Cooperative Visual-Inertial Sensor Fusion: Fundamental Equations

Agostino Martinelli, Alessandro Renzaglia

Cooperative Visual-Inertial Sensor Fusion: Fundamental Equations

Agostino Martinelli[1] and Alessandro Renzaglia[1,2]

***Abstract*— This paper provides a new theoretical and basic result in the framework of cooperative visual-inertial sensor fusion. Specifically, the case of two aerial vehicles is investigated. Each vehicle is equipped with inertial sensors (accelerometer and gyroscope) and with a monocular camera. By using the monocular camera, each vehicle can observe the other vehicle. No additional camera observations (e.g., of external point features in the environment) are considered. First, the entire observable state is analytically derived. This state includes the relative position between the two aerial vehicles (which includes the absolute scale), the relative velocity and the three Euler angles that express the rotation between the two vehicle frames. Then, the basic equations that describe this system are analytically obtained. In other words, both the dynamics of the observable state and all the camera observations are expressed only in terms of the components of the observable state and in terms of the inertial measurements. These are the fundamental equations that fully characterize the problem of fusing visual and inertial data in the cooperative case. The last part of the paper describes the use of these equations to achieve the state estimation through an EKF. In particular, a simple manner to limit communication among the vehicles is discussed. Results obtained through simulations show the performance of the proposed solution, and in particular how it is affected by limiting the communication between the two vehicles.**

I. INTRODUCTION

When a team of mobile robots cooperates to fulfill a mission, an optimal localization strategy must take advantage of relative observations (detection of other robots). This problem has been considered in the past by following different approaches and it is often referred as *Cooperative Localization*. In Cooperative Localization (CL), several communicating robots use relative measurements (such as distance, bearing and orientation between the robots) to jointly estimate their poses. This problem has been investigated for a long time and several approaches have been introduced in earlier works [5], [8], [17], [26], [35], [36], [37]. Then, a great effort has been devoted to decentralize the computation among the team members and, simultaneously, to minimize the communication among the robots without deteriorating the localization performance [2], [12], [18], [19], [20], [23], [24], [38]. Specific cases of cooperative localization have been considered both in 2D and in 3D. In the framework of Micro Aerial Vehicles (MAV), a critical issue is to limit the number of on-board sensors to reduce weight and power consumption. Several methods consider the use of bearing-only sensors [31], [34], [39], [40] or only range measurements [41]. A common setup is to combine a monocular camera with an Inertial Measurements Unit (IMU). On top of being cheap, these sensors have very interesting complementarities. Additionally, they can operate in indoor environments, where Global Positioning System (GPS) signals are shadowed.

The problem of fusing visual and inertial data for single robots has been extensively investigated in the past [1], [3], [9], [15], [22]. Recently, this sensor fusion problem has been successfully addressed by enforcing observability constraints [7], [11], and by using optimization-based approaches [4], [10], [14], [21], [25], [32], [33]. 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 closed-form solution able to automatically determine the state without initialization has been introduced [16], [28], [29].

Any estimation approach, either filter-based or optimization-based, is built upon the fundamental equations that characterize the considered sensor fusion problem. These equations are the differential equations that describe the dynamics of the observable state together with the equations that express the observations in terms of this observable state. Hence, to successfully solve a given estimation problem, the first step to be accomplished is the determination of the observable state. Regarding the single-vehicle visual-inertial sensor fusion problem, this state has been analytically derived by many authors and it consists of the absolute scale, the speed expressed in the local frame and the absolute roll and pitch angles. This result even holds if only a single point feature is available in the environment.

In this paper we investigate the visual-inertial sensor fusion problem in the cooperative case. We assume that the IMU provides bias-free measurements (the case of a bias is addressed in [30]). We investigate the extreme case where no point features are available. Additionally, we consider the critical case of only two aerial vehicles. 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 vehicle is equipped with an Inertial Measurement Unit (IMU) and a monocular camera. By using the monocular camera, each vehicle can observe the other vehicle. Note that, we do not assume that these camera observations contain metric information (due for instance to the known size of the observed vehicle). The two aerial vehicles can operate far from each other and a single camera observation only consists of the bearing of the

This work has been supported by the French National Research Agency ANR 2014 through the project VIMAD

[1] INRIA Rhone Alpes, Grenoble, France
[2] INSA Lyon CITI lab, France
e-mail: name.surname@inria.fr

observed vehicle in the frame of the observer. In other words, each vehicle acts as a moving point feature with respect to the other vehicle.

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 generally, 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 section II we provide a full answer to these fundamental questions.

