Closed-form solution to cooperative visual-inertial structure from motion

Agostino Martinelli
INRIA, Grenoble, France
agostino.martinelli@inria.fr

February 26, 2018

Abstract

This paper considers the problem of visual-inertial sensor fusion in the cooperative case and it provides new theoretical contributions, which regard its observability and its resolvability in closed form. The case of two agents is investigated. Each agent is equipped with inertial sensors (accelerometer and gyroscope) and with a monocular camera. By using the monocular camera, each agent can observe the other agent. No additional camera observations (e.g., of external point features in the environment) are considered. All the inertial sensors are assumed to be affected by a bias. First, the entire observable state is analytically derived. This state includes the absolute scale, the relative velocity between the two agents, the three Euler angles that express the rotation between the two agent frames and all the accelerometer and gyroscope biases. Second, the paper provides the extension of the closed-form solution given in [19] (which holds for a single agent) to the aforementioned cooperative case. The impact of the presence of the bias on the performance of this closed-form solution is investigated. As in the case of a single agent, this performance is significantly sensitive to the presence of a bias on the gyroscope, while, the presence of a bias on the accelerometer is negligible. Finally, a simple and effective method to obtain the gyroscope bias is proposed. Extensive simulations clearly show that the proposed method is successful. It is amazing that, it is possible to automatically retrieve the absolute scale and simultaneously calibrate the gyroscopes not only without any prior knowledge (as in [13]), but also without external point features in the environment.

Keywords: Visual Inertial aided Navigation; Closed form solution; Bias Calibration
Contents

1 Introduction 3

2 The system 4

3 Observable state and its dynamics 5
  3.1 The unbiased case ........................................ 5
  3.2 The biased case ........................................... 6

4 Closed-form solution 7
  4.1 The solution .............................................. 7
    4.1.1 Single camera ....................................... 8
    4.1.2 Two cameras synchronized ............................ 8
  4.2 Analytic Derivation ...................................... 10
    4.2.1 Single camera ....................................... 12
    4.2.2 Two cameras synchronized ............................ 12

5 Limitations of the Closed-Form Solution 13
  5.1 Simulation setup ........................................ 13
  5.2 Performance without bias ................................ 14
  5.3 Impact of accelerometer bias on the performance ......... 14
  5.4 Impact of gyroscope bias on the performance ............ 15

6 Estimating the Gyroscope Bias ($B_1^\Omega$ and $B_2^\Omega$) 15

7 Performance Overall Evaluation 17

8 Conclusion 17

A Observability Analysis 20

B Fundamental Equations 22

C Observability with bias 23
1 Introduction

The problem of fusing visual and inertial data has been extensively investigated in the past (e.g., [1, 2, 7, 12, 13]). Recently, this sensor fusion problem has been successfully addressed by enforcing observability constraints [6, 9], and by using optimization-based approaches [3, 8, 11, 13, 16, 21, 22]. These optimization methods outperform filter-based algorithms in terms of accuracy due to their capability of relinearizing past states. On the other hand, the optimization process can be affected by the presence of local minima. For this reason, a deterministic solution able to automatically determine the state without initialization has been introduced [13, 18, 19].

Visual and inertial sensors have also been used in a cooperative scenario (e.g., for cooperative mapping in [4]). However, in the cooperative case, a deterministic solution able to automatically determine the state without initialization (as in [13, 18, 19]), still misses. The goal of this paper is precisely to address this lack. We relax the assumption that one or more features are available from the environment: we investigate the extreme case where no point features are available. Additionally, we consider the critical case of only two agents. In other words, we are interested in investigating the minimal case. If we prove that the absolute scale is observable, we can conclude that it is observable in all the other cases. Each agent is equipped with an Inertial Measurement Unit (IMU) and a monocular camera. By using the monocular camera, each agent can observe the other one. Note that, we do not assume that these camera observations contain metric information (due for instance to the known size of the observed agent). The two agents can operate far from each other and a single camera observation only consists of the bearing of the observed agent in the frame of the observer. In other words, each agent acts as a moving point feature with respect to the other agent. The system is defined in section 2.

The first questions we wish to answer are: Is it possible to retrieve the absolute scale in these conditions? And the absolute roll and pitch angles? More in general, we want to determine the entire observable state, i.e., all the physical quantities that it is possible to determine by only using the information contained in the sensor data (from the two cameras and the two IMUs) during a short time interval. In [20] we provided the answers to these questions in the case when the inertial measurements are unbiased. These results are summarized in the first part of section 3. Then, in the second part of this section, we provide a full answer even in presence of biased measurements (both the ones from the accelerometers and the ones from the gyroscopes).

Section 4 provides a closed-form solution, able to obtain the observable state by only using visual and inertial measurements from the two agents delivered during a short time interval. This is precisely the extension of the closed form solution in [13, 19] to the cooperative case. Then, the paper demonstrates the efficiency of this solution. By nature, a closed-form solution is deterministic and, thus, does not require any initialization. We perform simulations with plausible MAV motions and synthetic noisy sensor data (section 5). This allows us to identify limitations of the solution and bring modifications to overcome them. In practice, we perform exactly the same investigation done in [13] for the case of a single agent. Specifically, we investigate the impact of biased inertial measurements. We show that a large bias on the accelerometer does not significantly worsen the performance (section 5.3). One major limitation is the impact of biased gyroscope measurements (section 5.4). In other words, the performance becomes very poor in presence of a bias on the gyroscopes of the two agents and, in practice, the overall method can only be successfully used with very precise - and expensive - gyroscopes. In section 6, we introduce a simple method that automatically estimates both these biases. By adding this new method for the bias estimation to the solution presented in section 4, we obtain results that are equivalent to the ones in absence of bias (section 7).

This method is suitable for MAVs that operate in absence of point features and need to recover the absolute scale in few seconds. The solution does not need initialization. Additionally,
is robust to the bias and automatically calibrates the gyroscopes.

2 The system

We consider two agents that move in a 3D–environment. Each agent is equipped with an Inertial Measurement Unit (IMU), which consists of three orthogonal accelerometers and three orthogonal gyroscopes. Additionally, each agent is equipped with a monocular camera. We assume that, for each agent, all the sensors share the same frame. Without loss of generality, we define the agent local frame as this common frame. The accelerometer sensors perceive both the gravity and the inertial acceleration in the local frame. The gyroscopes provide the angular speed in the local frame. Finally, the monocular camera mounted on the first agent (or the second agent) provides the bearing of the second agent (or the first agent) in its local frame.

Figure 1: The global frame and the two local frames (attached to the first and to the second agent, respectively). \( r_1 \) and \( r_2 \) are their position, expressed in the global frame. \( R \) is the relative position of the second agent in the local frame of the first agent.

We adopt the following notations:

- \( R \) is the position of the second agent in the local frame of the first agent;
- \( V \) is the relative velocity of the second agent with respect to the first agent, expressed in the frame of the first agent (note that this velocity is not simply the time derivative of \( R \) because of the rotations accomplished by the first agent);
• $O$ is the orthonormal matrix that characterizes the rotation between the two local frames; specifically, for a vector with given coordinates in the local frame of the second agent, we obtain its coordinates in the local frame of the first agent by pre multiplying by $O$.

