Cooperative Visual-Inertial Sensor Fusion: the Analytic Solution

Agostino Martinelli, Alexander Oliva, Bernard Mourrain

Cooperative Visual-Inertial Sensor Fusion: the Analytic Solution

Agostino Martinelli[1], Alexander Oliva[1], and Bernard Mourrain[2]

***Abstract*—This paper analyzes the visual-inertial sensor fusion problem in the cooperative case of two agents. The paper proves that, this sensor fusion problem, is equivalent to a simple polynomial equations system that consists of several linear equations and three polynomial equations of second degree. The analytic solution of this polynomial equations system is easily obtained by using an algebraic method. In other words, the paper provides the analytic solution to the visual-inertial sensor fusion problem in the case of two agents. The power of the analytic solution is twofold. From one side, it allows us to determine the relative state between the agents (i.e., relative position, speed and orientation) without the need of an initialization. From another side, it provides fundamental insights into all the theoretical aspects of the problem. This paper mainly focuses on the first issue. However, the analytic solution is also exploited to obtain basic structural properties of the problem that characterize the observability of the absolute scale and the relative orientation. Extensive simulations and real experiments show that the solution is successful in terms of precision and robustness.**

***Index Terms*—Sensor Fusion; Visual-Based Navigation; Multi-Robot Systems; Aerial Systems: Perception and Autonomy.**

I. Introduction

THE problem of fusing visual and inertial data has been extensively investigated in the past (e.g., [1], [2], [3], [4], [5]). In this context, methods to obtain the absolute scale in challenging conditions, have been proposed (e.g., [6], [7], [8]). Recently, this sensor fusion problem has been successfully addressed by enforcing observability constraints [9], [10], and by using optimization-based approaches [11], [12], [13], [14], [15], [16], [17]. 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. Deterministic solutions able to automatically determine the state without initialization have also been introduced [7], [18], and they can overcome this obstacle. Even more importantly, the solution provided in [7] is not simply a closed-form solution of the problem that does not need to be initialized. The analysis in [7] established that the visual-inertial sensor fusion problem is equivalent to a very simple Polynomial Equation System (PES). In particular, this PES consists of a single polynomial equation of second degree and several linear equations. This PES can be easily solved in closed-form and, this solution, is the analytic solution of the visual-inertial sensor fusion problem in the case of a single agent. This analytic solution contains all the structural properties of the problem. In particular, by studying this solution, the author of [7] obtained a detailed analysis of the problem by providing all the system singularities and minimal cases depending on the trajectory, on the number of camera images and on the features layout. The problem can have up to two distinct solutions in its minimal cases.

Visual and inertial sensors have also been used in a cooperative scenario (e.g., for cooperative mapping in [19]). Very recently, the visual-inertial sensor fusion problem, in the cooperative case of two agents, has been studied by focusing on the following two theoretical issues ([20], [21], [22]):

1) Investigating its observability properties in order to obtain the observable state in several conditions;
2) Obtaining a closed-form solution able to express the relative state at a given time in terms of the visual and inertial measurements acquired in a short time interval.

The first issue has fully been answered. The observable state only consists of relative states between the agents, i.e., the relative position, speed and orientation. This result, first obtained in [20], 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 above quantities. Estimating other physical quantities (e.g., the absolute roll and pitch angles of the first agent or of the second agent) is not possible. In [21], [22] it was also proved 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 change the observability properties with respect to the case of a single camera mounted on one of the two agents[1]. Additionally, the observable state remains the same even when the camera is a linear camera, i.e., it only provides the azimuth of the other agent in its local frame. Finally, this result is independent of the presence of a bias in the inertial sensors.

The second issue was dealt with in [21], [22]. In particular, [21], [22] provided a linear system where the unknowns are the components of the relative state and the coefficients (i.e., the matrix and the vector with the constant terms) only depend on the visual and inertial measurements delivered during a given time interval. By simply inverting this linear system, it is immediate to obtain a closed-form solution of our problem. This solution has the considerable advantage that it does not need to be initialized. However, this is not the analytic

[1] A. Martinelli and A. Oliva are with INRIA Rhone Alpes, Grenoble, France. agostino.martinelli@inria.fr.

[2] B. Mourrain is with INRIA Sophia Antipolis Mediterranean, Nice, France. bernard.mourrain@inria.fr.

