Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor

Junlin Song Antoine Richard Miguel Olivares-Mendez

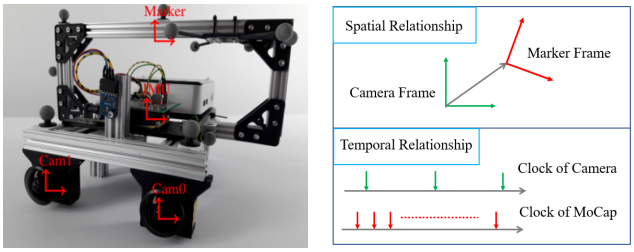
Abstract

*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.*

1. 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 [3, 6, 26]. The collection platform from [26] shown in Fig. 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. 1b).

This research was supported by the European Union's Horizon 2020 project SESAME (grant agreement No 101017258). The authors are with the Space Robotics (SpaceR) Research Group, Int. Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg.



(a) (b)

Figure 1. (a) Photo of the sensor setup, taken from [26]. (b) The spatial-temporal relationship between the camera measurements and the global pose measurements.

For instance, in Fig. 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 $\left\{ {}^{G}_{M}q \quad {}^{G}p_M \right\}$. Given the spatial-temporal 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 [9, 23, 24]. Target-based spatial-temporal hand-eye calibration was first presented in [10]. 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 [10] 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 cali-

bration 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 [22]. 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) [17, 21, 32]. Among them, the EKF-based methods are the most popular thanks to their computational efficiency. [17] 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 [31]. 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.

2. Notation

As shown in Fig. 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

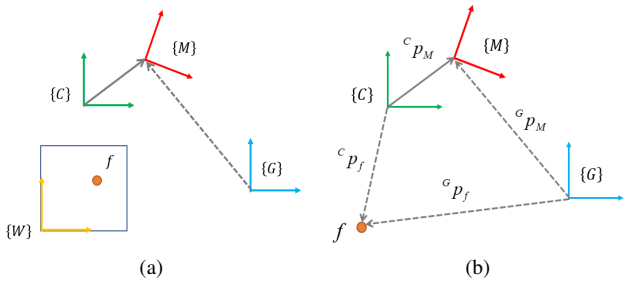


Figure 2. (a) Coordinate frames for the target-based method. (b) Coordinate frames for the target-less method.

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:

$$t_M = t_C + t_d \tag{1}$$

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 [29]. ${}^{M}_{G}q$ 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. 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 [20] as the calibration target, as shown in Fig. 3b. The coordinate frames involved in target-based method are depicted in Fig. 2a. Compared with Fig. 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:

$$\chi = \left\{ \begin{array}{cccccc} {}^{W}_{C_1}T & \cdots & {}^{W}_{C_N}T & {}^{G}_{W}T & {}^{C}_{M}T & t_d \end{array} \quad \varsigma \right\} \tag{2}$$

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:

$$\begin{aligned}\chi &= \arg\min\left\{\sum_{i=1}^{N}\sum_{j=1}^{K}\rho\left(r_{ij}\right)+\sum_{i=1}^{N}\rho\left(r_{gi}\right)\right\}\\ r_{ij} &= \pi\left({}^{C_i}_{W}T\,{}^{W}p_{f_j},\varsigma\right)-u_{ij}\\ r_{gi} &= Log\left({}^{M}_{G}T\left(t_i+t_d\right){}^{G}_{W}T\,{}^{W}_{C_i}T\,{}^{C}_{M}T\right)\end{aligned} \tag{3}$$

Where K is the corner numbers for each image. $\rho\left(\bullet\right)$ is a robust kernel function [4]. $\pi\left(\bullet,\bullet\right)$ is a fixed camera projection function [12, 30]. $Log\left(\bullet\right)$ maps the element on a Lie group to the tangent space vector [27].

${}^{M}_{G}T\left(t_i+t_d\right)$ 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:

$$\begin{aligned}&{}^{M}_{G}T\left(t_i+t_d\right) = Exp\left(\lambda Log\left({}^{M_b}_{G}T\,{}^{M_a}_{G}T^{-1}\right)\right){}^{M_a}_{G}T\\ &\lambda = \left(t_i+t_d-t_a\right)/\left(t_b-t_a\right)\end{aligned} \tag{4}$$

$Exp\left(\bullet\right)$ is the inverse operation of $Log\left(\bullet\right)$ [27].

Jacobians of residuals in Eq. (3) with respect to the optimization variables χ are calculated according to the chain rule and provided in Sec. 8 of supplementary material. The Levenberg-Marquardt algorithm is adpot to minimize Eq. (3) and update the optimal estimation iteratively.