Then, the basic equations that describe the cooperative visual-inertial sensor fusion problem are obtained (section III). These equations are:

- The differential equations that describe the dynamics of the observable state expressed only in terms of the components of the observable state and the accelerations and the angular speeds (i.e., the quantities measured by the two IMUs);
- The equations that provide the analytic expression of the two camera observations in terms of the components of the observable state.

These are the fundamental equations that fully characterize the problem of fusing visual and inertial data in the cooperative case. These equations can then be used to build any method (e.g., filter-based or optimization-based) to carry out the state estimation. In section IV we use them to introduce an EKF-based estimation method. Its performance is then evaluated through simulations and results are provided in section V. In particular, they show how the solution is affected by limiting the communication between the vehicles. Finally, in section VI we present our conclusion and the challenges that we are currently trying to address.

II. Observable State

A. The system

We consider two aerial vehicles that move in a $3D-$environment. Each vehicle is equipped with an Inertial Measurement Unit (IMU), which consists of three orthogonal accelerometers and three orthogonal gyroscopes. Additionally, each vehicle is equipped with a monocular camera, which is assumed to be calibrated. We assume that, for each vehicle, all the sensors share the same frame. Without loss of generality, we define the vehicle 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 of each vehicle provides the bearing of the other vehicle in its local frame (see figure 1 for an illustration). Additionally, we assume that the $z-$axis of the global frame is aligned with the direction of the gravity. 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 the two vehicles in the global frame;

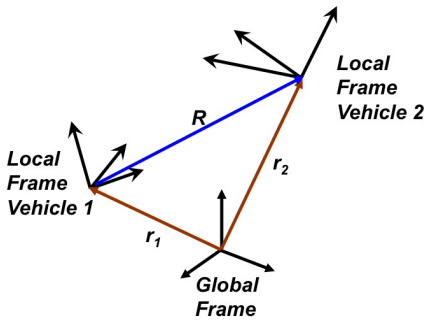

Fig. 1. The global frame and the two local frames (attached to the first and the second aerial vehicle, respectively). r_1 and r_2 are their position, expressed in the global frame. R is the relative position of the second vehicle with respect to the first vehicle, expressed in the local frame of the first vehicle.

- $v^1 = [v_x^1,\ v_y^1,\ v_z^1]$ and $v^2 = [v_x^2,\ v_y^2,\ v_z^2]$ are the velocities of the two vehicles in the global frame;
- $q^1 = q_t^1 + q_x^1 i + q_y^1 j + q_z^1 k$ and $q^2 = q_t^2 + q_x^2 i + q_y^2 j + q_z^2 k$ are the two unit quaternions that describe the rotations between the global and the two local frames, respectively[1].

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 the first vehicle, we have: $r_q^1 = 0 + x^1\ i + y^1\ j + z^1\ k$. Additionally, we denote by A^1, A^2, Ω^1 and Ω^2 the following physical quantities:

- $A^i = [A_x^i,\ A_y^i,\ A_z^i]$, $(i = 1, 2)$, is the vehicle acceleration perceived by the IMU mounted on the first and the second vehicle (this includes both the inertial acceleration and gravity);
- $\Omega^i = \left[\Omega_x^i\ \Omega_y^i\ \Omega_z^i\right]$, $(i = 1, 2)$, is the angular speed of the first and the second vehicle expressed in the respective local frame (and $\Omega_q^i = 0 + \Omega_x^i\ i + \Omega_y^i\ j + \Omega_z^i\ k$).

The dynamics of the first/second vehicle are:

$$\begin{bmatrix} \dot{r}_q^i = v_q^i \\ \dot{v}_q^i = q^i A_q^i (q^i)^* - gk \\ \dot{q}^i = \frac{1}{2} q^i \Omega_q^i \end{bmatrix} \tag{1}$$

where g is the magnitude of the gravity, $i = 1, 2$, and k is the fourth fundamental quaternion unit ($k = 0 + 0\ i + 0\ j + 1\ k$).

The monocular camera on the first vehicle provides the position of the second vehicle in the local frame of the first vehicle, up to a scale. The position of the second vehicle in the local frame of the first vehicle is given by the three components of the following imaginary quaternion:

$$p_q^1 = (q^1)^* (r_q^2 - r_q^1) q^1 \tag{2}$$

[1]A quaternion $q = q_t + q_x i + q_y j + q_z k$ is a unit quaternion if the product with its conjugate is 1, i.e.: $qq^* = q^* q = (q_t + q_x i + q_y j + q_z k)(q_t - q_x i - q_y j - q_z k) = (q_t)^2 + (q_x)^2 + (q_y)^2 + (q_z)^2 = 1$

Hence, the first camera provides the quaternion p_q^1 up to a scale. For the observability analysis, it is convenient to use the ratios of its components:

$$h^1 \triangleq [h_u^1,\ h_v^1]^T = \left[\frac{[p_q^1]_x}{[p_q^1]_z},\ \frac{[p_q^1]_y}{[p_q^1]_z}\right]^T \tag{3}$$