[1]Obviously, this does not mean that having an additional camera is useless (e.g., it improves the precision on the estimated state, it makes more likely that at least one agent observes the other, etc.)

solution of the problem because the unknowns that satisfy the aforementioned linear system are not independent.

The main goal of this paper is precisely to account for this issue. In section II we define our system. This consists of two agents equipped with an Inertial Measurement Unit (IMU) and a monocular camera. 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. In section III, we provide the main paper contribution. We prove that, by characterizing the relative state with independent elements, the vector of unknowns satisfies a PES instead of a linear system. In particular, we obtain that the cooperative visual-inertial sensor fusion problem, in the case of two agents, is equivalent to a PES that consists of several linear equations and three polynomial equations of second degree. This PES is solved by using the method based on the Macaulay resultant matrices [23] and can have up to eight distinct solutions in the minimal case. This is the analytic solution of our problem. Therefore, it allows us to obtain all the structural properties of the problem (e.g., how the minimal cases, singularities and degeneracies depend on the trajectories and on the number of camera images). In section IV we perform a preliminary analysis of the aforementioned PES by obtaining two basic structural properties of the problem. Note that this analysis allows us to investigate how the observability properties depend on the trajectories. This is fundamental in many applications and, so far, it has been discussed in the only-vision case [24]. Finally, we also obtain that the analytic solution outperforms the closed-form solution provided in [21], [22] in terms of precision and robustness. This is shown in section V through simulations. Finally, in section V we also evaluate the performance of the analytic solution by using real data.

II. The system

We consider two rigid bodies that move in a $3D-$environment. We denote them by $\mathcal{B}_1$ and $\mathcal{B}_2$. Each rigid body is equipped with an Inertial Measurement Unit (IMU), which consists of three orthogonal accelerometers and three orthogonal gyroscopes. Additionally, both $\mathcal{B}_1$ and $\mathcal{B}_2$ are equipped with a monocular camera. We assume that, for each rigid body, all the sensors share the same frame. Without loss of generality, we define the body 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 camera mounted on $\mathcal{B}_1$ (or $\mathcal{B}_2$) provides the bearing of $\mathcal{B}_2$ (or $\mathcal{B}_1$) in its local frame.

We adopt the following notations:

- P is the position of $\mathcal{B}_2$ in the local frame of $\mathcal{B}_1$;
- V is the relative velocity of $\mathcal{B}_2$ with respect to $\mathcal{B}_1$, expressed in the frame of $\mathcal{B}_1$ (note that this velocity is not simply the time derivative of P because of the rotations accomplished by $\mathcal{B}_1$);
- R is the rotation matrix that characterizes the rotation between the two local frames; specifically, for a vector with given coordinates in the local frame of $\mathcal{B}_2$, we obtain its coordinates in the local frame of $\mathcal{B}_1$ by pre multiplying by R.

Additionally, we denote by $A^1,\ A^2,\ \Omega^1$ and Ω^2 the accelerations and the angular speeds perceived by ideal (noiseless and unbiased) IMUs mounted on $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively. Regarding the acceleration, it includes both the inertial acceleration and the gravity.

The first camera provides the vector P, up to a scale. The scale is precisely the distance between $\mathcal{B}_1$ and $\mathcal{B}_2$ at the time of the camera measurement. The second camera provides the vector $-R^TP$, up to a scale. Note that, in the special case when the two cameras are synchronized, the scale coincides.

By using basic results on rigid body dynamics we obtain the time derivative of the previous physical quantities (a complete derivation is available in [21], [22]). They are:

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

where $\left[\Omega^1\right]_\times$ and $\left[\Omega^2\right]_\times$ are the skew-symmetric matrices associated to Ω^1 and Ω^2, respectively.

The cooperative visual-inertial sensor fusion problem (from now on CoVISF) is fully characterized by the dynamics equations given in (1) and the two observation functions given by the two vectors P and $-R^TP$, up to a scale.

III. The analytic solution

This section provides the main paper contribution. We show that the problem described in section II is equivalent to a very simple PES whose analytic solution can be easily obtained.

