-state constraint Kalman filter for vision-aided inertial navigation”. In: *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE. 2007, pp. 3565–3572.

[42] Ke Sun, Kartik Mohta, Bernd Pfrommer, Michael Watterson, Sikang Liu, Yash Mulgaonkar, Camillo J Taylor, and Vijay Kumar. “Robust stereo visual inertial odometry for fast autonomous flight”. In: *IEEE Robotics and Automation Letters* 3.2 (2018), pp. 965–972.

[43] Guoquan Huang, Michael Kaess, and John J Leonard. “Towards consistent visual-inertial navigation”. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 4926–4933.

[44] Tong Qin, Shaozu Cao, Jie Pan, and Shaojie Shen. “A general optimization-based framework for global pose estimation with multiple sensors”. In: *arXiv preprint arXiv:1901.03642* (2019).

[45] Ruben Mascaro, Lucas Teixeira, Timo Hinzmann, Roland Siegwart, and Margarita Chli. “Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1421–1428.

[46] Yang Yu, Wenliang Gao, Chengju Liu, Shaojie Shen, and Ming Liu. “A gps-aided omnidirectional visual-inertial state estimator in ubiquitous environments”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 7750–7755.

[47] Giovanni Cioffi and Davide Scaramuzza. “Tightly-coupled fusion of global positional measurements in optimization-based visual-inertial odometry”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 5089–5095.

[48] Woosik Lee, Kevin Eckenhoff, Patrick Geneva, and Guoquan Huang. “Intermittent gps-aided vio: Online initialization and calibration”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 5724–5731.

[49] Yonggang Tang, Yuanxin Wu, Meiping Wu, Wenqi Wu, Xiaoping Hu, and Lincheng Shen. “INS/GPS integration: Global observability analysis”. In: *IEEE Transactions on Vehicular Technology* 58.3 (2008), pp. 1129–1142.

[50] Yuanxin Wu, Hongliang Zhang, Meiping Wu, Xiaoping Hu, and Dewen Hu. “Observability of strapdown INS alignment: A global perspective”. In: *IEEE Transactions on Aerospace and Electronic Systems* 48.1 (2012), pp. 78–102.

[51] Robert Hermann and Arthur Krener. “Nonlinear controllability and observability”. In: *IEEE Transactions on automatic control* 22.5 (1977), pp. 728–740.

[52] Junlin Song, Pedro J Sanchez-Cuevas, Antoine Richard, and Miguel Olivares-Mendez. “GPS-aided Visual Wheel Odometry”. In: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2023, pp. 375–382.

[53] Shaozu Cao, Xiuyuan Lu, and Shaojie Shen. “Gvins: Tightly coupled gnss–visual–inertial fusion for smooth and consistent state estimation”. In: *IEEE Transactions on Robotics* (2022).

[54] Simon Boche, Xingxing Zuo, Simon Schaefer, and Stefan Leutenegger. “Visual-inertial slam with tightly-coupled dropout-tolerant gps fusion”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 7020–7027.

[55] Joan Sola, Jeremie Deray, and Dinesh Atchuthan. “A micro Lie theory for state estimation in robotics”. In: *arXiv preprint arXiv:1812.01537* (2018).

[56] Giovanni Cioffi, Titus Cieslewski, and Davide Scaramuzza. “Continuous-time vs. discrete-time vision-based SLAM: A comparative study”. In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 2399–2406.

[57] William Talbot, Julian Nubert, Turcan Tuna, Cesar Cadena, Frederike Dümbgen, Jesus Tordesillas, Timothy D Barfoot, and Marco Hutter. “Continuous-time state estimation methods in robotics: A survey”. In: *IEEE Transactions on Robotics* (2025).

[58] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. “IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation”. In: *Robotics: Science and Systems XI*. 2015.

[59] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. “On-manifold preintegration for real-time visual–inertial odometry”. In: *IEEE Transactions on Robotics* 33.1 (2016), pp. 1–21.