Differentiate from [10], 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.

4. 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. 2b.

4.1. State Vector

The EKF state vector inspired by MSCKF [19] includes the marker state, the spatial-temporal calibration parameters, the camera intrinsic parameters, augmented N marker states and up to L augmented features:

$$\begin{aligned}x &= \begin{bmatrix} x_M^T & x_{calib}^T & x_c^T & x_f^T \end{bmatrix}^T\\ x_M &= \begin{bmatrix} {}^{M}_{G}q^T & {}^{G}p_M^T & \omega^T & {}^{G}v_M^T \end{bmatrix}^T\\ x_{calib} &= \begin{bmatrix} {}^{C}_{M}q^T & {}^{C}p_M^T & t_d & \varsigma \end{bmatrix}^T\\ x_c &= \begin{bmatrix} x_{c_1}^T & \cdots & x_{c_N}^T \end{bmatrix}^T \qquad x_{c_i} = \begin{bmatrix} {}^{M_i}_{G}q^T & {}^{G}p_{M_i}^T \end{bmatrix}^T\\ x_f &= \begin{bmatrix} {}^{G}p_{f_1}^T & \cdots & {}^{G}p_{f_L}^T \end{bmatrix}^T\end{aligned} \tag{5}$$

Where x_M is the current marker state at the camera clock. Calibration parameter x_{calib} includes the 6DoF transformation $\left\{ {}^{C}_{M}q \quad {}^{C}p_M \right\}$, 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 [11, 14, 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. (9)).

4.2. Constant Velocity Propagation

Referring to previous study on trajectory estimation [7, 25], 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:

$$\begin{aligned}&{}^{M}_{G}\dot{q} = \tfrac{1}{2}\Omega\left(\omega\right){}^{M}_{G}q, \quad {}^{G}\dot{p}_M = {}^{G}v_M\\ &\dot{\omega} = n_\omega, \quad {}^{G}\dot{v}_M = n_v\end{aligned} \tag{6}$$

$\Omega\left(\omega\right) = \begin{bmatrix} -[\omega]_\times & \omega \\ -\omega^T & 0 \end{bmatrix}$. $n_{[\bullet]}$ represents the zero mean Gaussian noise of $[\bullet]$, which is a hyperparameter. These hyperparameters can be determinated in advance using existing approaches [2, 7].

By linearizing Eq. (6) at the current state estimation, the state transition matrix from time t_0 to time t_k can be analytically calculated as follows:

$$\begin{aligned}\Phi_M\left(t_k,t_0\right) &= \begin{bmatrix} A & 0_3 & B & 0_3 \\ 0_3 & I_3 & 0_3 & I_3\Delta t \\ 0_3 & 0_3 & I_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & I_3 \end{bmatrix}\\ A &= {}^{M_k}_{G}R\,{}^{M_0}_{G}R^T\\ B &= {}^{M_k}_{G}R\,{}^{M_0}_{G}R^T J_r\left(-\omega\Delta t\right)\Delta t\end{aligned} \tag{7}$$

Where $J_r\left(\bullet\right)$ is the right Jacobian of SO(3) [1].

4.3. 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. (1), the corresponding marker timestamp is $t+t_d$. The new cloned marker pose is:

$$x_{c_{new}} = \begin{bmatrix} {}^{M}_{G}q\left(t+t_d\right) \\ {}^{G}p_M\left(t+t_d\right) \end{bmatrix} \tag{8}$$

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:

$$H_{aug} = \begin{bmatrix} I_3 & 0_3 & \omega \\ 0_3 & I_3 & {}^{G}v_M \end{bmatrix} \tag{9}$$

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:

$$\begin{aligned} z_f &= \pi\left({}^{C}p_f, \varsigma\right) \\ {}^{C}p_f &= {}^{C}_{M}R^{M}_{G}R\left({}^{G}p_f - {}^{G}p_M\right) + {}^{C}p_M \end{aligned} \tag{10}$$

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

$$x_s = \begin{bmatrix} {}^{M}_{G}q^T & {}^{G}p^T_M & {}^{C}_{M}q^T & {}^{C}p^T_M & {}^{G}p^T_f \end{bmatrix}^T \tag{11}$$

The feature measurement Jacobian is calculated as:

$$\begin{aligned} H_f &= \frac{\partial z_f}{\partial {}^{C}p_f}{}^{C}_{M}R^{M}_{G}R \begin{bmatrix} J_1 & -I_3 & J_2 & {}^{G}_{M}R^{M}_{C}R & I_3 \end{bmatrix} \\ J_1 &= \left[\left({}^{G}p_f - {}^{G}p_M\right)\right]_{\times}{}^{G}_{M}R \\ J_2 &= \left[\left({}^{G}p_f - {}^{G}p_M\right)\right]_{\times}{}^{G}_{M}R{}^{M}_{C}R \end{aligned} \tag{12}$$

More details about feature detection, tracking, outlier rejection, triangulation, sliding window update scheme and covariance management can be found in [11].