Let us consider a given time interval $(t_A,\ t_B)$. Let us denote by P_A, V_A and R_A, the values of P, V and R at time t_A. These will be precisely the unknowns of the aforementioned equations system. The fundamental feature of this PES is that all its parameters (i.e., the coefficients of the system, and the constant terms) only depend on the visual and inertial measurements delivered in the time interval $(t_A,\ t_B)$. As a result, by solving this PES, we obtain the analytic expression of P_A, V_A and R_A in terms of the visual and inertial measurements delivered in the time interval $(t_A,\ t_B)$. Note that, for a practical application, it is in general more convenient to compute the state at t_B (instead of t_A). Since all the equations are reversible in time, by using the unknowns P_B, V_B and R_B, i.e., the values of P, V and R at time t_B, we would obtain a comparable PES. On the other hand, we believe that it is easier to follow the analytic derivation to obtain the PES in the unknowns P_A, V_A and R_A. Once obtained this PES, we easily obtain the PES in P_B, V_B and R_B (see section III-C).

We start our derivation by introducing a new local frame for each rigid body (i.e., one new frame for $\mathcal{B}_1$ and one new frame for $\mathcal{B}_2$). 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 II, as to the *original*

frame. Figure 1 displays the original and the new frame of $\mathcal{B}_i$ $(i = 1, 2)$. 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.

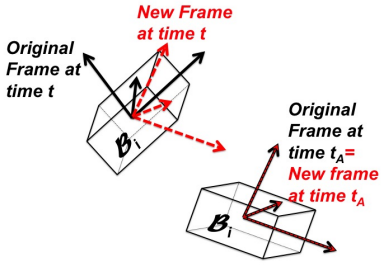

Fig. 1. Original and new local frame of $\mathcal{B}_i$ $(i = 1, 2)$. The original frame is attached to the rigid body and rotates with it. At time t_A the two frames coincide. The new frame does not rotate and its origin coincides with the origin of the original frame at any time.

Let us introduce the following notation:

- ξ is the position of $\mathcal{B}_2$ in the new local frame of $\mathcal{B}_1$;
- η is the relative velocity of $\mathcal{B}_2$ with respect to $\mathcal{B}_1$, expressed in the new local frame of $\mathcal{B}_1$;
- $M^1(t)$ 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:

$$\xi_A \equiv \xi(t_A) = P_A \qquad \eta_A \equiv \eta(t_A) = V_A \tag{2}$$

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

$$\dot{M}_1 = \left[\Omega^1\right]_\times^T M_1 \quad \dot{M}_2 = \left[\Omega^2\right]_\times^T M_2 \tag{3}$$

with initial conditions: $M_1(t_A) = M_2(t_A) = I_3$, where I_3 is the 3×3 identity matrix (note that these two matrices can be easily obtained from the measurements delivered by the gyroscopes in the considered time interval).

From (1) we obtain the following dynamics:

$$\begin{bmatrix} \dot{\xi} &= \eta \\ \dot{\eta} &= R_A\mathcal{A}^2 - \mathcal{A}^1 \\ \dot{R}_A &= 0 \end{bmatrix} \tag{4}$$

where:

- $\mathcal{A}^1$ is the acceleration (gravitational and inertial) of $\mathcal{B}_1$ expressed in the first new local frame (i.e., $\mathcal{A}^1 = M^1A^1$);
- similarly, $\mathcal{A}^2 = M^2A^2$.

Let us introduce the following notation $(i = 1, 2)$:

$$\beta^i(t) = [\beta^i_x(t),\ \beta^i_y(t),\ \beta^i_z(t)]^T = \int_{t_A}^{t}\int_{t_A}^{t'} \mathcal{A}^i(\tau)d\tau dt' \tag{5}$$

Note that these quantities are directly provided by the IMU measurements delivered in the interval $(t_A,\ t)$.

By integrating the second equation in (4) between t_A and a given $t' \in [t_A, t_B]$ and by substituting in the first equation in (4) and integrating again, we obtain:

$$\xi(t) = \xi_A + \eta_A(t - t_A) + R_A\beta^2(t) - \beta^1(t) \tag{6}$$

Note that this equation provides $\xi(t)$ as a linear expression of 15 unknowns, which are the components of ξ_A, η_A and the matrix R_A. These unknowns are not independent because the matrix R_A is orthonormal (i.e., it is characterized by only three parameters instead of nine). We obtain the analytic solution of CoVISF in two separate steps. In the former, we build a linear system in these unknowns together with the unknown distances when the cameras perform the measurements (section III-A). We will call this linear system, the *linear system associated to CoVISF*. It will be denoted by Σ_{Lin}. In the latter step, we exploit the fact that R_A is orthonormal and we end up with a polynomial equation system that consists of three polynomial equations of second degree and several linear equations (section III-B). This PES will be denoted by $\mathcal{P}$.