[60] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. “The EuRoC micro aerial vehicle datasets”. In: *The International Journal of Robotics Research* 35.10 (2016), pp. 1157–1163.

[61] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. “The TUM VI benchmark for evaluating visual-inertial odometry”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1680–1687.

[62] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. “Are we ready for autonomous drone racing? the UZH-FPV drone racing dataset”. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 6713–6719.

[63] Nikolas Trawny and Stergios I Roumeliotis. “Indirect Kalman filter for 3D attitude estimation”. In: *University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep* 2 (2005), p. 2005.

[64] Edwin Olson. “AprilTag: A robust and flexible visual fiducial system”. In: *2011 IEEE international conference on robotics and automation*. IEEE. 2011, pp. 3400–3407.

[65] Nived Chebrolu, Thomas Läbe, Olga Vysotska, Jens Behley, and Cyrill Stachniss. “Adaptive robust kernels for non-linear least squares problems”. In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2240–2247.

[66] Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. “The double sphere camera model”. In: *2018 International Conference on 3D Vision (3DV)*. IEEE. 2018, pp. 552–560.

[67] Lionel Heng, Bo Li, and Marc Pollefeys. “Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry”. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 1793–1800.

[68] Mingyang Li and Anastasios I Mourikis. “Optimization-based estimator design for vision-aided inertial navigation”. In: *Robotics: Science and Systems*. Berlin Germany. 2013, pp. 241–248.

[69] Mingyang Li. *Visual-inertial odometry on resource-constrained systems*. University of California, Riverside, 2014.

[70] Jing Dong, Mustafa Mukadam, Byron Boots, and Frank Dellaert. “Sparse Gaussian processes on matrix lie groups: A unified framework for optimizing continuous-time trajectories”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 6497–6504.

[71] David Schubert, Nikolaus Demmel, Vladyslav Usenko, Jorg Stuckler, and Daniel Cremers. “Direct sparse odometry with rolling shutter”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 682–697.

[72] Tim D Barfoot, Chi Hay Tong, and Simo Särkkä. “Batch Continuous-Time Trajectory Estimation as Exactly Sparse Gaussian Process Regression.” In: *Robotics: Science and Systems*. Vol. 10. Citeseer. 2014, pp. 1–10.

[73] Timothy D Barfoot. *State estimation for robotics*. Cambridge University Press, 2024.

[74] Zhe Chen, Ke Jiang, and James C Hung. “Local observability matrix and its application to observability analyses”. In: *[Proceedings] IECON’90: 16th Annual Conference of IEEE Industrial Electronics Society*. IEEE. 1990, pp. 100–103.

[75] Mingyang Li and Anastasios I Mourikis. “High-precision, consistent EKF-based visual-inertial odometry”. In: *The International Journal of Robotics Research* 32.6 (2013), pp. 690–711.

[76] Joel A Hesch, Dimitrios G Kottas, Sean L Bowman, and Stergios I Roumeliotis. “Consistency analysis and improvement of vision-aided inertial navigation”. In: *IEEE Transactions on Robotics* 30.1 (2013), pp. 158–176.

[77] Jiajun Lv, Xingxing Zuo, Kewei Hu, Jinhong Xu, Guoquan Huang, and Yong Liu. “Observability-Aware Intrinsic and Extrinsic Calibration of LiDAR-IMU Systems”. In: *IEEE Transactions on Robotics* (2022).

[78] Davide Falanga, Kevin Kleber, Stefano Mintchev, Dario Floreano, and Davide Scaramuzza. “The foldable drone: A morphing quadrotor that can squeeze and fly”. In: *IEEE Robotics and Automation Letters* 4.2 (2018), pp. 209–216.

[79] Woosik Lee, Kevin Eckenhoff, Patrick Geneva, and Guoquan Huang. “Gps-aided visualinertial navigation in large-scale environments”. In: *Robot Perception and Navigation Group (RPNG), University of Delaware, Tech. Rep* (2019).