4.4. 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:

$$z_g = \begin{bmatrix} {}^{M}_{G}q \\ {}^{G}p_M \end{bmatrix} \tag{13}$$

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

$$H_g = \begin{bmatrix} I_3 & 0_3 \\ 0_3 & I_3 \end{bmatrix} \tag{14}$$

5. 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 [5] derived 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.

[1]The camera intrinsic ς is omitted here because it does not affect the subsequent observability analysis in Sec. 5.

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 [13, 15]. The system state vector becomes:

$$x = \begin{bmatrix} {}^{M}_{G}q^T & {}^{G}p^T_M & {}^{C}_{M}q^T & {}^{C}p^T_M & t_d & {}^{G}p^T_f \end{bmatrix}^T \tag{15}$$

The state transition matrix becomes:

$$\Phi\left(t_k, t_0\right) = \begin{bmatrix} A & \\ & I_{13} \end{bmatrix} \tag{16}$$

A is defined in Eq. (7).

H_{aug} in Eq. (9), H_f in Eq. (12) and H_g in Eq. (14) are stacked to construct the general Jacobian of the state:

$$H_k = \begin{bmatrix} I_3 & 0_3 & 0_3 & 0_3 & \omega & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & {}^{G}v_M & 0_3 \\ J_1 & -I_3 & J_2 & {}^{G}_{M}R^{M}_{C}R & 0_{3\times1} & I_3 \\ I_3 & 0_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \end{bmatrix} \tag{17}$$

The common factor $\frac{\partial z_f}{\partial {}^{C}p_f}{}^{C}_{M}R^{M}_{G}R$ in Eq. (12) is ignored here because it does not affect the observability analysis. Now the observability matrix would be constructed as [5]:

$$\begin{aligned} O &= \begin{bmatrix} \cdots & O^T_k & \cdots \end{bmatrix}^T \\ O_k &= H_k\Phi\left(t_k, t_0\right) \\ &= \begin{bmatrix} A & 0_3 & 0_3 & 0_3 & \omega & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & {}^{G}v_M & 0_3 \\ J_1A & -I_3 & J_2 & {}^{G}_{M}R^{M}_{C}R & 0_{3\times1} & I_3 \\ A & 0_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_{3\times1} & 0_3 \end{bmatrix} \end{aligned} \tag{18}$$

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

$$N_1 = \begin{bmatrix} 0_{3\times9} & I_3 & 0_{3\times1} & -\left({}^{G}_{M}R^{M}_{C}R\right)^T \end{bmatrix}^T \tag{19}$$

Proof. The fact that N_1 is indeed the right null space of O can be verified by multiplying O_k with N_1. $O_kN_1 = 0$ 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 5.2. *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:*

$$N_2 = \begin{bmatrix} 0_{1\times 9} & \left({}^C_M R\omega_2\right)^T & 0 & -\left({}^G_M R\omega_2\right)^T \end{bmatrix}^T \quad (20)$$

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:

$$\frac{d\left({}^G_M R\omega_2\right)}{dt} = \left(\frac{d\left({}^G_M R\right)}{dt}\right)\omega_2 = {}^G_M R[\omega]_\times \omega_2 = 0 \quad (21)$$

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 5.2. 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.

6. 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. 5 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.

6.1. 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. 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 [17, 18, 32].

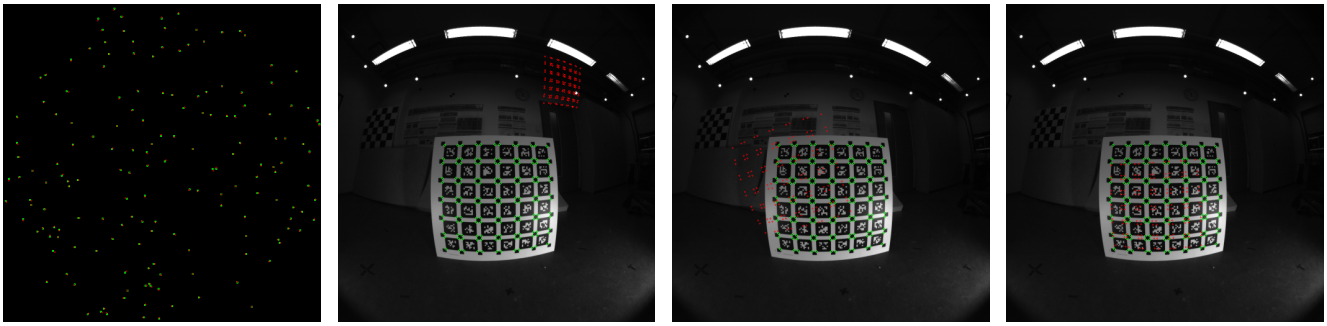
To validate the observability assertion in Sec. 5, we set ${}^C_M R$ as I_3, and design five rotation motion cases.