Additionally, we denote by $A^1$, $\Omega^1$ the acceleration and the angular speed of the first agent expressed in the first agent local frame. Similarly, we denote by $A^2$, $\Omega^2$ the acceleration and the angular speed of the second agent expressed in the second agent local frame. Regarding the accelerations, they include both the inertial acceleration and the gravity. We denote by $B^1_A$, $B^2_A$, $B^1_\Omega$ and $B^2_\Omega$ the biases that affect the measurements from the two accelerometers and the two gyroscopes.

We conclude this section by providing the analytic expressions of the two camera observations. The first camera provides the vector $R$, up to a scale. The scale is precisely the distance between the first agent and the second agent at the time of the camera measurement. The second camera provides the vector $-O^T R$, up to a scale. Note that, in the special case when the two cameras are synchronized, the scale coincides.

3 Observable state and its dynamics

3.1 The unbiased case

In appendix A we analyzed the observability properties of the system defined in section 2, in the case when the inertial measurements are unbiased. We showed that all the observable states are:

• the three components of $R$;
• the three components of $V$;
• the three Euler angles that uniquely define the matrix $O$.

Basically, this result tells us that, starting from the measurements delivered by the two IMUs and the two cameras during a given time interval, we can reconstruct the previous three quantities. Estimating other physical quantities (e.g., the absolute roll and pitch angles of the first agent or the second agent) is not possible. Note that we can retrieve the absolute scale, even when no feature is available in the environment. In appendix A we also proven that, the same observable state, characterizes the case when only one of the agents is equipped with a camera. Namely, the presence of two cameras does not improve the observability properties with respect to the case of a single camera mounted on one of the two agents. In appendix B we derived the analytic expression of the dynamics of the observable states. They are:

$$\begin{align*}
\dot{R} &= [\Omega^1]_\times R + V \\
\dot{V} &= [\Omega^1]_\times V + O A^2 - A^1 \\
\dot{O} &= [\Omega^1]_\times O + O [\Omega^2]_\times
\end{align*}$$

(1)

where $[\Omega^{1/2}]_\times$ are the skew-symmetric matrices associated to the vectors $\Omega^{1/2}$:

$$
[\Omega^i]_\times = \begin{bmatrix}
0 & \Omega^i_z & -\Omega^i_y \\
-\Omega^i_z & 0 & \Omega^i_x \\
\Omega^i_y & -\Omega^i_x & 0
\end{bmatrix}, \quad i = 1, 2
$$

(2)

Finally, the two cameras provide the two vectors, $R$ and $-O^T R$, up to a scale.
3.2 The biased case

The goal of this subsection is to obtain the observable state when the inertial measurements are corrupted by the biases $B_1^1$, $B_1^2$, $B_1^3$, and $B_2^2$. Since the presence of the bias cannot improve the observability properties, we characterize our system by including in the observable state that holds in absence of bias (i.e., the state mentioned in the previous subsection), all the 12 components of the 4 bias vectors. If we prove that this state is observable, we can conclude that it is the entire observable state, i.e., any other physical quantity independent from its components is unobservable. Additionally, we will consider the case when only the first agent is equipped with a camera. Again, by proving that in these conditions the previous state is observable, we can conclude that the same observable state characterizes the case of two cameras.

The bias is time dependent. However, it changes very slowly with time. In particular, in the interval of few seconds, it can be considered constant. Since we will consider time intervals no longer than 4 seconds (and, as it will be shown, this will allow us to auto calibrate the inertial sensors with very high accuracy), we can assume that the bias is constant during the considered time interval (the same assumption is made in [13]).

We characterize our system by the following state:

$$X = [R \ V \ q \ B_1^1 \ B_1^2 \ B_2^2]^T \quad (3)$$

where $q$ is the unit quaternion associated to the rotation matrix $O$.

The dimension of this state is equal to 22. Actually, the components of this state are not independent, since $q$ is a unit quaternion. In other words, we have: $(q_t)^2 + (q_x)^2 + (q_y)^2 + (q_z)^2 = 1$.

The dynamics of the state defined in (27) are given by the following equations:

$$\begin{align*}
\dot{R} &= [\Omega^1]^T \times R + V \\
\dot{V} &= [\Omega^1]^T V + O \dot{A} - A' \\
\dot{q} &= -\frac{1}{2} \Omega^1 q + \frac{1}{2} q \Omega^2 \\
\dot{B}_1^1 &= \dot{B}_1^2 = \dot{B}_2^1 = \dot{B}_2^2 = 0
\end{align*}$$

(4)

where:

- $\Omega^1 = \Omega^1 + B_1^1$, $A' = A^1 + B_1^1$, $\Omega^2 = \Omega^2 + B_2^2$, $A' = A^2 + B_2^2$.

- The matrix $O$, can be uniquely expressed in terms of the components of the quaternion $q$.

- $\Omega_1^1$ is the imaginary quaternion associated with $\Omega^1$, i.e.: $\Omega_1^1 = 0 + \Omega_1^1 i + \Omega_1^1 j + \Omega_1^1 k$. The same holds for $\Omega_1^2$.

The observation functions are the two scalar functions $h_u$, $h_v$:

$$h \triangleq [h_u, h_v]^T = \begin{bmatrix} R_x & R_y \\ R_z & R_z \end{bmatrix}$$

(5)

Additionally, we need to add the observation function that expresses the constraint that $q$ is a unit quaternion. The additional observation is:

$$h_{const}(X) \triangleq (q_t)^2 + (q_x)^2 + (q_y)^2 + (q_z)^2$$

(6)

The analytic derivation of this system observability is provided in appendix C. We summarize its result:
For the system defined in section 2 the observable state is the one given in (27). This holds both in the case when both the agents are equipped with a camera and in the case when only one agent is equipped with a camera. Additionally, the observable state remains the same even in the case when the camera is a linear camera, i.e., it only provides the azimuth of the other agent in its local frame.

4 Closed-form solution

In this section we will refer to the two agents as to two rigid bodies and we denote them by $B_1$ and $B_2$. This is to emphasize the generality of the closed form solution here introduced, which is not limited to the framework of aerial navigation.

Let us consider a given time interval $(t_A, t_B)$. Let us denote by $R_A, V_A$ and $O_A$, the values of $R, V$ and $O$ at time $t_A$. The goal of this section is to obtain these values in closed form, only in terms of the measurements provided during the considered time interval (both from the cameras and the two IMUs). Note that, the length of the considered time interval (i.e., $t_B - t_A$) is very small (about 2, 3 seconds).

In section 4.1 we directly provide the solution. All the analytic derivations are given later, in section 4.2.

4.1 The solution

We distinguish the case when only $B_1$ is equipped with a camera, from the case when both $B_1$ and $B_2$ are equipped with a camera and the observations are synchronized. In particular, we assume that, during our time interval, the camera (or the two cameras) performs $n$ observations at the times $t_j, (j = 1, \cdots, n)$, with $t_1 = t_A$ and $t_n = t_B$. In both cases (single camera and two cameras synchronized), we will obtain the components of $R_A, V_A$ and $O_A$ by simply solving the linear system:

$$\Xi x = b$$

where:

- $\Xi$ is a matrix with dimension $3n \times (15 + n)$ in the case of a single camera and $6n \times (21 + n)$ in the case of two cameras synchronized.