[80] William J Terrell. “Local observability of nonlinear differential-algebraic equations (DAEs) from the linearization along a trajectory”. In: *IEEE Transactions on Automatic Control* 46.12 (2001), pp. 1947–1950.

[81] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. “SVO: Semidirect visual odometry for monocular and multicamera systems”. In: *IEEE Transactions on Robotics* 33.2 (2016), pp. 249–265.

[82] RPG_SVO_PRO_OPEN. *Known issues and possible improvements*. Available at https://github.com/uzh-rpg/rpg_svo_pro_open/blob/master/doc/known_issues_and_improvements.md.

[83] Agostino Martinelli. “Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination”. In: *IEEE Transactions on Robotics* 28.1 (2011), pp. 44–60.

[84] Woosik Lee, Patrick Geneva, Yulin Yang, and Guoquan Huang. “Tightly-coupled GNSS-aided visual-inertial localization”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 9484–9491.

[85] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. “Keyframe-based visual–inertial odometry using nonlinear optimization”. In: *The International Journal of Robotics Research* 34.3 (2015), pp. 314–334.

[86] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. “Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback”. In: *The International Journal of Robotics Research* 36.10 (2017), pp. 1053–1072.

[87] Rik Girod. “State Estimation and Mission Planning for Precision-critical Aerial Field Robotics”. PhD thesis. ETH Zurich, 2022.

[88] Francesco Crocetti, Enrico Bellocchio, Alberto Dionigi, Simone Felicioni, Gabriele Costante, Mario L Fravolini, and Paolo Valigi. “ARD-VO: Agricultural robot data set of vineyards and olive groves”. In: *Journal of Field Robotics* 40.6 (2023), pp. 1678–1696.

[89] Pau Vial, Joan Solà, Narcís Palomeras, and Marc Carreras. “On Lie group IMU and linear velocity preintegration for autonomous navigation considering the Earth rotation compensation”. In: *IEEE Transactions on Robotics* (2024).

[90] Xingxing Zuo, Nathaniel Merrill, Wei Li, Yong Liu, Marc Pollefeys, and Guoquan Huang. “CodeVIO: Visual-inertial odometry with learned optimizable dense depth”. In: *2021 ieee international conference on robotics and automation (icra)*. IEEE. 2021, pp. 14382–14388.

[91] Kejie Qiu, Tong Qin, Wenliang Gao, and Shaojie Shen. “Tracking 3-D motion of dynamic objects using monocular visual-inertial sensing”. In: *IEEE Transactions on Robotics* 35.4 (2019), pp. 799–816.

[92] Kevin Eckenhoff, Patrick Geneva, Nathaniel Merrill, and Guoquan Huang. “Schmidt-EKF-based visual-inertial moving object tracking”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 651–657.

[93] Drew Hanover, Antonio Loquercio, Leonard Bauersfeld, Angel Romero, Robert Penicka, Yunlong Song, Giovanni Cioffi, Elia Kaufmann, and Davide Scaramuzza. “Autonomous drone racing: A survey”. In: *IEEE Transactions on Robotics* (2024).

[94] Ivan Alberico, Jeff Delaune, Giovanni Cioffi, and Davide Scaramuzza. “Structure-Invariant Range-Visual-Inertial Odometry”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 10613–10620.

[95] Johannes L Schonberger and Jan-Michael Frahm. “Structure-from-motion revisited”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4104–4113.

[96] Hannes Ovrén and Per-Erik Forssén. “Spline error weighting for robust visual-inertial fusion”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 321–329.

[97] Hannes Ovrén and Per-Erik Forssén. “Trajectory representation and landmark projection for continuous-time structure from motion”. In: *The International Journal of Robotics Research* 38.6 (2019), pp. 686–701.

[98] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361.