- 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. 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 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. 4. The similar analysis also applies to Case 4 and Case 5.

6.2. 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. 6.1 show that it is advised to choose the fully excited 6DoF trajectory. The experiments in [32] also inspire us to utilize the fully excited hand-held TUM-VI Dataset [26] instead of under-actuated dataset, such as EuRoC MAV Dataset [3]. TUM-VI Dataset contains multiple sequences with or without 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



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

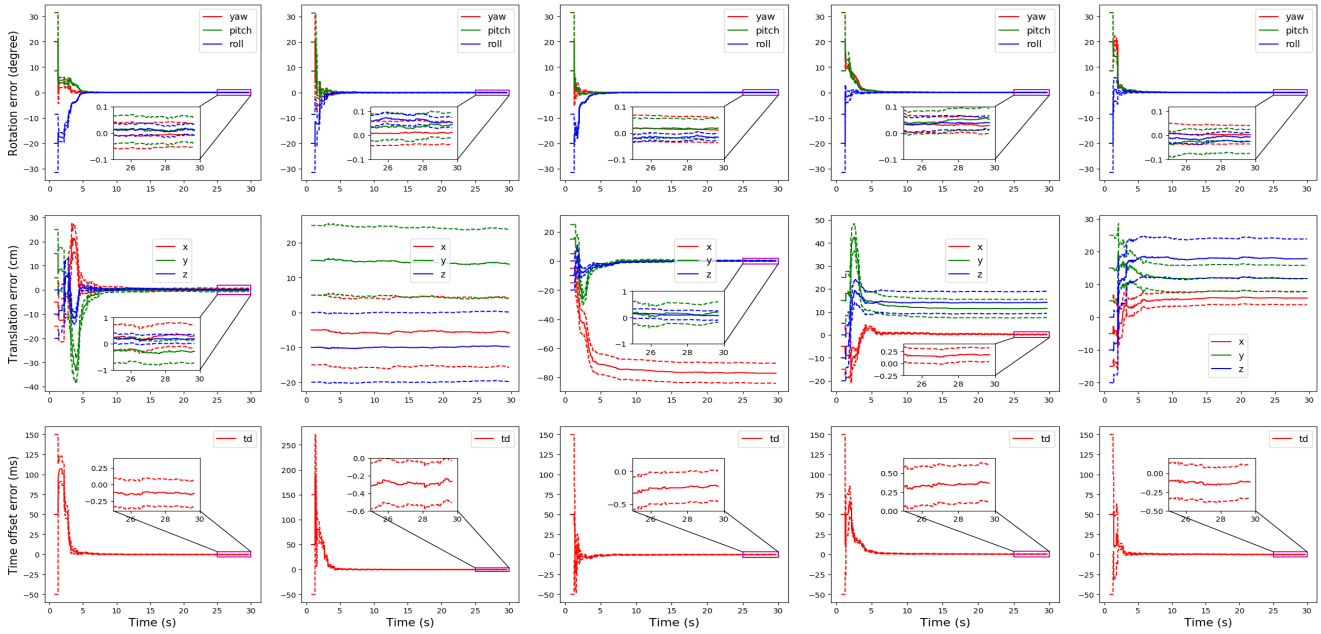


Figure 4. Errors (solid lines) and 1σ bounds (dashed lines) of the spatial-temporal calibration parameters. x-axis represents time in seconds. Left to right corresponds to Case1 to Case5 in Sec. 6.1. The estimation error of the rotation and temporal calibration parameters perfectly approach to zero for any cases. While the convergence results of the translation calibration parameter are varied from case to case.

is the reference value of the temporal parameter. As [26] has leveraged IMU to align the marker frame to the IMU frame, the transformation from IMU to camera [28], is also the reference value of the interested spatial parameter.