A. Linear system (Σ_{Lin})

We distinguish the case when only $\mathcal{B}_1$ is equipped with a camera, from the case when both $\mathcal{B}_1$ and $\mathcal{B}_2$ are equipped with a camera. In this latter case we further distinguish the case when the observations from the two cameras are synchronized from the case where they are not.

1) Single camera: The camera on $\mathcal{B}_1$ provides the vector $P(t) = M^1(t)\xi(t)$, up to a scale. We denote by $\lambda(t)$ this scale (this is the distance between $\mathcal{B}_1$ and $\mathcal{B}_2$ 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 $\mathcal{B}_1$) provide precisely the unit vector $\mu(t)$. Indeed, the camera provides the unit vector along $P(t)$; then, to obtain $\mu(t)$, it suffices to pre multiply this unit vector by $[M^1(t)]^T$.

We assume 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 at the time t_j by the subscript j (e.g., $\lambda_j = \lambda(t_j)$). In this notation, equation (6) becomes:

$$\lambda_j\mu_j = \xi_A + \eta_A(t_j - t_A) + R_A\beta^2_j - \beta^1_j \tag{7}$$

This is a vector equation, 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. The unknowns are: (i) The distances $\lambda_1, \cdots, \lambda_n$; (ii) The three components of ξ_A; (iii) The three components of η_A; (iv) The nine entries of the matrix R_A. The linear system given in (7) is precisely Σ_{Lin} when only $\mathcal{B}_1$ is equipped by a camera.

2) Two cameras: Let us consider now the case when also $\mathcal{B}_2$ is equipped with a camera and, the measurements made by this camera, occur at the times $t_{j'}$ $(j' = 1, \cdots, n')$.

By proceeding as in III-A1 we obtain the following additional set of linear equations

$$\lambda_{j'}\nu_{j'} = \xi'_{A'} + \eta'_{A'}(t_{j'} - t_{A'}) + R'_{A'}\beta^1_{j'} - \beta^2_{j'} \qquad (8)$$

where the quantities $\nu,\ \xi',\ \eta',\ R'$ are defined as $\mu,\ \xi,\ \eta,\ R$ but by changing $\mathcal{B}_1$ with $\mathcal{B}_2$. In this case the solution consists of the solutions of two independent problems. The only interesting case occurs when the two cameras are synchronized. In this case $\lambda_j\nu_j = -R_A^T\xi_j$. By substituting in this equation $\xi_j = \lambda_j\mu_j$ we obtain:

$$R_A\nu_j = -\mu_j \qquad (9)$$

Hence, in this case Σ_{Lin} is characterized by the linear equations in (7) and (9).

We conclude this section by providing Σ_{Lin} in matrix form. We have $Mx = b$ with:

$$x = [\xi_A^T,\ \eta_A^T,\ \lambda_1,\ \cdots,\ \lambda_n,\ (E^1)^T,\ (E^2)^T,\ (E^3)^T]^T,$$
$$b = [0_3^T,\ (\beta_2^1)^T,\ \cdots,\ (\beta_n^1)^T,\ -\mu_1^T,\ \cdots,\ -\mu_n^T]^T$$

$$M = \left[\begin{array}{c|c|c|c|c|c|c} I_3 & 0_{3\times3} & -\mu_1 & 0_3 & \cdots & 0_3 & 0_{3\times9} \\ I_3 & \Delta_2 I_3 & 0_3 & -\mu_2 & \cdots & 0_3 & \Gamma_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ I_3 & \Delta_n I_3 & 0_3 & 0_3 & \cdots & -\mu_n & \Gamma_n \\ 0_{3\times3} & 0_{3\times3} & 0_3 & 0_3 & \cdots & 0_3 & \mathcal{V}_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0_{3\times3} & 0_{3\times3} & 0_3 & 0_3 & \cdots & 0_3 & \mathcal{V}_n \end{array}\right] \qquad (10)$$

where 0_3 is the zero 3×1 vector, $0_{3\times3}$ the zero 3×3 matrix, E^1, E^2, E^3 are the three columns of R_A, $\Delta_j = t_j - t_A$, Γ_j is the 3×9 matrix block $\Gamma_j \triangleq \left[\begin{array}{c|c|c} \beta^2_{xj} I_3 & \beta^2_{yj} I_3 & \beta^2_{zj} I_3 \end{array}\right]$ and $\mathcal{V}_j$ is the 3×9 matrix block $\mathcal{V}_j \triangleq \left[\begin{array}{c|c|c} \nu_{xj} I_3 & \nu_{yj} I_3 & \nu_{zj} I_3 \end{array}\right]$. In the case when only $\mathcal{B}_1$ is equipped by a camera, M and b only include the first $3n$ lines.