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

$$h^2 \triangleq [h_u^2,\ h_v^2]^T = \left[\frac{[p_q^2]_x}{[p_q^2]_z},\ \frac{[p_q^2]_y}{[p_q^2]_z}\right]^T \tag{4}$$

where p_q^2 is the imaginary quaternion whose three components are the position of the first vehicle in the local frame of the second, namely:

$$p_q^2 = (q^2)^*(r_q^1 - r_q^2)q^2 \tag{5}$$

Note that, using the ratios in (3) and (4) as observations, can provide problems due to singularities and, when the camera measurements are used to estimate a state, it is more preferable to adopt different quantities (e.g., the two bearing angles, i.e., the azimuth and the zenith). For the observability analysis, this problem does not arise.

B. Observability Analysis

The goal of this subsection is to obtain the entire observable state for the system defined in section II-A. First of all, we characterize this system by the following state:

$$X = [r^1\ v^1\ q^1\ r^2\ v^2\ q^2]^T \tag{6}$$

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: $(q_t^1)^2 + (q_x^1)^2 + (q_y^1)^2 + (q_z^1)^2 = (q_t^2)^2 + (q_x^2)^2 + (q_y^2)^2 + (q_z^2)^2 = 1$.

The dynamics of the state defined in (6) are given by (1). The observation functions are the four scalar functions $h_u^1\ h_v^1\ h_u^2\ h_v^2$ given by equations (2-5). 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:

$$h_{const}^i(X) \triangleq (q_t^i)^2 + (q_x^i)^2 + (q_y^i)^2 + (q_z^i)^2 = 1 \tag{7}$$

$i = 1, 2$. 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 [6]. The dynamics are affine in the inputs, i.e., they have the expression

$$\dot{X} = f_0(X) + \sum_{i=1}^{12} f_i(X)u_i \tag{8}$$

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 Ω^1;
- u_7, u_8, u_9 the three components of A^2;
- u_{10}, u_{11}, u_{12} the three components of Ω^2;

Then, by comparing (1) with (8) it is immediate to obtain the analytic expression of all the vector fields $f_0, f_1, \cdots f_{12}$; for instance, we have:

$$f_0 = [v_x^1, v_y^1, v_z^1, 0, 0, -g, 0_4, v_x^2, v_y^2, v_z^2, 0, 0, -g, 0_4]^T$$

$$f_1 = [0_3, (q_t^1)^2 + (q_x^1)^2 - (q_y^1)^2 - (q_z^1)^2, 2(q_t^1 q_z^1 + q_x^1 q_y^1),$$
$$2(q_x^1 q_z^1 - q_t^1 q_y^1), 0_{14}]^T$$

$$f_4 = \frac{1}{2}[0_6, -q_x^1, q_t^1, q_z^1, -q_y^1, 0_{10}]^T$$

where 0_n is the n-line zero vector.

For systems with the dynamics given in (8) the application of the observability rank condition can be automatically done by a recursive algorithm. In particular, this algorithm automatically returns the observable codistribution[2] by computing the Lie derivatives of all the system outputs along all the vector fields that characterize the dynamics. In the sequel, we provide a very simple description of the observability rank condition for systems with the dynamics given in (8), i.e., dynamics nonlinear in the state and affine in the inputs (for a detailed description the reader is addressed to the first chapter of [13]). In accordance with the observability rank condition, the observable codistribution provides all the observability properties. The dimension of this vector space (the observable coditribution) cannot exceed the dimension of the state X. If this dimension is equal to the dimension of the state X, this means that the entire state is observable (actually, weakly locally observable [6]). If this dimension is smaller than the dimension of the state X, the entire state is not observable and it is possible to detect the observable states by computing its Killing vectors in order to obtain the system symmetries [27]. The recursive algorithm that returns the observable codistribution, for systems with the dynamics given in (8), is the following:

Algorithm 1 Observable codistribution Ω

1) $\Omega_0 = span\{\nabla h_u^1, \nabla h_v^1, \nabla h_u^2, \nabla h_v^2, \nabla h_{const}^1, \nabla h_{const}^2\}$;
2) $\Omega_m = \Omega_{m-1} + \mathcal{L}_{f_0}\Omega_{m-1} + \sum_{i=1}^{12} \mathcal{L}_{f_i}\Omega_{m-1}$