Table 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

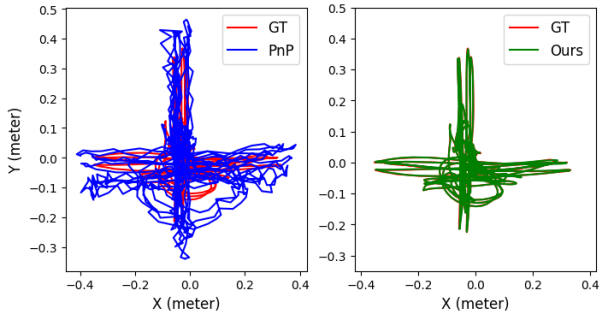


Figure 5. $imu1$ is used. GT: groundtruth trajectory output from motion capture system. PnP: camera trajectory output from PnP algorithm. Ours: refined camera trajectory ${}^{W}_{C_i}T, i = 1 \cdots N$.

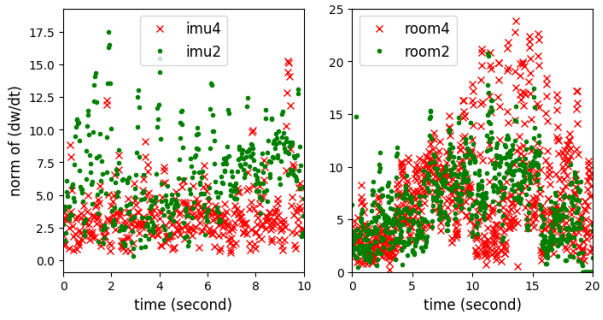


Figure 6. Norm of $d\omega/dt$.

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.

6.2.1 Environments with target

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

[10] can not work for these sequences due to the relatively large trajectory noise output by PnP algorithm, as shown in Fig. 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. 9 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:

$$\begin{aligned} z &= \pi\left({}^{C}p_f, \varsigma\right) \\ {}^{C}p_f &= {}^{C}_{M}T\,{}^{M}_{G}T\,(t+t_d)\,{}^{G}_{W}T\,{}^{W}p_f \end{aligned} \tag{22}$$

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. (2).

For a specific run of the target-based method, the iterative update results are visualized from Fig. 3b to Fig. 3f. After 5 iterations, all predicted feature positions are perfectly close to expected feature positions. Fig. 3g shows the feature points update of the target-less method. The predicted feature position is obtained via Eq. (10).

When using the left camera, the RMSE of the calibration results are shown in Tab. 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. 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

Table 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

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. 5 and Sec. 6.1, it is reasonable to examine the rotation excitation to reveal the behind reason. Fig. 6 depicts the norm of the angular velocity difference. $imu2$ has more sufficient rotation excitation, improving the observability of the translation calibration parameter.

6.2.2 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. 2. Compared with the sequence with calibration target (see Tab. 1), the estimation of the calibration parameter 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. 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 [8] 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.

$$ {}^{M'}_{G}T = {}^{M'}_{M}T\,{}^{M}_{G}T \qquad {}^{M'}_{C}T = {}^{M'}_{M}T\,{}^{M}_{C}T \tag{23} $$

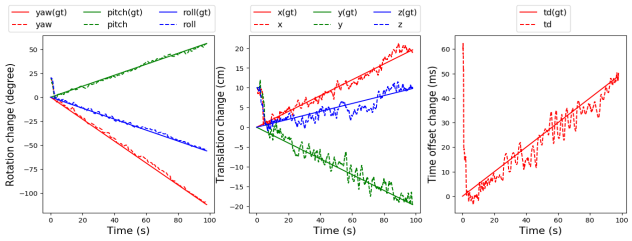


Figure 7. Groundtruth (solid lines) and estimation (dashed lines) of the time-varying change of the spatial-temporal parameters.

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. 7, the time-varying quantity of spatial-temporal parameter is designed to 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.

7. CONCLUSIONS

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.

References

[1] Timothy D Barfoot. *State estimation for robotics*. Cambridge University Press, 2017. 3, 1

[2] 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*, pages 1–10. Citeseer, 2014. 3

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35 (10):1157–1163, 2016. 1, 5

[4] Nived Chebrolu, Thomas Läbe, Olga Vysotska, Jens Behley, and Cyrill Stachniss. Adaptive robust kernels for non-linear least squares problems. *IEEE Robotics and Automation Letters*, 6(2):2240–2247, 2021. 3

[5] 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*, pages 100–103. IEEE, 1990. 4

[6] 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)*, pages 6713–6719. IEEE, 2019. 1

[7] 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)*, pages 6497–6504. IEEE, 2018. 3

[8] Davide Falanga, Kevin Kleber, Stefano Mintchev, Dario Floreano, and Davide Scaramuzza. The foldable drone: A morphing quadrotor that can squeeze and fly. *IEEE Robotics and Automation Letters*, 4(2):209–216, 2018. 8