- $x$ is the vector that contains all the unknowns; in the case of a single camera it contains the components of $R_A, V_A$ and $O_A$ and the $n$ distances between $B_1$ and $B_2$ at the times $t_1, t_2, \cdots, t_n$; in the case of two cameras synchronized, it also contains the components of the two vectors $O_AR_A$ and $O_AV_A$.

- $b$ is a vector with dimension $3n$ in the case of a single camera and $6n$ in the case of two cameras synchronized.

In the next two subsections we provide the expressions of the vector $x$, the matrix $\Xi$ and the vector $b$, in the case of a single camera and two synchronized cameras, respectively. Then, in section 4.2 we provide all the analytic steps to obtain these expressions and the linear system in (7).
4.1.1 Single camera

The vector $x$, i.e., the vector that includes all the unknowns is:

$$x \equiv [R_A^T, V_A^T, O_{A1}, O_{A21}, O_{A31}, O_{A12}, O_{A22}, O_{A13}, O_{A23}, O_{A33}, \lambda_1, \ldots, \lambda_n]^T \tag{8}$$

where $\lambda_1, \ldots, \lambda_n$, are the distances between $B_1$ and $B_2$ at the times $t_1, \ldots, t_n$. Hence, $x$ contains the components of $R_A$, the components of $V_A$, all the entries of the matrix $O_A$ and all the aforementioned distances. Since $O_A$ is orthonormal, by including all its entries in $x$ we are ignoring six quadratic constraints. The matrix $\Xi$ is:

$$\Xi \equiv \begin{bmatrix}
I_3 & 0_{33} & 0_{33} & 0_{33} & -\mu_1 & 0_3 & \cdots & \cdots & \cdots & \cdots & 0_3 \\
I_3 & \Delta_3 I_3 & \beta_{x3} I_3 & \beta_{y3} I_3 & \beta_{z3} I_3 & 0_3 & -\mu_2 & 0_3 & \cdots & \cdots & 0_3 \\
I_3 & \Delta_3 I_3 & \beta_{x2} I_3 & \beta_{y2} I_3 & \beta_{z2} I_3 & 0_3 & 0_3 & -\mu_3 & 0_3 & \cdots & 0_3 \\
\vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
I_3 & \Delta_n I_3 & \beta_{x1} I_3 & \beta_{y1} I_3 & \beta_{z1} I_3 & 0_3 & \cdots & 0_3 & -\mu_j & 0_3 & \cdots & 0_3 \\
\vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
I_3 & \Delta_n I_3 & \beta_{xn} I_3 & \beta_{yn} I_3 & \beta_{zn} I_3 & 0_3 & \cdots & \cdots & 0_3 & -\mu_{n-1} & 0_3 & \cdots & 0_3 \\
0_3 & 0_3 & 0_3 & -\mu_n & 0_3 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{bmatrix} \tag{9}$$

where:

- $I_3$ is the identity $3 \times 3$ matrix, $0_{33}$ is the $3 \times 3$ zero matrix, $0_3$ the zero $3 \times 1$ vector;
- $\mu_1, \ldots, \mu_n$ are the unit vectors provided by the camera (i.e., the directions of $B_2$ in the frame of $B_1$ at times $t_1, \ldots, t_n$) rotated by pre-multiplying them by the matrix $M^1(t_j)^T$ (this matrix is defined in section 4.2 by equation (15) and can be obtained by only using the gyroscope measurements from the IMU of $B_1$ in the considered time interval);
- $\Delta_j \equiv t_j - t_1 = t_j - t_A$ ($j = 2, \ldots, n$);
- $\beta_{xj}$, $\beta_{yj}$ and $\beta_{zj}$ ($j = 2, \ldots, n$) are defined in section 4.2 and can be obtained by only using the gyroscope and accelerometer measurements from the IMU of $B_2$ in the time interval $(t_A, t_j)$.

The vector $b$ is:

$$b \equiv [\beta_1^1 T, \beta_1^2 T, \ldots, \beta_1^n T, \beta_2^1 T, \ldots, \beta_2^n T]^T \tag{10}$$

where $\beta_j^1$ ($j = 2, \ldots, n$) are defined in section 4.2 and can be obtained by only using the gyroscope and accelerometer measurements from the IMU of $B_1$ in the time interval $(t_A, t_j)$.

4.1.2 Two cameras synchronized

The vector $x$ also includes the components of the two vectors $O_AR_A$ and $O_AV_A$. These vectors are actually dependent on the remaining components of $x$. Since we obtain the solution by inverting the linear system in (7), we are ignoring six further quadratic equations. The vector $x$ is:

$$x \equiv [R_A^T, V_A^T, (O_AR_A)^T, (O_AV_A)^T, O_{A1}, O_{A21}, O_{A31}, O_{A12}, O_{A22}, O_{A13}, O_{A23}, O_{A33}, \lambda_1, \ldots, \lambda_n]^T \tag{11}$$
The matrix $\Xi$ is:

\[
\Xi = \begin{bmatrix}
\nu_1 & \cdots & \nu_n \\
\beta_1 & \cdots & \beta_n \\
\end{bmatrix}
\]

(12)

where:

- $\nu_1, \cdots, \nu_n$ are the unit vectors provided by the second camera (i.e., the directions of $B_2$ in the frame of $B_1$ at times $t_1, \cdots, t_n$) rotated by pre-multiplying them by the matrix $M^2(t_j)^T$ (this matrix is defined in section 4.2 by equation 15) and can be obtained by only using the gyroscope measurements from the IMU of $B_2$ in the time interval $(t_A, t_j)$;

- $[\beta_j^{1\text{up}}] = \begin{bmatrix} \beta_j^1_{x} & \beta_j^1_{y} & \beta_j^1_{z} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $[\beta_j^{1\text{center}}] = \begin{bmatrix} 0 & 0 & 0 \\ \beta_j^1_{x} & \beta_j^1_{y} & \beta_j^1_{z} \\ 0 & 0 & 0 \end{bmatrix}$ and $[\beta_j^{1\text{down}}] = \begin{bmatrix} 0 & 0 & 0 \\ \beta_j^1_{x} & \beta_j^1_{y} & \beta_j^1_{z} \\ 0 & 0 & 0 \end{bmatrix}$.

The vector $b$ is:

\[
b \equiv [\beta_1^1 T, \beta_2^1 T, \beta_3^1 T, \beta_4^1 T, \cdots, \beta_n^1 T, \beta_2^2 T, \cdots, \beta_n^2 T, \beta_n^3 T]^T
\]

(13)

We conclude this section by remarking that all the components of the matrix $\Xi$ and the vector $b$ depend only on the measurements from the IMUs and the camera (or the two cameras) delivered during the time interval $(t_A, t_B)$. As a result, the solution is able to obtain the entire observable state without any prior knowledge (e.g., initialization). In particular, it provides the state as a simple expression of the measurements delivered during the time interval $(t_A, t_B)$. In addition, since the time interval is very short (2, 3 seconds), the solution is also drift-free.