We remind the reader that the Lie derivative of a scalar function $h(x)$ along the vector field $f(x)$ is defined as follows:

$$\mathcal{L}_f h \triangleq \frac{\partial h}{\partial x} f$$

[2]The reader non-familiar with the concept of *codistribution*, as it is used in [13], should not be afraid by the term *distribution* and the term *codistribution*. Very simply speaking (and this is enough to understand the theory of nonlinear observability) they are both vector spaces. Specifically, a distribution is the span of a set of column-vector functions. A codistribution is the span of a set of line-vector functions. Hence, both a distribution and a codistribution can be regarded as vector spaces that change by moving on the space of the states (X), namely, vector spaces that depend on X.

which is the product of the row vector $\frac{\partial h}{\partial x}$ with the column vector f. Hence, it is a scalar function. Additionally, by definition of Lie derivative of covectors, we have: $\mathcal{L}_f \nabla h = \nabla \mathcal{L}_f h$. Finally, given two vector spaces V_1 and V_2, we denoted by $V_1 + V_2$ their sum, i.e., the span of all the generators of both V_1 and V_2.

In [13] it is proven that algorithm 1 converges. In particular, it is proven that it has converged when $\Omega_m = \Omega_{m-1}$. From this, it is easy to realize that the convergence is achieved in at most $n-1$ steps[3], where n is the dimension of the state.

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: $\mathcal{L}^0 h^1_u,\ \mathcal{L}^0 h^1_v,\ \mathcal{L}^0 h^2_u,\ \mathcal{L}^0 h^2_v,\ \mathcal{L}^0 h^1_{const},\ \mathcal{L}^0 h^2_{const},\ \mathcal{L}^1_{f_0} h^1_u,\ \mathcal{L}^1_{f_0} h^1_v,\ \mathcal{L}^1_{f_0} h^2_u,\ \mathcal{L}^2_{f_0 f_0} h^1_u,\ \mathcal{L}^2_{f_0 f_1} h^1_u$.

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, [27]), we detect the following independent observable modes:

- The position of the second vehicle in the local frame of the first vehicle (three observable modes);
- The velocity of the second vehicle in the local frame of the first vehicle (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. It must be possible to express the dynamics of this state only in terms of its components and the twelve system inputs. Additionally, also the camera observations must be expressed only in terms of these nine components. This is actually trivial, since the first camera provides the first three components of this state, up to a scale. The second camera, provides the same unit vector rotated according to the previous three Euler angles. Regarding the dynamics, its derivation is a bit more complex. We provide all these analytic expressions in the next section.

We conclude this section with the following remark. The absolute roll and pitch angles of each vehicle are not observable. This is a consequence of the fact that no feature in the environment has been considered. The observation consists only of the bearing angles of each vehicle in the local frame of the other vehicle. The presence of the gravity, which determines the observability of the absolute roll and pitch in the case of a single vehicle, acts in the same way on the two IMU's and its effect on the system observability vanishes.

[3]This is a consequence of lemmas 1.9.1, 1.9.2 and 1.9.6 in [13].

III. Fundamental Equations

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

$$S = [R\ V\ q]^T \tag{9}$$

where:

- R is the position of the second vehicle in the local frame of the first vehicle;
- V is the velocity of the second vehicle in the frame of the first vehicle (note that this velocity is not simply the time derivative of R because of the rotations accomplished by the first local frame);
- q is the unit quaternion that describes the relative rotation between the two local frames.

In other words, the imaginary quaternions associated to R and V are:

$$R_q = (q^1)^*(r^2_q - r^1_q)q^1 \tag{10}$$

$$V_q = (q^1)^*(v^2_q - v^1_q)q^1 \tag{11}$$

and

$$q = (q^1)^* q^2 \tag{12}$$

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 (1) in order to express the dynamics in terms of the components of the state in (9) and the components of $A^1, A^2, \Omega^1, \Omega^2$. After some analytic computation, we obtain:

$$\begin{cases} \dot{R}_q = \frac{1}{2}(\Omega^1_q)^* R_q + \frac{1}{2} R_q \Omega^1_q + V_q \\ \dot{V}_q = \frac{1}{2}(\Omega^1_q)^* V_q + \frac{1}{2} V_q \Omega^1_q + q A^2_q q^* - A^1_q \\ \dot{q} = \frac{1}{2}(\Omega^1_q)^* q + \frac{1}{2} q \Omega^2_q \end{cases} \tag{13}$$