[9] 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*, pages 1280–1286. IEEE, 2013. 1

[10] 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*, pages 145–159. Springer, 2018. 1, 3, 7, 2

[11] 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)*, pages 4666–4672. IEEE, 2020. 3, 4

[12] 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*, pages 1793–1800. IEEE, 2013. 3, 1

[13] Joel A Hesch, Dimitrios G Kottas, Sean L Bowman, and Stergios I Roumeliotis. Consistency analysis and improvement of vision-aided inertial navigation. *IEEE Transactions on Robotics*, 30(1):158–176, 2013. 4

[14] Mingyang Li. *Visual-inertial odometry on resource-constrained systems*. University of California, Riverside, 2014. 3

[15] Mingyang Li and Anastasios I Mourikis. High-precision, consistent ekf-based visual-inertial odometry. *The International Journal of Robotics Research*, 32(6):690–711, 2013. 4

[16] Mingyang Li and Anastasios I Mourikis. Optimization-based estimator design for vision-aided inertial navigation. In *Robotics: Science and Systems*, pages 241–248. Berlin Germany, 2013. 3

[17] Mingyang Li and Anastasios I Mourikis. Online temporal calibration for camera–imu systems: Theory and algorithms. *The International Journal of Robotics Research*, 33(7):947–964, 2014. 2, 5

[18] Jiajun Lv, Xingxing Zuo, Kewei Hu, Jinhong Xu, Guoquan Huang, and Yong Liu. Observability-aware intrinsic and extrinsic calibration of lidar-imu systems. *IEEE Transactions on Robotics*, 2022. 5

[19] 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*, pages 3565–3572. IEEE, 2007. 3

[20] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011. 2

[21] 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)*, pages 3662–3669. IEEE, 2018. 2

[22] Kejie Qiu, Tong Qin, Jie Pan, Siqi Liu, and Shaojie Shen. Real-time temporal and rotational calibration of heterogeneous sensors using motion correlation analysis. *IEEE Transactions on Robotics*, 37(2):587–602, 2020. 2

[23] 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)*, pages 4304–4311. IEEE, 2016. 1

[24] Joern Rehder, Roland Siegwart, and Paul Furgale. A general approach to spatiotemporal calibration in multisensor systems. *IEEE Transactions on Robotics*, 32(2):383–398, 2016. 1

[25] 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)*, pages 682–697, 2018. 3

[26] 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)*, pages 1680–1687. IEEE, 2018. 1, 5, 6, 2

[27] Joan Sola, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. 3

[28] 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*, pages 11148–11156, 2020. 6

[29] Nikolas Trawny and Stergios I Roumeliotis. Indirect kalman filter for 3d attitude estimation. *University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep*, 2:2005, 2005. 2

[30] Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. The double sphere camera model. In *2018 International Conference on 3D Vision (3DV)*, pages 552–560. IEEE, 2018. 3, 1, 2

[31] Yulin Yang, Patrick Geneva, Kevin Eckenhoff, and Guoquan Huang. Degenerate motion analysis for aided ins with online spatial and temporal sensor calibration. *IEEE Robotics and Automation Letters*, 4(2):2070–2077, 2019. 2

[32] Yulin Yang, Patrick Geneva, Xingxing Zuo, and Guoquan Huang. Online imu intrinsic calibration: Is it necessary? In *Robotics: Science and Systems*, 2020. 2, 5

Joint Spatial-Temporal Calibration for Camera and Global Pose Sensor

Supplementary Material

8. Analytical on-manifold Jacobians for the target-based method

The optimization function (Eq. (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.

8.1. Jacobians of pixel measurement residual

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

$$\begin{aligned} r_{ij} &= \pi\left({}^{C_i}p_{f_j}, \varsigma\right) - u_{ij} \\ {}^{C_i}p_{f_j} &= {}^{C_i}_{W}T\,{}^{W}p_{f_j} \end{aligned} \tag{24}$$

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

$$\chi_{s_1} = \left\{ \begin{array}{cc} {}^{W}_{C_i}T & \varsigma \end{array} \right\} \tag{25}$$

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 [12, 30]. The Jacobian of the pixel residual r_{ij} with respect to the camera pose ${}^{W}_{C_i}T$ is:

$$\begin{aligned} \frac{\partial r_{ij}}{\partial {}^{W}_{C_i}T} &= \frac{\partial r_{ij}}{\partial {}^{C_i}p_{f_j}} \frac{\partial {}^{C_i}p_{f_j}}{\partial {}^{C_i}_{W}T} \frac{\partial {}^{C_i}_{W}T}{\partial {}^{W}_{C_i}T} \\ \frac{\partial {}^{C_i}p_{f_j}}{\partial {}^{C_i}_{W}T} &= \left({}^{C_i}_{W}T\,{}^{W}p_{f_j}\right)^{\odot} \\ \frac{\partial {}^{C_i}_{W}T}{\partial {}^{W}_{C_i}T} &= -I \end{aligned} \tag{26}$$

Where $\odot$ is an operator for the homogeneous coordinate [1, 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. (26) and $\frac{\partial r_{ij}}{\partial \varsigma}$.