Note that we include in $x$ all the entries of the matrix $O_A$ and, in the case of two synchronized cameras, we also include the components of the two vectors $O_A R_A$ and $O_A V_A$. This means that by obtaining the observable state through the inversion of the linear system in (7), we are considering independent the entries of $O_A$ and, in the case of two cameras synchronized, the entries of $O_A R_A$ and $O_A V_A$ independent from the remaining components of $x$. The fact that the matrix $O_A$ is orthonormal, means that we are ignoring 6 quadratic equations, i.e., the equations that express the fact that each column of the matrix is a unit vector (3 equations) and that the three columns are orthogonal one each other (3 equations). In the case of two cameras synchronized, including the components of the two vectors $O_A R_A$ and $O_A V_A$ in $x$, means that we are ignoring six further quadratic equations. All these quadratic equations could be exploited in a second step to improve the precision (e.g., by minimizing the cost function defined as the norm of the residual of both the linear system in (7) and the system of the aforementioned quadratic equations).
4.2 Analytic Derivation

We start our derivation by introducing a new local frame for each rigid body (i.e., one new frame for \( B_1 \) and one new frame for \( B_2 \)). Each new frame is defined as follows. It shares the same origin with the original local frame. Additionally, it does not rotate and its orientation coincides with the one of the original frame at the time \( t_A \). From now on, we will refer to this frame as to the new frame. Additionally, we will refer to the original local frame, namely the one defined at the beginning of section 2, as to the original frame. Figure 2 illustrates the original and the new frame of \( B_i \) (\( i = 1, 2 \)). Specifically, in this figure, the considered rigid body accomplishes a translation and rotation between time \( t_A \) and \( t \). The original frame is in black. The new frame is in red dashed line. The two frames coincide at time \( t_A \).

![Figure 2: Original and new local frame of \( B_i \) (\( i = 1, 2 \)). The original frame is attached to the rigid body and rotates with it. At time \( t_A \) the two frames coincide. The new frame does not rotate and its origin coincides with the origin of the original frame at any time.](image)

Let us introduce the following notation:

- \( \xi \) is the position of \( B_2 \) in the new local frame of \( B_1 \);
- \( \eta \) is the relative velocity of \( B_2 \) with respect to \( B_1 \), expressed in the new local frame of \( B_1 \);
Additionally, we have:

- $M^1(t)$ is the orthonormal matrix that characterizes the rotation made by $\mathcal{B}_1$ between $t_A$ and $t \in (t_A, t_B)$; in other words, it describes the difference in orientation between the original and the new frame of $\mathcal{B}_1$ at a given time $t \in (t_A, t_B)$;

- $M^2(t)$ is defined as $M^1(t)$, but for $\mathcal{B}_2$.

By construction we have:

$$
\xi_A \equiv \xi(t_A) = R_A \quad \eta_A \equiv \eta(t_A) = V_A
$$

Additionally, $M^1(t)$ and $M^2(t)$ can be computed by integrating the following first order differential equations:

$$
\dot{M}_1 = [\Omega^1]_x M_1 \quad \dot{M}_2 = [\Omega^2]_x M_2
$$

with initial conditions: $M_1(t_A) = M_2(t_A) = I_3$, $(I_3$ is the $3 \times 3$ identity matrix) and $[\Omega^1]_x$ and $[\Omega^2]_x$ are the matrices defined in (13). Note that, since $t_B - t_A$ does not exceed 3 seconds, these two matrices can be obtained with very high accuracy by using the measurements from the gyroscopes delivered in the considered time interval. In particular, the drift due to the noise in the gyroscope measurements, is negligible. Regarding the bias, we will show that it can be removed (BIAS).

From (14) we have the following dynamics in the new coordinates:

$$
\begin{bmatrix}
\dot{\xi} = \eta \\
\dot{\eta} = O_A \mathcal{A}_2 - \mathcal{A}_1 \\
\dot{O}_A = 0
\end{bmatrix}
$$

where:

- $\mathcal{A}_1$ is the acceleration (gravitational and inertial) of $\mathcal{B}_1$ expressed in the first new local frame (i.e., $\mathcal{A}_1 = M^1 \mathcal{A}$);

- similarly, $\mathcal{A}_2 = M^2 \mathcal{A}^2$.

Let us introduce the following notation:

- $w^1, w^2$ and $w^3$ are the three columns of the matrix $O_A$, i.e., $O_A = \begin{bmatrix} w^1 & w^2 & w^3 \end{bmatrix}$;

- $\alpha^1(t) = [\alpha^1_x(t), \alpha^1_y(t), \alpha^1_z(t)]^T = \int_{t_A}^t \mathcal{A}_1(\tau)d\tau$;  

- $\beta^1(t) = [\beta^1_x(t), \beta^1_y(t), \beta^1_z(t)]^T = \int_{t_A}^t \alpha^1(\tau)d\tau$;

- $\alpha^2(t) = [\alpha^2_x(t), \alpha^2_y(t), \alpha^2_z(t)]^T = \int_{t_A}^t \mathcal{A}_2(\tau)d\tau$;

- $\beta^2(t) = [\beta^2_x(t), \beta^2_y(t), \beta^2_z(t)]^T = \int_{t_A}^t \alpha^2(\tau)d\tau$;

Note that all the quantities $\alpha^1(t)$, $\alpha^2(t)$, $\beta^1(t)$ and $\beta^2(t)$ are directly provided by the IMU measurements delivered in the interval $(t_A, t)$.

Let us integrate the second equation in (16) between $t_A$ and a given $t \in [t_A, t_B]$. We obtain:

$$
\eta(t) = \eta_A + w^1 \alpha^2_x(t) + w^2 \alpha^2_y(t) + w^3 \alpha^2_z(t) - \alpha^1(t)
$$

(17)
and by substituting in the first equation in (16) and integrating again, we obtain:

\[ \xi(t) = \xi_A + \eta_A(t - t_A) + w^1 \beta_2^1(t) + w^2 \beta_2^2(t) + w^3 \beta_2^3(t) - \beta_1^1(t) \]  

(18)

Note that this equation provides \( \xi(t) \) as a linear expression of 15 unknowns, which are the components of the 5 vectors: \( \xi_A, \eta_A, w^1, w^2 \) and \( w^3 \). In the sequel, we build a linear system in these unknowns together with the unknown distances when the cameras perform the measurements.

4.2.1 Single camera

The camera on \( B_1 \) provides the vector \( R(t) = M^1(t)\xi(t) \), up to a scale. We denote by \( \lambda(t) \) this scale (this is the distance between \( B_1 \) and \( B_2 \) at the time \( t \)). We have \( \xi(t) = \lambda(t)\mu(t) \), where \( \mu(t) \) is the unit vector with the same direction of \( \xi(t) \). Note that our sensors (specifically, the camera together with the gyroscope on \( B_1 \)) provide precisely the unit vector \( \mu(t) \): the camera provides the unit vector along \( R(t) \); then, to obtain \( \mu(t) \) it suffices to pre multiply this unit vector by \([M^1(t)]^T\).

We remind the reader that the camera performs \( n \) observations at the times \( t_j \), \( (j = 1, \cdots, n) \), with \( t_1 = t_A \) and \( t_n = t_B \). For notation brevity, for a given time dependent quantity (e.g., \( \lambda(t) \)), we will denote its value a the time \( t_j \) by the pedix \( j \) (e.g., \( \lambda_j = \lambda(t_j) \)). In this notation, equation (18) becomes:

\[ \lambda_j \mu_j = \xi_A + \eta_A(t_j - t_A) + w^1 \beta_2^1_j + w^2 \beta_2^2_j + w^3 \beta_2^3_j - \beta_1^1 \]  

(19)

This is a linear equation in \( 15 + n \) unknowns. The unknowns are:

- The distances \( \lambda_1, \cdots, \lambda_n \).
- The three components of \( \xi_A \).
- The three components of \( \eta_A \).
- The components of the vectors \( w_1, w_2 \) and \( w_3 \), i.e., the nine entries of the matrix \( O_A \).