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 the vehicles). Finally, the camera observations can be immediately expressed in terms of the state in (9). The first camera provides the vector R up to a scale. Regarding the second camera, we first need the position of the first vehicle in the second local frame. 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 section 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 (13) it is immediate to obtain the dynamics of this state. They are:

$$\begin{cases} \dot{R} = \left[\Omega^1\right]_\times R + V \\ \dot{V} = \left[\Omega^1\right]_\times V + OA^2 - A^1 \\ \dot{O} = \left[\Omega^1\right]_\times^T O + O\left[\Omega^2\right]_\times \end{cases} \tag{14}$$

where $\left[\Omega^i\right]_\times$, $i = 1, 2$, are the skew-symmetric matrices associated to the vectors Ω^i:

$$\left[\Omega^i\right]_\times = \begin{bmatrix} 0 & \Omega_z^i & -\Omega_y^i \\ -\Omega_z^i & 0 & -\Omega_x^i \\ \Omega_y^i & -\Omega_x^i & 0 \end{bmatrix}$$

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

The cooperative visual-inertial sensor fusion problem is fully characterized by the dynamics equations given in (14) and the two observations given by R and $-O^T R$, up to a scale. These equations allow us to build any estimation strategy (filter-based, optimization-based or a closed-form solution, i.e. a deterministic solution that extends the solution given in [29] to the cooperative case). In the next section, we discuss the possibility of using these equations to build up an EKF. Future works will be devoted to obtain a closed-form solution starting from these equations.

IV. Estimation based on an EKF

The goal of this section is to build an EKF that estimates a state that includes the two 3D vectors R and V, and the three Euler angles (α, β, γ) that characterize the matrix O, introduced in the previous section. Specifically, we set:

$$O = O_x(\alpha) O_z(\beta) O_x(\gamma) \tag{15}$$

where $O_x(\alpha), O_z(\beta), O_x(\gamma)$ are the matrices representing the rotations about the axes x, z and x of α, β and γ, respectively.

The filter estimates the state $[R, V, \alpha, \beta, \gamma]^T$. The dynamics of the system can be computed starting from (14). By discretizing the dynamics, we obtain the Jacobians necessary to build up the prediction phase of the filter.

The two cameras provide the two vectors R and $-O^T(\alpha, \beta, \gamma)R$, up to a scale. We consider, as observations, the two scalar functions obtained by computing the two independent bearing angles of the corresponding vector.

Therefore, the filter has four scalar functions as observation functions. By computing their Jacobian with respect to the state, we have all the ingredients to implement this EKF.

We remark that this EKF presents an important drawback. The two vehicles must communicate to implement both the prediction and the estimation phase. This becomes an important issue because the data provided by the IMU are delivered at a very high frequency (more than $100Hz$). Note that, in cooperative localization, most of the proposed strategies need communication among the team members only at the frequency of the exteroceptive measurements (see for instance [36]). In the sequel, we propose a possible implementation that allows implementing the EKF by limiting the communication between the vehicles, in terms of frequency and amount of data. This solution is based on a suitable approximation, which regards the acceleration of the second vehicle.

Let us consider the case where the EKF runs on the first vehicle and let us highlight the quantities, measured by the second vehicle, needed to implement the prediction phase of the filter. The dynamics of α, β and γ depend on Ω^2 (see the third equation in (14)). The dynamics of V depend on A^2 and O (second equation in (14)). Finally, the dynamics of R depend on V (first equation in (14)). The most critical quantity is O, since it impacts the dynamics of all the remaining quantities and depends on the angular speed of the second vehicle.

Our idea is to eliminate the effect of the rotations by exploiting the fact that, during short time intervals, the IMU provides the angular speed with very high accuracy (provided that it is calibrated with respect to the bias). Let us denote by T the length of a time interval for which the drift on the rotation, obtained by integrating the angular speed provided by an IMU, is negligible. Our method consists in defining new local frames (for both the vehicles) that do not rotate during each time interval of length T. For each time interval, an EKF is implemented in these coordinates. Then, at the end of each time interval, new local frames are re-defined and a new EKF is initialized.

Let us consider a given time interval $(T_0,\ T_0 + T)$. At the initial time (T_0), the two vehicles have given (unknown) orientations with respect to the global frame. The two new local frames maintain these orientations during the entire time interval. Let us denote by ξ the position of the vehicle in the first new local frame and by η its speed. We have:

$$\begin{cases} \dot{\xi} &= \eta \\ \dot{\eta} &= O_0(\alpha_0, \beta_0, \gamma_0)\mathcal{A}^2 - \mathcal{A}^1 \\ \dot{\alpha}_0 = \dot{\beta}_0 = \dot{\gamma}_0 &= 0 \end{cases} \tag{16}$$

where:

- $\alpha_0, \beta_0, \gamma_0$ are the three Euler angles that define the rotation between the first vehicle and the second vehicle at the initial time of the considered time interval (T_0) and $O_0(\alpha_0, \beta_0, \gamma_0)$ is the corresponding rotation matrix;
- $\mathcal{A}^1$ is the acceleration (gravitational and inertial) of the first vehicle expressed in the first new local frame (in other words, it is the acceleration A^1 rotated by integrating Ω^1 from the time T_0 up to the current time);
- $\mathcal{A}^2$ is the acceleration (gravitational and inertial) of the second vehicle expressed in the second new local frame (i.e., the acceleration A^2 rotated by integrating Ω^2 from the time T_0 up to the current time).