8.2. Jacobians of global pose measurement residual

Next, we analyze the Jacobians involved in the global pose measurement residual r_{gi} (Eq. (3)). To simplify the description, we define the following intermediate quantities:

$$\begin{aligned} {}^{M}_{G}\hat{T} &\triangleq {}^{M}_{G}T\left(t_i + t_d\right) \\ {}^{W}_{C}T &\triangleq {}^{W}_{C_i}T \\ {}^{M_b}_{M_a}\theta &\triangleq Log\left({}^{M_b}_{G}T\,{}^{M_a}_{G}T^{-1}\right) \end{aligned} \tag{27}$$

Therefore

$$\begin{aligned} r_{gi} &= Log\left({}^{M}_{G}\hat{T}\,{}^{G}_{W}T\,{}^{W}_{C}T\,{}^{C}_{M}T\right) \\ {}^{M}_{G}\hat{T} &= Exp\left(\lambda\,{}^{M_b}_{M_a}\theta\right){}^{M_a}_{G}T \\ \lambda &= (t_i + t_d - t_a)/(t_b - t_a) \end{aligned} \tag{28}$$

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

$$\chi_{s_2} = \left\{ \begin{array}{cccc} {}^{W}_{C_i}T & {}^{G}_{W}T & {}^{C}_{M}T & t_d \end{array} \right\} \tag{29}$$

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

$$\frac{\partial r_{gi}}{\partial {}^{C}_{M}T} = J_r^{-1}\left(r_{gi}\right) \tag{30}$$

Where $J_r\left(\bullet\right)$ is the right Jacobian of SE(3) [1].

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

$$\frac{\partial r_{gi}}{\partial {}^{W}_{C}T} = J_r^{-1}\left(r_{gi}\right) Ad\left({}^{C}_{M}T^{-1}\right) \tag{31}$$

Where $Ad\left(\bullet\right)$ is the adjoint of SE(3) [1].

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

$$\frac{\partial r_{gi}}{\partial {}^{G}_{W}T} = J_r^{-1}\left(r_{gi}\right) Ad\left(\left({}^{W}_{C}T\,{}^{C}_{M}T\right)^{-1}\right) \tag{32}$$

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

$$\frac{\partial r_{gi}}{\partial {}^{M}_{G}\hat{T}} = J_r^{-1}\left(r_{gi}\right) Ad\left(\left({}^{M}_{G}\hat{T}\,{}^{G}_{W}T\,{}^{W}_{C}T\,{}^{C}_{M}T\right)^{-1}\right) \tag{33}$$

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

$$\frac{\partial {}^{M}_{G}\hat{T}}{\partial \lambda} = Ad\left(Exp\left(\lambda\,{}^{M_b}_{M_a}\theta\right)\right) J_r\left(\lambda\,{}^{M_b}_{M_a}\theta\right){}^{M_b}_{M_a}\theta \tag{34}$$

The Jacobian of λ with respect to t_d is:

$$\frac{\partial \lambda}{\partial t_d} = \frac{1}{t_b - t_a} \tag{35}$$

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

$$\frac{\partial r_{gi}}{\partial t_d} = \frac{\partial r_{gi}}{\partial {}^{M}_{G}\hat{T}} \frac{\partial {}^{M}_{G}\hat{T}}{\partial \lambda} \frac{\partial \lambda}{\partial t_d} \tag{36}$$

In summary, the Jacobians of the global pose measurement residual r_{gi} with respect to χ_{s_2} can be computed via Eq. (30), Eq. (31), Eq. (32) and Eq. (36).