B. Polynomial equation system ($\mathcal{P}$) and its solution

So far, we have obtained a system of linear equations in $15+n$ unknowns. In the case of a single camera, the equations are the $3n$ scalar equations given in (7). In the case of two cameras synchronized we also have the $3n$ equations given in (9). On the other hand, the unknowns are not independent since the matrix R_A is orthonormal. As a result, it is defined by three parameters instead of nine. To account for this, we proceed with the following three steps:

1) Elimination of ξ_A, η_A and $\lambda_1, \cdots, \lambda_n$ from the linear system by using part of its equations.
2) Quaternion parametrization of R_A.
3) Reduction to a quadratic system in three unknowns.

First step: We start by eliminating ξ_A, η_A and $\lambda_1, \cdots, \lambda_n$. This can be done by following several approaches. Basically, we need to use $6 + n$ independent equations of the linear system to express ξ_A, η_A and $\lambda_1, \cdots, \lambda_n$ in terms of the entries of R_A and then substitute these expressions in the remaining equations of the linear system. As a result, we obtain a linear system in nine unknowns, which are the entries of R_A. We obtain this elimination, by using a QR factorization of our linear system. Then, we use the last $6 + n$ equations to obtain the components of ξ_A, η_A and $\lambda_1, \cdots, \lambda_n$ in terms of the nine entries of R_A. In the case of a single camera, the resulting system in the nine entries of R_A consists of $3n - (6 + n) = 2n - 6$ equations. In the case of two cameras, it consists of $5n - 6$ equations.

Second step: We use the following parametrization: $R_A =$

$$\left[\begin{array}{ccc} a^2+b^2-c^2-d^2 & 2(bc-ad) & 2(ac+bd)i \\ 2(ad+bc) & a^2-b^2+c^2-d^2 & 2(cd-ab)i \\ 2(bd-ac) & 2(ab+cd) & a^2-b^2-c^2+d^2 \end{array}\right]$$

with $a^2 + b^2 + c^2 + d^2 = 1$.

Third step: By using this expression for the nine entries of R_A in the linear system obtained at the first step, we would obtain a system of polynomial equations of second degree in the four unknowns a, b, c, d. However, it is much more preferable to eliminate one of the unknowns by proceeding as follows. We set $d = 1$ and we obtain a system of polynomial equations of second degree in the three unknowns a, b, c. By doing this, the resulting matrix R_A remains orthogonal, but not orthonormal. We enforce this last condition by normalizing the columns of the matrix at the end. Hence, we need to solve a system of $2n - 6$ (or $5n - 6$ in the case of two cameras) polynomial equations of second degree in the three unknowns a, b, c. For that, we proceed into 2 separate steps. In the first we extract three equations. We solve the system of three polynomial in three unknowns by using the method in [23] (see also [25]), which is based on the Macaulay resultant matrices. Obviously, we discard all the solutions which are not real. The number of real solutions are up to 8. Then, it suffices to use a further equation (independent from the three extracted at the beginning) to obtain a unique solution. In the second step, we use this unique solution to initialize the minimization of a cost function that is the square of the residual of the entire equations system. Once a, b, c have been determined, we obtain an orthogonal matrix with the expression given above (where $d = 1$). We obtain the matrix R_A by normalizing the columns of this matrix. By substituting its entries in the expressions obtained in the first (elimination) step, we finally obtain also ξ_A, η_A and $\lambda_1, \cdots, \lambda_n$.

C. Polynomial equations system in P_B, V_B and R_B

In order to obtain the PES in P_B, V_B and R_B, we can proceed exactly as above with the following two changes:

1) All the time integrals must be computed from t_B instead of t_A. As a result, they are computed in the reversal time direction.
2) The new frame on each body coincides with the local frame at the final time t_B instead of t_A (and, consequently, $\xi_B = P_B$ and $\eta_B = V_B$).