Note that, in accordance with the observability analysis carried out in section II-B, the state $[\xi, \eta, \alpha_0, \beta_0, \gamma_0]^T$ is

observable[4].

Now, the implementation of the prediction phase of an EKF that estimates $[\xi, \eta, \alpha_0, \beta_0, \gamma_0]^T$ only requires that the second vehicle communicates the 3D vector $\mathcal{A}^2$. Our approximation consists in communicating this vector not at the frequency of the IMU of the second vehicle but at a slower frequency. Let us denote by τ the elapsed time between two consecutive communications. The second vehicle provides the mean value of $\mathcal{A}^2$, where the mean is computed on the time interval τ. In other terms, the vector provided by the second vehicle is $\bar{\mathcal{A}}^2(t) = \frac{1}{\tau}\int_{t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} \mathcal{A}^2(s)ds$.

In this settings, the communication needed to implement the prediction phase of the filter is limited to three scalars for each time interval τ. Namely, if $\tau = 0.1\ s$, the frequency is $10\ Hz$. Additionally, the communication to implement the estimation phase only consists of two scalars (the two bearing angles provided by the second camera) at the frequency of the camera observations.

V. Simulations

A. Simulated trajectories and sensors

The trajectories are simulated as follows. The equations in (1) are discretized with a time step of $0.01\ s$. Each trial lasts $100\ s$. The initial vehicle speed is set to $[0.1,\ 0.1,\ 0.1] ms^{-1}$, for both vehicles. The initial position of the first vehicle is set equal to $[0,\ 0,\ 0]m$. For each trial, the initial position of the second vehicle and the initial roll, pitch and yaw angles of the two vehicles are randomly initialized. For the specific scenario reported in this section, the second vehicle initial position is $[0.20,\ -1.58,\ -0.68]m$ and the first and second vehicle angles are $[-0.2\pi,\ 0.3\pi, -0.8\pi]$ and $[0.7\pi,\ 0.2\pi, 0.4\pi]$ respectively.

The vehicle trajectories are randomly generated. The angular speeds, i.e. Ω^1 and Ω^2, are Gaussian. Specifically, their values at each step follow a zero-mean Gaussian distribution with covariance matrix equal to $(1\ deg)^2 I_3$, where I_3 is the 3×3 identity matrix. At each time step, the two vehicle speeds are incremented by adding a random vector with zero-mean Gaussian distribution. In particular, the covariance matrix of this distribution is set equal to $\sigma^2 I_3$, with $\sigma = 2 * 10^{-3} m$. Typical trajectories obtained with this setting, are displayed in figure 2.

The vehicles are equipped with inertial sensors able to measure at each time step 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 at every time step of $0.01\ s$ 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 $0.01\ ms^{-2}$ for the accelerometer and $1\ deg\ s^{-1}$ for the gyroscope.

Regarding the camera measurements, they are generated at a lower frequency. Specifically, the measurements are generated each $0.2\ s$. 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 $1\ deg^2$.

[4]This result could also be obtained by computing the Lie derivatives of the system outputs along the vector fields that define the dynamics in (16).

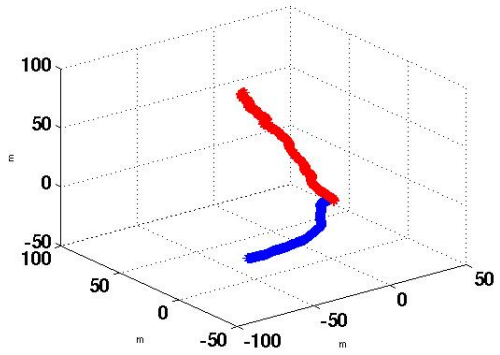

Fig. 2. Typical simulated trajectories. In blue the first vehicle, in red the second vehicle.

B. Estimation results

We implement the EKF described in section IV. The parameter T is set equal to $5\ s$ for all the trials. We find similar results for values of T that do not exceed $20\ s$. Regarding the parameter τ, we consider three distinct values: $\tau = 0.01\ s$, $\tau = 0.1\ s$ and $\tau = 0.2\ s$. In practice, the first case ($\tau = 0.01\ s$) corresponds to an implementation without approximation. The vector $\mathcal{A}^2$ is provided to the first vehicle at the same frequency of the inertial sensor. In the last case ($\tau = 0.2\ s$), the communication occurs at the same frequency of the camera observations.