[99] Ashish Devadas Nair, Julien Kindle, Plamen Levchev, and Davide Scaramuzza. “Hilti SLAM Challenge 2023: Benchmarking Single + Multi-Session SLAM Across Sensor Constellations in Construction”. In: *IEEE Robotics and Automation Letters* 9.8 (2024), pp. 7286–7293.

[100] TaeYoung Kim, Gyuhyeon Pak, and Euntai Kim. “GRIL-Calib: Targetless Ground Robot IMU-LiDAR Extrinsic Calibration Method using Ground Plane Motion Constraints”. In: *IEEE Robotics and Automation Letters* (2024).

[101] Todd Lupton and Salah Sukkarieh. “Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions”. In: *IEEE Transactions on Robotics* 28.1 (2011), pp. 61–76.

[102] Yifu Wang, Yonhon Ng, Inkyu Sa, Alvaro Parra, Cristian Rodriguez-Opazo, Taojun Lin, and Hongdong Li. “MAVIS: Multi-Camera Augmented Visual-Inertial SLAM using SE 2 (3) Based Exact IMU Pre-integration”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 1694–1700.

[103] Giulio Delama, Alessandro Fornasier, Robert Mahony, and Stephan Weiss. “Equivariant IMU Preintegration with Biases: a Galilean Group Approach”. In: *IEEE Robotics and Automation Letters* (2024).

[104] Yonggen Ling and Shaojie Shen. “High-precision online markerless stereo extrinsic calibration”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 1771–1778.

[105] Ryan Nemiroff, Kenny Chen, and Brett T Lopez. “Joint On-Manifold Gravity and Accelerometer Intrinsics Estimation for Inertially Aligned Mapping”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 1388–1394.

[106] Elmar Mair, Michael Fleps, Michael Suppa, and Darius Burschka. “Spatio-temporal initialization for IMU to camera registration”. In: *2011 IEEE International Conference on Robotics and Biomimetics*. IEEE. 2011, pp. 557–564.

[107] Zichao Zhang and Davide Scaramuzza. “A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 7244–7251.

[108] Fangcheng Zhu, Yunfan Ren, and Fu Zhang. “Robust real-time lidar-inertial initialization”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 3948–3955.

[109] Alex D Jordan, Jacob C Johnson, Timothy W McLain, and Randal W Beard. “Offline GNSS/camera extrinsic calibration using RTK and fiducials”. In: *IEEE Robotics and Automation Letters* 8.6 (2023), pp. 3454–3461.

[110] Shuolong Chen, Xingxing Li, and Liu Yuan. “eKalibr-Inertial: Continuous-Time Spatiotemporal Calibration for Event-Based Visual-Inertial Systems”. In: *arXiv preprint arXiv:2509.05923* (2025).

[111] Florian Mahlknecht, Daniel Gehrig, Jeremy Nash, Friedrich M Rockenbauer, Benjamin Morrell, Jeff Delaune, and Davide Scaramuzza. “Exploring event camera-based odometry for planetary robots”. In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 8651–8658.

[112] Junkai Niu, Sheng Zhong, Xiuyuan Lu, Shaojie Shen, Guillermo Gallego, and Yi Zhou. “Esvo2: Direct visual-inertial odometry with stereo event cameras”. In: *IEEE Transactions on Robotics* (2025).

[113] ESA. *Event-based Lunar OPtical flow Egomotion estimation (ELOPE) challenge*. https://kelvins.esa.int/elope/home.

[114] Xiaolei Lang, Jiajun Lv, Jianxin Huang, Yukai Ma, Yong Liu, and Xingxing Zuo. “Ctrl-VIO: Continuous-time visual-inertial odometry for rolling shutter cameras”. In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 11537–11544.

[115] Xiaolei Lang, Chao Chen, Kai Tang, Yukai Ma, Jiajun Lv, Yong Liu, and Xingxing Zuo. “Coco-lic: continuous-time tightly-coupled lidar-inertial-camera odometry using non-uniform b-spline”. In: *IEEE Robotics and Automation Letters* 8.11 (2023), pp. 7074–7081.