Note that equation (19) is a vector equations, providing 3 scalar equations. Since this holds for each \( j = 1, \cdots, n \), we obtain a linear system of \( 3n \) equations in \( 15 + n \) unknowns. This is precisely the linear system given in (7) with the vector \( x \) given in (8), the matrix \( A \) given in (9) and the vector \( b \) given in (10).

4.2.2 Two cameras synchronized

Let us consider now the case when also \( B_2 \) is equipped with a camera and, the measurements made by this camera occur at the same times \( t_1, \cdots, t_n \) (synchronized cameras). We define \( \nu_j \) the unit vector such that

\[ \lambda_j \nu_j = -O_A^T \xi_j \]

Note that \(-O_A^T \xi_j\) is the position of \( B_1 \) in the new local frame of \( B_2 \). The second camera provides this vector up to a scale and rotated by the rotation defined by the matrix \( M^2(t_j) \). Since this matrix is known, we conclude that the camera and the IMU measurements on \( B_2 \) provide \( \nu_1, \cdots, \nu_n \).

Let us pre multiply both members of (18) by \(-O_A^T\). We obtain, for \( t = t_j \):

\[ \lambda_j \nu_j = \xi_A' + \eta_A'(t_j - t_A) - \beta_2' + v_1^T \beta_2^1 + v_2^T \beta_2^2 + v_3^T \beta_2^3 \]  

(20)

where:
\( v_1, v_2 \) and \( v_3 \) are the three lines of the matrix \( O_A \);

\( \xi'_A \equiv -O_T^T \xi_A; \)

\( \eta'_A \equiv -O_T^T \eta_A \)

Equations (19) and (20) provide a linear system of \( 6n \) equations in \( 21 + n \) unknowns. With respect to the case of a single camera, we have 6 new unknowns (i.e., the components of the two vectors \( \xi'_A \) and \( \eta'_A \)) but we gain \( 3n \) further equations. The equations in (19) for \( j = 1, \cdots, n \) are precisely the linear system in (7) with the vector \( x \) given in (11), the matrix \( A \) given in (12) and the vector \( b \) given in (13).

5 Limitations of the Closed-Form Solution

The goal of this section is to find out the limitations of the solution provided in section 4 when it is adopted in a real scenario. In particular, special attention will be devoted to the case of an agent equipped with low-cost camera and IMU sensors. For this reason, this section evaluates the impact of the bias on the performance.

5.1 Simulation setup

We simulate two agents as point particles executing random trajectories. The trajectories are simulated as follows. Each trial lasts 4 s. The first agent starts at the origin. The second agent starts at the position \((1.0, 1.0, 1.0) m\). The initial velocities are set equal to \((0.1, 0.1, 0.1) ms^{-1}\) and \((0.2, 0.8, 0.1) m s^{-1}\), for the first and the second agent, respectively. Finally, the initial orientations are characterized by the following values of the roll, pitch and yaw angles, for the two agents: \((0.2\pi, -0.3\pi, 0.8\pi) rad\) and \((0.2\pi, 0.3\pi, -0.8\pi) rad\). For each trial the agents move randomly. The angular speeds, i.e. \( \Omega^1 \) and \( \Omega^2 \), are Gaussian. Specifically, their values at each step of 0.1 s follow a zero-mean Gaussian distribution with covariance matrix equal to \((30 deg)^2 I_3\), where \( I_3 \) is the identity \( 3 \times 3 \) matrix. At each time step, the two agent inertial accelerations are generated as random vectors with zero-mean Gaussian distribution. In particular, the covariance matrix of this distribution is set equal to \((1 ms^{-2})^2 I_3\).

The agents are equipped with inertial sensors able to measure at the frequency of \(0.5 kHz\) the acceleration (the sum of the gravity and the inertial acceleration) and the angular speed. These measurements are affected by errors. Specifically, each measurement is generated by adding to the true value a random error that follows a Gaussian distribution. The mean value of this error is zero and the standard deviation is \(0.03 ms^{-2}\) for the accelerometer and \(0.1 deg s^{-1}\) for the gyroscope. Regarding the camera measurements, they are generated at a lower frequency. Specifically, the measurements are generated at 5Hz. Also these measurements are affected by errors. Specifically, each measurement is generated by adding to the true value a random error that follows a zero mean Gaussian distribution, with variance \(1 deg^2\). All the inertial measurements are corrupted by a bias (see sections 5.3 and 5.4).

We measure our error on the absolute scale by computing the mean error over all estimated distances between the agents at times \( t_1, \cdots, t_n \). In our case, \( n = \frac{4 \times 5}{0.2} = 20 \). We define the relative error as the euclidean distance between the estimation and the ground truth, normalized by the ground truth. For the speed, the error is computed by averaging on the three components of the relative speed. Finally, for the relative orientation, the error is computed by averaging on the roll pitch and yaw, that define the relative rotation between the two local frames.
In the next subsections, we will present the results obtained with the closed-form solution provided in section 4 on the simulated data mentioned, with different sensor bias settings. Our goal is to identify its performance limitations and introduce modifications to overcome them.

5.2 Performance without bias

In Fig. 3, we display the performance of the Closed-Form solution in estimating absolute scale, relative speed and relative orientation. All the values are averaged on 1000 trials. Note how the evaluations get better as we increase the integration time. Note that a robust estimation of the relative orientation requires a shorter duration of integration than the speed and the absolute scale.

![Figure 3: Closed-form solution estimations in absence of bias on the inertial measurements. In blue the relative error on the absolute scale, in red on the relative speed and in black on the relative orientation. The two agents observes one each other over a variable duration of integration.](image)

5.3 Impact of accelerometer bias on the performance

In order to visualize the impact of the accelerometer bias on the performance, we corrupt the accelerometer measurements by a bias (Fig. 4). Despite a high accelerometer bias, the closed-form solution still provides good results. As in Fig 3 all the values are averaged on 1000 trials. Note that, even in the case of a bias with magnitude $0.1 ms^{-1}$ (yellow line in Fig. 4), the error
attains its minimum after 1.5s and it is less than 3% for the scale and less than 10% for the relative speed (note that, the larger error on the speed is due to its smaller absolute value, which is $\approx 0.8ms^{-1}$, meaning that a 10% error corresponds to an error of 0.08$m/s$).

![Graph showing impact of accelerometer bias on performance](image)

**Figure 4**: Impact of the accelerometer bias on the performance of the closed-form solution. The two agents observe one each other over a variable duration of integration.

### 5.4 Impact of gyroscope bias on the performance

To visualize the impact of the gyroscope bias on the performance, we corrupt the gyroscope measurements by an artificial bias (Fig. 5). As in Fig 3, all the values are averaged on 1000 trials. As seen in Fig. 5, the performance becomes very poor in presence of a bias on the gyroscope and, in practice, the overall method could only be successfully used with a very precise - and expensive - gyroscope.

### 6 Estimating the Gyroscope Bias ($B_{\Omega}^1$ and $B_{\Omega}^2$)

In this section, we propose an optimization approach to estimate the gyroscope bias using the closed-form solution.