Figures 3-5 display the estimated relative position (the trajectory of the second vehicle in the local frame of the first vehicle, i.e., the vector R). The true trajectory is in blue. The black line is the trajectory obtained by only using the inertial measurements, while the red line is the trajectory obtained by the proposed EKF. Figure 3 is obtained by setting $\tau = 0.01\ s$, figure 4 by setting $\tau = 0.1\ s$ and figure 5 by setting $\tau = 0.2\ s$.

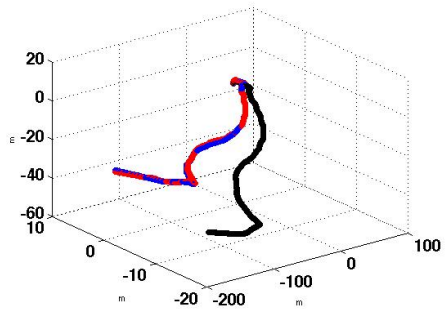

Fig. 3. Estimated trajectory of the second vehicle in the local frame of the first vehicle. In blue the ground truth, in black the trajectory obtained by only using the inertial measurements, in red the trajectory obtained by the proposed EKF. The parameter τ is set equal to $0.01\ s$.

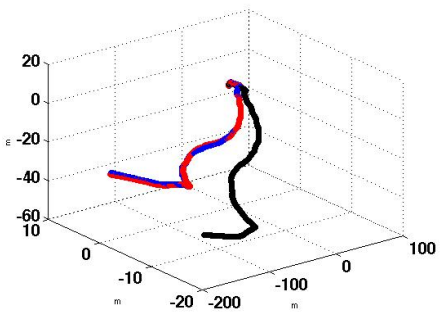

Fig. 4. As in figure 3 but with $\tau = 0.1\ s$.

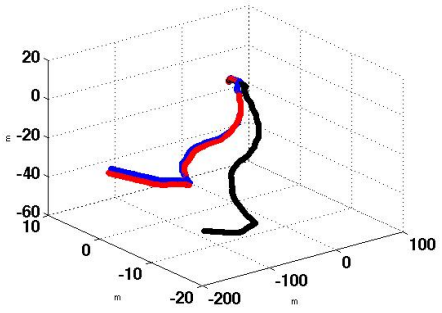

Fig. 5. As in figure 3 but with $\tau = 0.2\ s$.

The results shown in the previous figures regard a single trial. In table I we report the error on the final position averaged on 1000 trials.

VI. Conclusion

In this paper we provided a new theoretical and basic result in the framework of cooperative visual-inertial sensor fusion. We considered two sensor suites that accomplish independent 3D motions (in general, these can be the sensor suites of two independent aerial vehicles). Each sensor suite consists of a monocular camera and an Inertial Measurement Unit. We assumed that, by using the monocular camera, each vehicle can observe the other vehicle. No additional camera observations (e.g., of external point features in the environment) were considered.

The first theoretical result was the analytic derivation of the entire observable state. Specifically, this state consists of the relative position between the two aerial vehicles (which includes the absolute scale), the relative velocity and the three Euler angles that express the rotation between the two vehicle frames. The absolute roll and pitch are not observable (note that this is true only when no point feature is observed in the environment).

τ (in s)	Position Error (in m)
0.01	0.67
0.1	2.0
0.2	4.7

TABLE I

Error on the final value of R averaged on 1000 trials.

The second theoretical result was the derivation of the fundamental equations that describe cooperative visual-inertial sensor fusion. Both the dynamics of the observable state and all the camera observations were expressed exclusively in terms of the components of the observable state and in terms of the inertial measurements. These are the fundamental equations that fully characterize the problem of fusing visual and inertial data in the cooperative case.

The last part of the paper provided a first simple use of these equations to perform the state estimation through an EKF. In particular, a simple manner to limit communication among the vehicles was proposed and discussed. Results obtained through simulations assessed the performance of this approach and the effect of the limitation in the communication between the two vehicles.

The importance of the results provided by this paper is that it is possible to retrieve the absolute scale, even when no feature is available in the environment.

Our current work is devoted to use these equations in order to derive a closed-form solution [30]. This is the extension of the closed-form solution provided in [29] to the cooperative case.

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] L. C. Carrillo-Arce, E. D. Nerurkar, J. L. Gordillo, and S. I. Roumeliotis, Decentralized multi-robot cooperative localization using covariance intersection, in IEEE/RSJ Int. Conf. on Intelligent Robots & Systems, (Tokyo, Japan), pp. 14121417, 2013.

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

[4] 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.

[5] R. Grabowski, L.E. Navarro-Serment, C.J.J. Paredis, P.K. Khosla, 2000, Heterogeneous Teams of Modular Robots for Mapping and Exploration, Autonomous Robots, Vol. 8, no. 3, 293-308, June 2000

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

[7] 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.