In practice, instead of equation (7) we have:

$$\lambda_j\mu_j = \xi_B + \eta_B(t_j - t_B) + R_B\beta^2_j - \beta^1_j \qquad (11)$$

where β^1_j and β^2_j are given in (5) but with t_B instead of t_A and the matrices $M_1(t)$ and $M_2(t)$ ($t \in (t_A,\ t_B)$) are obtained by integrating (3) from t_B to t. Finally, equation (9) remains the same (with R_B instead of R_A) and the procedure in section III-B remains the same.

IV. Basic Structural Properties

On the basis of the analytic results derived in the previous section it is possible to obtain all the structural properties of CoVISF. In particular, it is possible to study how the number of solutions of CoVISF depends on the relative motion between the agents and on the number of camera images. Specifically, this is obtained by studying the analytic properties of $\mathcal{P}$. Note that, for this analysis, we do not need to optimally solve $\mathcal{P}$. Indeed, the structural properties of CoVISF are independent of the measurements' noise. In this section, we provide two fundamental structural properties. The former is a necessary condition that characterizes the observability of the absolute scale. The latter regards the observability of the relative orientation. This latter property only holds in the case of two synchronized cameras and it enormously simplifies the analysis of the structural properties of CoVISF. In particular, it allows us to simply refer to Σ_{Lin}(instead of $\mathcal{P}$), in the case of two synchronized cameras.

Property 1 (Scale invariance) *If the relative inertial acceleration between the two bodies is null, the CoVISF is scale invariant (i.e., the absolute scale cannot be determined).*

Proof: When the relative inertial acceleration is null, the relative motion is characterized by a constant velocity, which is the velocity at the initial time. Hence, we have:

$$\xi_j = \xi_A + \eta_A(t_j - t_A), \qquad j = 2, \cdots, n$$

Hence, the following vector is a killing vector of Σ_{Lin} (i.e., it belongs to the null space of the matrix in (10)):

$$n = [\xi_A^T,\ \eta_A^T,\ \lambda_1,\ \lambda_2,\ \lambda_3, \cdots,\ \lambda_n,\ 0_3^T,\ 0_3^T,\ 0_3^T]^T \quad (12)$$

The existence of this killing vector reveals the scale invariance of CoVISF. Indeed, if x_0 solves $\mathcal{P}$, any $x = x_0 + \gamma n$ solves $\mathcal{P}$, for any scalar γ. In other words, the solutions of $\mathcal{P}$ are invariant under the transform $\xi_A \rightarrow (1+\gamma)\xi_A$, $\eta_A \rightarrow (1+\gamma)\eta_A$, $\lambda_j \rightarrow (1+\gamma)\lambda_j$, $j = 1, \cdots, n$.

To obtain the second property, we need to introduce the concept of *collinear* relative motion. Specifically, with *collinear* relative motion, we mean that $\mu_1 = \mu_2 = \cdots = \mu_n$. This condition is equivalent to the condition $\nu_1 = \nu_2 = \cdots = \nu_n$. In the case of two synchronized cameras, the following fundamental property holds:

Property 2 (Separation) *In the case of two synchronized cameras, if the relative motion is not collinear, R_A is fully observable and its determination can be separated from the determination of the remaining unknowns.*

Proof: This is a trivial consequence of the fact that, by using equation (9) with two distinct unit vectors (ν_i and ν_j) we can uniquely determine R_A.

This property is very important. It allows us to obtain all the singularities and minimal cases of CoVISF in the case of two synchronized cameras by simply studying the rank of a linear system. Indeed, once R_A is known, the remaining unknowns are only conditioned by the linear system that is characterized by the matrix that is the upper left $3n \times (6+n)$ block of the matrix M in (10). By studying how the rank of this matrix depends on the relative trajectory and the number of camera images, we obtain all the structural properties of CoVISF. This will be the matter of a future work together with the study of the more difficult case that occurs when only one agent is equipped by a camera (or when the two cameras are not synchronized).

V. Performance Evaluation

In this section, we evaluate the performance of the analytic solution of CoVISF in presence of noise. In particular, we compare its performance with the one of the closed-form solution introduced in [21], [22]. From now on, we denote this latter solution by $\mathcal{C}$ and the analytic solution of CoVISF by $\mathcal{A}$. We evaluate the performance of $\mathcal{C}$ and $\mathcal{A}$ for changing lengths of the time interval $(t_A,\ t_B)$. Additionally, we evaluate their robustness against noisy measurements and also against a non perfect synchronization between the two cameras. Finally, we evaluate their performance also against the trajectories accomplished by the agents. In particular, on the basis of the theoretical result stated by Property 1, we consider trajectories characterized by different accelerations. This analysis is carried out by using synthetic measurements (section V-A). In section V-B we also provide the results obtained with real measurements.