Since system in (7) is overconstrained, inverting it is equivalent to finding the vector $x$ that minimizes the residual $|\Xi x - b|^2$. We define the following cost function:

$$Cost(B) = |\Xi x - b|^2$$  \hspace{1cm} (21)
Figure 5: Impact of the gyroscope bias on the performance of the closed-form solution. The two agents observe one each other over a variable duration of integration.
with:

- \( B \) is a vector with six components, which are the components of the bias of the first and the second gyroscope, i.e., \( B = [B_1^1, B_2^2] \).

- \( \Xi \) and \( b \) are computed by removing from the measurements provided by the two gyroscopes, the corresponding components of \( B \).

By minimizing this cost function, we recover the gyroscope bias \( B \) and the vector \( x \). Since our cost function requires an initialization and is non-convex, the optimization process can be stuck in local minima. However, by running extensive simulations we found that the cost function is convex around the true value of the bias. Hence, we can initialize the optimization process with \( B = 0 \) since the bias is usually rather small.

7 Performance Overall Evaluation

This section analyzes the performance of the closed form solution completed with the bias estimator introduced in section 6. The setup is the one described in section 5.4. Also in this case, the results are averaged on 1000 trials. We consider the same five values of the bias of the gyroscopes considered in Fig. 5. Finally, we set the magnitude of accelerometer bias equal to zero (Fig. 6) and equal to \( 0.1 \text{ms}^{-2} \) (Fig. 7). Fig. 6 shows a performance comparable to the one exhibited in Fig. 3. On the other hand, Fig. 7 shows a performance better than the one exhibited in Fig. 4. This demonstrates that the effect of the bias has fully overcome.

8 Conclusion

In this paper we extended the results presented in [13,18] to the cooperative case. Specifically, the case of two agents was investigated. Each agent was equipped with inertial sensors (accelerometer and gyroscope) and with a monocular camera. By using the monocular camera, each agent can observe the other agent. No additional camera observations (e.g., of external point features in the environment) were considered. All the inertial sensors were assumed to be affected by a bias.

First, the entire observable state was analytically derived. To this regard, we proved that the entire observable state consists of the following independent physical quantities:

- The position of one of the agents in the local frame of the other agent (this means that the absolute scale is observable).

- The relative speed between the two agents expressed in the local frame of one of them.

- The three Euler angles that characterize the rotation between the two local frames attached to the two agents.

- All the bias that affect the inertial measurements (both the accelerometers and the gyroscopes).

Note that this result holds even in the case when only one of the two agents is equipped with a camera and, very surprisingly, even when this camera is a linear camera, i.e., it only provides the azimuth of the other agent in its local frame.

Then, the paper provided a deterministic solution, able to provide the observable state by only using visual and inertial measurements delivered in a short time interval (2, 3 seconds).
Figure 6: Impact of the gyroscope bias on the performance of the closed-form solution completed with the bias estimator. The accelerometers are unbiased. The two agents observe one each other over a variable duration of integration.
Figure 7: As in Fig. [6] but the magnitude of the accelerometer bias is set to $0.1 m/s^{-2}$ for both the agents.
This solution extended the solution in [15, 19] to the cooperative case. Note that it is possible to retrieve the absolute scale even when no point features are available in the environment.

Following the analysis conducted in [13], the paper focused on investigating all the limitations that characterize this solution when used in a real scenario. Specifically, the impact of the presence of the bias on the performance of this closed-form solution was investigated. As in the case of a single agent, this performance is significantly sensitive to the presence of a bias on the gyroscope, while, the presence of a bias on the accelerometer is negligible. A simple and effective method to obtain the gyroscope bias was proposed. Extensive simulations clearly showed that the proposed method is successful. Note that, it is possible to automatically retrieve the absolute scale and simultaneously calibrate the gyroscopes not only without any prior knowledge (as in [13] for a single agent), but also without external point features in the environment.

Future works will be focused on the analysis of all the singularities that characterize this solution when used in a real scenario. Specifically, the impact of the presence of the bias on the performance of this closed-form solution was investigated. As in the previous case of a single agent, this performance is significantly sensitive to the presence of a bias on the gyroscope, while, the presence of a bias on the accelerometer is negligible. A simple and effective method to automatically retrieve the absolute scale even when no point features are available in the environment.

A Observability Analysis

The goal of this appendix is to detect all the observable quantities (or observable modes) of the system defined in section 2. These scalar quantities will be included in a single vector that will be named the observable state.

We adopt the following notations:

- \( r^1 = [x^1, y^1, z^1] \) and \( r^2 = [x^2, y^2, z^2] \) are the positions of \( B_1 \) and \( B_2 \) in the global frame;
- \( v^1 = [v^1_x, v^1_y, v^1_z] \) and \( v^2 = [v^2_x, v^2_y, v^2_z] \) are the velocities of \( B_1 \) and \( B_2 \) in the global frame;
- \( q^1 = q^1_i + q^1_j + q^1_k \) and \( q^2 = q^2_i + q^2_j + q^2_k \) are the two unit quaternions that describe the rotations between the global and the two local frames, respectively.

In the sequel, for each vector defined in the 3D space, the subscript \( q \) will be adopted to denote the corresponding imaginary quaternion. For instance, regarding the position of \( B_1 \), we have:

\( r^1_q = 0 + x^1 i + y^1 j + z^1 k \)

The dynamics of the first/second rigid body are:

\[
\begin{align*}
\dot{r}^{1/2}_q &= v^{1/2}_q \\
\dot{v}^{1/2}_q &= \frac{q^{1/2}}{2} A^{1/2}_q (q^{1/2})^\ast - gk \\
\dot{q}^{1/2}_q &= \frac{1}{2} q^{1/2} \Omega^{1/2}_q
\end{align*}
\]

(22)

where \( g \) is the magnitude of the gravity and \( k \) is the fourth fundamental quaternion unit (\( k = 0 + 0 i + 0 j + 1 k \)).

The monocular camera on \( B_1 \) provides the position of \( B_2 \) in the local frame of \( B_1 \), up to a scale. The position of \( B_2 \) in the local frame of \( B_1 \) is given by the three components of the following imaginary quaternion:

\[
p^1_q = (q^1)^\ast(r^2_q - r^1_q)q^1
\]

(23)

A quaternion \( q = qt + q_2 i + q_3 j + q_4 k \) is a unit quaternion if the product with its conjugate is 1, i.e.: \( qq^\ast = q^\ast q = (qt + q_2 i + q_3 j + q_4 k)(qt - q_2 i - q_3 j - q_4 k) = (qt)^2 + (q_2)^2 + (q_3)^2 + (q_4)^2 = 1 \)
Hence, the first camera provides the ratios of its components:

$$h^1 \triangleq [h^1_u, h^1_v]^T = \left[ \frac{[p]^1_x}{[p]^1_z}, \frac{[p]^1_y}{[p]^1_z} \right]^T$$  \hspace{1cm} (24)

where the subscripts $x$, $y$ and $z$ indicate respectively the $i$, $j$ and $k$ component of the corresponding quaternion. Similarly, the second camera provides:

$$h^2 \triangleq [h^2_u, h^2_v]^T = \left[ \frac{[p]^2_x}{[p]^2_z}, \frac{[p]^2_y}{[p]^2_z} \right]^T$$ \hspace{1cm} (25)

where $p^2_0$ is the imaginary quaternion whose three components are the position of $B_1$ in the local frame of $B_2$, namely:

$$p^2_0 = (q^2)^*(-r^1_q - r^2_q)q^2$$ \hspace{1cm} (26)

We want to obtain the entire observable state. First of all, we characterize our system by the following state:

$$X = [v^1 \ v^2 \ q^1 \ r^2 \ v^2 \ q^2]^T$$ \hspace{1cm} (27)

The dimension of this state is equal to 20. Actually, the components of this state are not independent. Both $q^1$ and $q^2$ are unit quaternions. In other words, we have: $(q^1_t)^2 + (q^1_i)^2 + (q^1_j)^2 + (q^1_k)^2 = (q^2_t)^2 + (q^2_i)^2 + (q^2_j)^2 + (q^2_k)^2 = 1$.