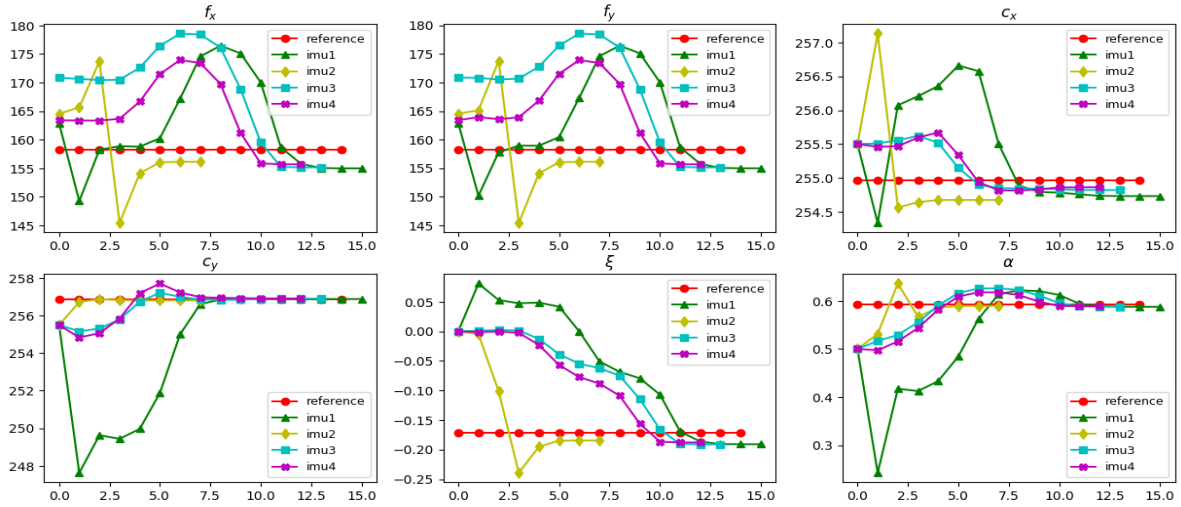


Figure 8. Iterative process of calibrating left camera intrinsic from scratch. x-axis represents iteration steps.

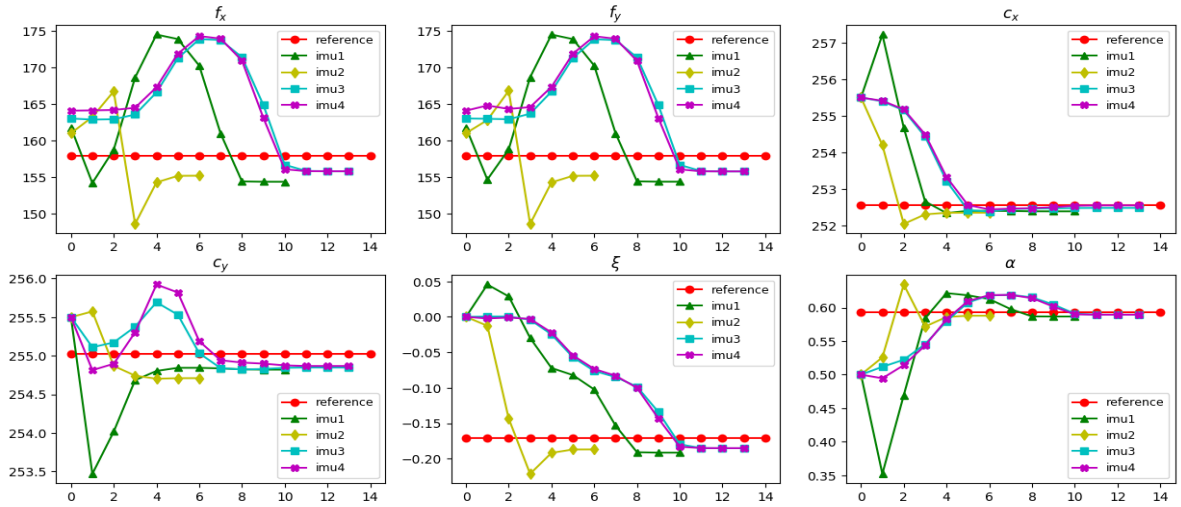


Figure 9. Iterative process of calibrating right camera intrinsic from scratch. x-axis represents iteration steps.

9. Additional comparison results

Compared to [10], our proposed target-based method has another benefit, in addition to iterative optimization of camera trajectory. Prior to perform spatial-temporal hand-eye calibration, [10] 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, [10] 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. 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 [26], and double sphere camera model [30] is adopted. Regarding the initialization method and reference values for camera intrinsic parameters, we refer to [30]. In Fig. 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. 9. Final average reprojection error and position error in Eq. (3) are smaller than 0.1 pixel and 0.1 cm for left and right camera from each sequence. Results from Fig. 8 and Fig. 9 demonstrate the ability of calibrating optimal camera intrinsic from scratch for each sequence with the target-based method.