A. Simulations

1) Simulated trajectories and sensors: The trajectories are simulated as follows. First, we characterize the configuration of each agent with its position (p) its speed (v) and its orientation in a global frame. The orientation is characterized by a unit quaternion q. In this notation, the dynamics of each agent satisfy the following 3 differential equations: $\dot{p} = v$, $\dot{v} = qAq^* - gk$ and $\dot{q} = \frac{1}{2}q\Omega$, where g is the magnitude of the gravity, k is the fourth fundamental quaternion unit ($k = 0 + 0\,i + 0\,j + 1\,k$), A is the acceleration in the local frame (which includes the gravity) and Ω is the angular speed in the local frame. These equations are discretized with a time step of $0.002\ s$. For each trial, the initial position of the first agent is set always equal to $[0,\ 0,\ 0]m$ while the initial position of the second agent is randomly generated, with a normal distribution, centered at the origin, and with covariance matrix $1\ m^2 I_3$. The initial velocities of both the agents are randomly generated. Specifically, their values are normally distributed, with zero mean, and covariance matrix $1\ (m/s)^2 I_3$. Finally, the initial orientations are characterized by the roll, pitch and yaw angles. These are also randomly generated, with zero mean and covariance matrix $(50\ deg)^2 I_3$. The trajectories of both the agents are also randomly generated. The angular speeds are Gaussian. Specifically, their values at each step follow a zero-mean Gaussian distribution with covariance matrix equal to $(1\ deg)^2 I_3$. At each time step, the agents' speeds are incremented by setting the inertial acceleration 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 $\sigma = 1\ ms^{-2}$, unless otherwise indicated.

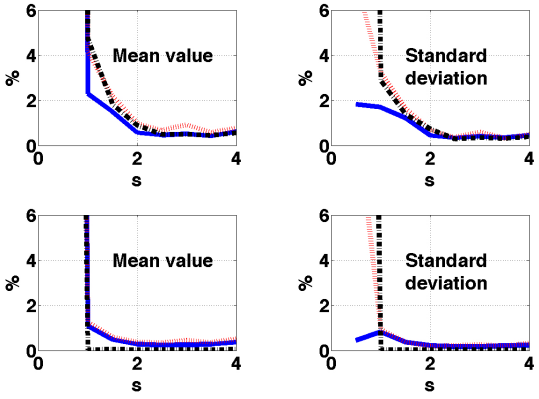

Fig. 2. Relative error of the analytic solution in determining the absolute scale (solid blue), the relative speed (dotted red) and the relative orientation (dashed black). The upper two plots show the results obtained with $\mathcal{C}$, while the lower two plots the results obtained with $\mathcal{A}$. All the values are obtained by running 1000 trials (mean value of the relative error on the left and standard deviation on the right). The two agents observe one each other over a variable duration of integration $(t_B - t_A)$. $\sigma_{Accel} = 0.03\ ms^{-2}$ and $\sigma_{Gyro} = 0.1\ deg\ s^{-1}$.

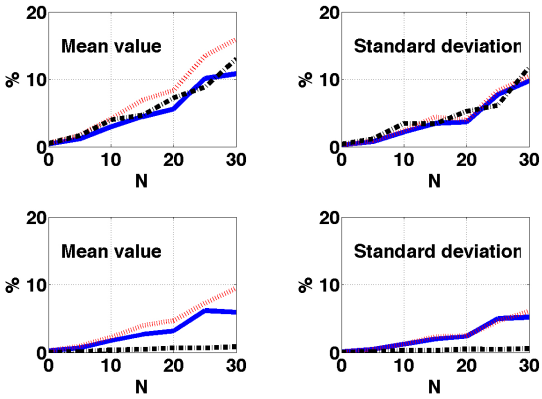

Fig. 3. As in Fig. 2 but for a variable noise on the inertial measurements ($\sigma_{Accel} = (N \cdot 0.03)\ ms^{-2}$ and $\sigma_{Gyro} = (N \cdot 0.1)\ deg\ s^{-1}$). The two agents observe one each other over 3 seconds.

The agents 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 $0.002\ s$ by adding to the true value a random error that follows a Gaussian distribution. The mean value of this error is zero. The standard deviation will be denoted by σ_{Accel} for the accelerometer and σ_{Gyro} for the gyroscope (these values will be specified for each result). 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 $1\ deg^2$.