The dynamics of the state defined in (27) are given by (22). The observation functions are the four scalar functions $h^1_u$, $h^1_v$, $h^2_u$, $h^2_v$ given by equations (23-26). Additionally, we need to add the two observation functions, which express the constraint that the two quaternions, $q^1$ and $q^2$, are unit quaternions. The two additional observations are:

$$h_{const}^{1/2}(X) \triangleq (q^1_t)^{1/2} + (q^1_i)^{1/2} + (q^1_j)^{1/2} + (q^1_k)^{1/2} + (q^2_t)^{1/2} + (q^2_i)^{1/2} + (q^2_j)^{1/2} + (q^2_k)^{1/2}$$ \hspace{1cm} (28)

We investigate the observability properties of this system. Since both the dynamics and the six observations are nonlinear with respect to the state, we use the observability rank condition in [5]. The dynamics are affine in the inputs, i.e., they have the expression

$$\dot{X} = f_0(X) + \sum_{i=1}^{12} f_i(X)u_i$$ \hspace{1cm} (29)

where $u_i$ are the system inputs, which are the quantities measured by the two IMUs. Specifically, we set:

- $u_1, u_2, u_3$ the three components of $A^1$;
- $u_4, u_5, u_6$ the three components of $\Omega^1$;
- $u_7, u_8, u_9$ the three components of $A^2$;
- $u_{10}, u_{11}, u_{12}$ the three components of $\Omega^2$;

Then, by comparing (22) with (25) it is immediate to obtain the analytic expression of all the vector fields $f_0, f_1, \cdots, f_{12}$; for instance, we have:

$$f_0 = [v^-_y, v^-_z, 0, -g, 0, 0, 0, v^+_z, v^+_y, v^+_x, 0, 0, -g, 0, 0, 0, 0]^T$$
For systems with the dynamics given in (33) the application of the observability rank condition can be automatically done by a recursive algorithm. In particular, this algorithm automatically returns the observable codistribution by computing the Lie derivatives of all the system outputs along all the vector fields that characterize the dynamics (see chapter 1 of [19]). For the specific case, we obtain that the algorithm converges at the third step, i.e., the observable codistribution is the span of the differentials of the previous Lie derivatives up to the second order. In particular, its dimension is 11 and, a choice of eleven Lie derivatives is: \( L^0 h^1_u, L^0 h^2_u, L^0 h^1_{\text{const}}, L^0 h^2_{\text{const}}, L^1_f h^1_u, L^1_f h^2_u, L^2_f h^1_u, L^2_f h^2_u \). Note that the choice of these eleven independent Lie derivatives is not unique. In particular, it is possible to avoid the Lie derivatives of the functions \( h^2_u \) and \( h^2_{\text{const}} \). Specifically, a possible choice is: \( L^0 h^1_u, L^0 h^1_{\text{const}}, L^0 h^2_{\text{const}}, L^1_f h^1_u, L^1_f h^2_u, L^2_f h^1_u, L^2_f h^2_u, L^2_{f, f'} h^1_u, L^2_{f, f'} h^2_u \). This means that we obtain the same observability properties when only \( B_1 \) (or only \( B_2 \)) is equipped with a camera. In other words, the presence of two cameras does not improve the observability properties with respect to the case of a single camera mounted on one of the two rigid bodies.

Once we have obtained the observable codistribution, the next step is to obtain the observable state. This state has eleven components. Obviously, a possible choice would be the state that contains the previous eleven Lie derivatives. On the other hand, their expression is too complex and it is much more preferable to find an easier state, whose components have a clear physical meaning. By analytically computing the continuous symmetries of our system (i.e., the killing vectors of the previous observable codistribution, [17]), we detect the following independent observable modes:

- The position of \( B_2 \) in the local frame of \( B_1 \) (three observable modes);
- The velocity of \( B_2 \) in the local frame of \( B_1 \) (three observable modes);
- The three Euler angles that characterize the rotation between the two local frames (three observable modes);
- Trivially, the norm of the two quaternions (two observable modes).

Therefore, we can fully characterize our system by a state whose components are the previous observable modes.

### B Fundamental Equations

In accordance with the observability analysis carried out in appendix A, we characterize our system by the following state:

\[
S = [R \ V \ q]^T
\]

where \( q \) is the unit quaternion that describes the relative rotation between \( B_1 \) and \( B_2 \).

The imaginary quaternions associated to \( R \) and \( V \) are:

\[
R_q = (q^1)^* (r^2_q - r^1_q) q^1
\]

\[
V_q = (q^1)^* (v^2_q - v^1_q) q^1
\]

22
and
\[ q = (q^1)^*q^2 \] (33)
The fundamental equations of the cooperative visual-inertial sensor fusion problem are obtained by differentiating the previous three quantities with respect to time and by using (22) in order to express the dynamics in terms of the components of the state in (30) and the components of \( A^1, A^2, \Omega^1, \Omega^2 \). After some analytic computation, we obtain:
\[
\begin{align*}
\dot{R}_q &= \frac{1}{2}(\Omega^1)^*R_q + \frac{1}{2}R_q \Omega^1_q + V_q \\
\dot{V}_q &= \frac{1}{2}(\Omega^1)^*V_q + \frac{1}{2}V_q \Omega^1_q + qA^2_q + A^1_q \\
\dot{q} &= \frac{1}{2}(\Omega^1)^*q + \frac{1}{2}q \Omega^2_q
\end{align*}
\] (34)
As desired, the dynamics of the state is expressed only in terms of the components of the state and the system inputs (the angular speeds and the accelerations of both \( B_1 \) and \( B_2 \)). Finally, the camera observations can be immediately expressed in terms of the state in (30). The first camera provides the vector \( R \) up to a scale. Regarding the second camera, we first need the position of \( B_1 \) in the local frame of \( B_2 \). The components of this position are the components of the following imaginary quaternion: \(-q^* R_q q\). The second camera provides this position up to a scale.

In the last part of this appendix we provide the same equations, without using quaternions. We characterize our system by the two 3D vectors \( R \) and \( V \), as before. Instead of the quaternion \( q \), we use the matrix \( O \) that characterizes the rotation between the two local frames. From (34) it is immediate to obtain the dynamics of this state. They are the equations in (34).