[8] Howard A., Mataric M.J. and Sukhatme G.S., "Localization for Mobile Robot Teams Using Maximum Likelihood Estimation", *International Conference on Inteligent Robot and Systems (IROS02)*, Volume: 3 , 30 Sept.-5 Oct. 2002 Pages:2849 - 2854, Lausanne.

[9] 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. 44014408.

[10] 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. 6572.

[11] G. Huang, M. Kaess, and J. J. Leonard, Towards consistent visual-inertial navigation, in International Conference on Robotics and Automation (ICRA), 2015.

[12] Indelman, V., Gurfil, P., Rivlin, E., & Rotstein, H. (2012). Graph-based distributed cooperative navigation for a general multi-robot measurement model. The International Journal of Robotics Research, 31(9), 1057-1080.

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

[14] 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. 721738, 2013.

[15] 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.

[16] 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)

[17] K. Kato, H. Ishiguro, M. Barth, "Identifying and Localizing Robots in a Multi-Robot System Environment" *International Conference on Inteligent Robot and Systems (IROS99)* 1999

[18] S. Kia, S. Rounds, and S. Martnez, Cooperative localization for mobile agents: a recursive decentralized algorithm based on Kalman filter decoupling, IEEE Control Systems Magazine, vol. 36, no. 2, pp. 86101, 2016.

[19] Kim, B., Kaess, M., Fletcher, L., Leonard, J., Bachrach, A., Roy, N., & Teller, S. (2010, May). Multiple relative pose graphs for robust cooperative mapping. In Robotics and Automation (ICRA), 2010 IEEE International Conference on (pp. 3185-3192). IEEE.

[20] K. Y. K. Leung, T. D. Barfoot, and H. H. T. Liu, Decentralized localization of sparsely-communicating robot networks: A centralized-equivalent approach, IEEE Transactions on Robotics, vol. 26, no. 1, pp. 6277, 2010.

[21] 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.

[22] 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. 690711, 2013.

[23] H. Li and F. Nashashibi, Cooperative multi-vehicle localization using split covariance intersection filter, IEEE Intelligent Transportation Systems Magazine, vol. 5, no. 2, pp. 3344, 2013.

[24] L. Luft, T. Schubert, S. I. Roumeliotis, and W. Burgard, Recursive decentralized collaborative localization for sparsely communicating robots, in Proceedings of Robotics: Science and Systems, (AnnArbor, Michigan), June 2016.

[25] 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. 6176, 2012.

[26] Martinelli A., Pont F. and Siegwart R., "Multi-Robot Localization Using Relative Observations" *International Conference on Robotics and Automation*, April 2005, Barcellona, Spain.

[27] 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

[28] 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.

[29] 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.

[30] A. Martinelli, "Closed form solution to cooperative visual inertial sensor fusion" https://sites.google.com/site/agobotix/home/publications

[31] G Michieletto, A Cenedese, A Franchi, Bearing rigidity theory in SE (3), (CDC), 2016.

[32] 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.

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

[34] G. Philippe, I. Rekleitis, and M. Latulippe, "I see you, you see me: Cooperative localization through bearing-only mutually observing robots." Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012.

[35] Rekleitis, I.M., Dudek, G., Milios, E.E., "Multi-robot cooperative localization: a study of trade-offs between efficiency and accuracy " *International Conference on Inteligent Robot and Systems (IROS02)* Lausanne, 2002, Switzerland

[36] S.I. Roumeliotis and G.A. Bekey, 2002, Distributed Multirobot Localization, IEEE Transaction On Robotics And Automation Vol 18, No.5, pp. 781–795, October 2002

[37] J.R. Spletzer and C.J. Taylor, "A Bounded Uncertainty Approach to Multi-Robot Localization" *International Conference on Inteligent Robot and Systems (IROS03)* Las Vegas, USA, 2003

[38] N. Trawny, S. I. Roumeliotis, and G. B. Giannakis, Cooperative multi-robot localization under communication constraints, in IEEE Int. Conf. on Robotics and Automation, (Kobe, Japan), pp. 43944400, May 2009.

[39] R Tron, L Carlone, F Dellaert, K Daniilidis, "Rigid Components Identification and Rigidity Enforcement in Bearing-Only Localization using the Graph Cycle Basis", American Control Conference (ACC), 2015.

[40] D Zelazo, PR Giordano, A Franchi, Bearing-only formation control using an se(2) rigidity theory, (CDC), 2015.

[41] X.S. Zhou, S.I. Roumeliotis, "Robot-to-robot relative pose estimation from range measurements", IEEE Transactions on Robotics, 24(6), 1379–1393, 2008