### C Observability with bias

We analytically obtain the observability properties of the system defined by the state in (27), the dynamics in (4) and the three observations in (24) and (28). Since both the dynamics and the observations are nonlinear with respect to the state, we use the observability rank condition in [5]. The dynamics are affine in the inputs, i.e., they have the expression
\[
\dot{X} = f_0(X) + \sum_{i=1}^{12} f_i(X)u_i
\] (35)
where \( u_i \) are the system inputs, which are the quantities measured by the two IMUs. Specifically, we set:
- \( u_1, u_2, u_3 \) the three components of \( \Omega^1 \);
- \( u_4, u_5, u_6 \) the three components of \( A^1 \);
- \( u_7, u_8, u_9 \) the three components of \( \Omega^2 \);
- \( u_{10}, u_{11}, u_{12} \) the three components of \( A^2 \);

Then, by comparing (4) with (35) it is immediate to obtain the analytic expression of all the vector fields \( f_0, f_1, \ldots, f_{12} \); for instance, we have:
\[
f_4 = [0, 0, 0, q_1^2 + q_2^2 - q_3^2 - q_4^2, 2q_1 q_5 + 2q_2 q_6, 2q_3 q_5 - 2q_4 q_6, 0, 0, 0, 0, 0, 0]^T
\]
\[ f_2 = [0, -R_z, R_y, 0, -V_z, V_y, q_z/2, -q_y/2, q_z/2, -q_y/2, 0, 12]^T \]

For systems with the dynamics given in (33), the application of the observability rank condition can be automatically done by a recursive algorithm. In particular, this algorithm automatically returns the observable codistribution by computing the Lie derivatives of all the system outputs along all the vector fields that characterize the dynamics (see chapter 1 of [10]). For the specific case, we obtain that the algorithm converges at the fourth step, i.e., the observable codistribution is the span of the differentials of the previous Lie derivatives up to third order. In particular, its dimension is 22 meaning that all the state components are observable. A choice of 22 Lie derivatives is: \(e^{0}_{h_v}, e^{0}_{h_u}, e^{0}_{h_{const}}, e^{1}_{h_u}, e^{2}_{h_u}, e^{3}_{h_u}, e^{4}_{h_u}, e^{5}_{h_u}, e^{6}_{h_u}, e^{7}_{h_u}, e^{8}_{h_u}, e^{9}_{h_u}, e^{10}_{h_u}, e^{11}_{h_u}, e^{12}_{h_u}, e^{13}_{h_u}, e^{14}_{h_u}, e^{15}_{h_u}, e^{16}_{h_u}, e^{17}_{h_u}, e^{18}_{h_u}, e^{19}_{h_u}, e^{20}_{h_u}, e^{21}_{h_u}, e^{22}_{h_u} \). Note that the choice of these 22 independent Lie derivatives is not unique. In particular, it is possible to avoid the Lie derivatives of the functions \(h_v\). Specifically, in the previous choice, only the last two Lie derivatives are Lie derivatives of the function \(h_v\). It is possible to avoid these two functions. On the other hand, in this case we need to include fourth order Lie derivatives of \(h_u\). For instance, we can replace the last two functions with \(e^{22}_{h_u} \). This means that we obtain the same observability properties when the first agent is equipped with a linear camera able to only provide the azimuth of the second agent in its local frame.

References

[1] L. Armesto, J. Tornero, and M. Vincze, Fast ego-motion estimation with multi-rate fusion of inertial and vision, The International Journal of Robotics Research (IJRR), vol. 26, no. 6, pp. 577–589, 2007.

[2] C. Forster, M. Pizzoli, and D. Scaramuzza, Svo: Fast semi-direct monocular visual odometry, in International Conference on Robotics and Automation (ICRA), 2014.

[3] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation, in Robotics: Science and Systems (RSS), 2015.

[4] H. X. Guo, K. Sartipi, R. C. DuToit, G. A. Georgiou, R. Li, J. Oleary, E. D. Nerurkar, J. A. Hesch, and S. I. Roumeliotis, Large-Scale Cooperative 3D Visual-Inertial Mapping in a Manhattan World, 2016 IEEE International Conference on Robotics and Automation (ICRA), May 2016, Stockholm, Sweden.

[5] Hermann R. and Krener A.J., 1977, Nonlinear Controllability and Observability, Transaction On Automatic Control, AC-22(5): 728–740

[6] J. Hesch, D. Kottas, S. Bowman, and S. Roumeliotis, Consistency analysis and improvement of vision-aided inertial navigation, Transactions on Robotics (T-RO), vol. 30, no. 1, pp. 158–176, 2014.

[7] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, On the complexity and consistency of ukf-based slam, in International Conference on Robotics and Automation (ICRA), 2009, pp. 4401–4408.

[8] G. P. Huang, A. Mourikis, S. Roumeliotis, et al., An observability-constrained sliding window filter for slam, in Intelligent Robots and Systems (IROS), 2011, pp. 65–72.
[9] G. Huang, M. Kaess, and J. J. Leonard, Towards consistent visual-inertial navigation, in International Conference on Robotics and Automation (ICRA), 2015.

[10] Isidori A., Nonlinear Control Systems, 3rd ed., London, Springer Verlag, 1995.

[11] V. Indelman, S. Williams, M. Kaess, and F. Dellaert, Information fusion in navigation systems via factor graph based incremental smoothing, Robotics and Autonomous Systems, pp. 721–738, 2013.

[12] E. Jones and S. Soatto, Visual-inertial navigation, mapping and localization: A scalable real-time causal approach, The International Journal of Robotics Research, vol. 30, no. 4, pp. 407–430, Apr. 2011.

[13] J. Kaiser, A. Martinelli, F. Fontana and D. Scaramuzza, Simultaneous State Initialization and Gyroscope Bias Calibration in Visual Inertial Aided Navigation, IEEE Robotics and Automation Letters (Volume: 2, Issue: 1, Jan. 2017).

[14] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, Keyframe-based visual-inertial odometry using nonlinear optimization, International Journal of Robotics Research (IJRR), 2014.

[15] M. Li and a. I. Mourikis, High-precision, consistent ekf-based visual-inertial odometry, The International Journal of Robotics Research (IJRR), vol. 32, no. 6, pp. 690–711, 2013.

[16] T. Lupton and S. Sukkarieh, Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions, Transactions on Robotics (T-RO), vol. 28, no. 1, pp. 61–76, 2012.

[17] A. Martinelli, State Estimation Based on the Concept of Continuous Symmetry and Observability Analysis: the Case of Calibration, IEEE Transactions on Robotics, vol. 27, no. 2, pp. 239–255, 2011.

[18] A. Martinelli, Vision and imu data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination, Transactions on Robotics (T-RO), vol. 28, no. 1, pp. 44–60, 2012.

[19] A. Martinelli, Closed-form solution of visual-inertial structure from motion, International Journal of Computer Vision (IJCV), vol. 106, no. 2, pp. 138–152, 2014.

[20] A. Martinelli and A. Renzaglia, Cooperative Visual-Inertial Sensor Fusion: Fundamental Equations, MRS 2017.

[21] A. Mourikis, S. Roumeliotis, et al., A multi-state constraint kalman filter for vision-aided inertial navigation, in International Conference on Robotics and Automation (ICRA), 2007, pp. 3565–3572.

[22] A. Mourikis, S. Roumeliotis, et al., A dual-layer estimator architecture for long-term localization, in Computer Vision and Pattern Recognition Workshops (CVPRW), 2008.