2) Estimation results: We provide the precision of $\mathcal{C}$ and $\mathcal{A}$ in estimating the absolute scale the relative speed and the relative orientation. For these three quantities, we provide the relative error (in $\%$), which is obtained by performing 1000 trials. We provide both the mean value and the standard

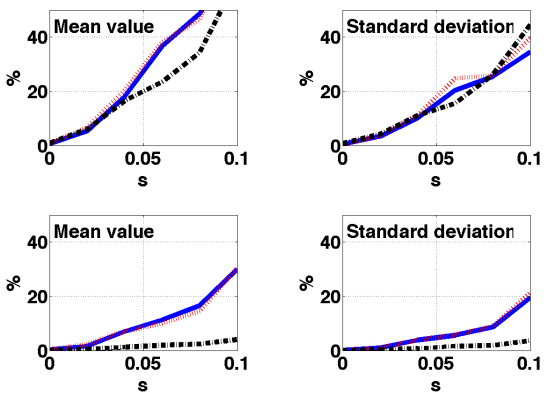

Fig. 4. As in Fig. 2 but for a variable synchronization error (Δ_t, in seconds). The two agents observe one each other over 3 seconds

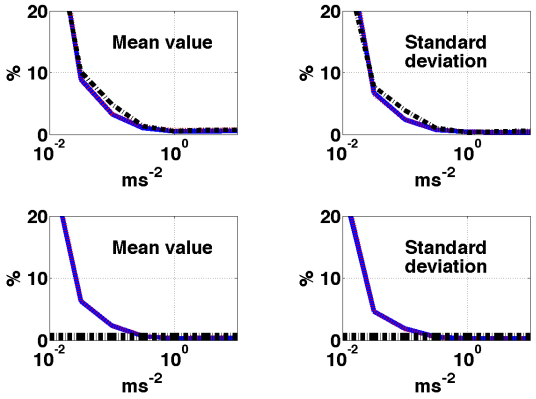

Fig. 5. As in Fig. 2 but for a variable σ, i.e., the standard deviation adopted to randomly generate the acceleration of the two agents' trajectories.

deviation. Figure 2 provides this relative error for a variable duration of the considered interval $(t_A,\ t_B)$. In other words, we provide the relative error vs the length $t_B - t_A$. In this case, we set $\sigma_{Accel} = 0.03\ ms^{-2}$ and $\sigma_{Gyro} = 0.1\ deg\ s^{-1}$. The upper two plots show the results obtained with $\mathcal{C}$, while the lower two plots the results obtained with $\mathcal{A}$. Note how the evaluations get better as we increase the integration time. The relative orientation is determined with a final error of 0.047% for $\mathcal{A}$ and 0.42% for $\mathcal{C}$, the relative speed with a final error of 0.67% for $\mathcal{A}$ and 0.83% for $\mathcal{C}$ and the relative position with a final error of 0.41% for $\mathcal{A}$ and 0.54% for $\mathcal{C}$. In the case of $\mathcal{A}$, a time interval of 1.5 seconds is sufficient to achieve the aforementioned precision. Regarding $\mathcal{C}$, it is necessary a longer interval (about 3 seconds). $\mathcal{A}$ always outperforms $\mathcal{C}$. The improvement is very significant for the relative orientation (one order of magnitude). For the relative position and speed, the improvement is of about 20%.

Regarding the computational complexity, both $\mathcal{A}$ and $\mathcal{C}$ are very efficient. For $t_B - t_A = 4s$, the time of computation on MATLAB r2011a, 3.1 GHz Intel Core i7, 8GB RAM, OS X El Capitan is $1.1\ 10^{-3}\ s$ for $\mathcal{A}$ and $0.46\ 10^{-3}\ s$ for $\mathcal{C}$.

Fig 3 displays the relative error for the same quantities showed in Fig. 2 but for a variable noise on the inertial measurements. Specifically, $\sigma_{Accel} = (N \cdot 0.03)\ ms^{-2}$ and

$\sigma_{Gyro} = (N \cdot 0.1)\ deg\ s^{-1}$. In this case, the two agents observe one each other over 3 seconds. The general behaviour remains the same. Note that the noise is very large (standard sensors are characterized by $N \simeq 1$). The performance remains very good also for very large noise. We remark that $\mathcal{A}$ outperforms $\mathcal{C}$ and the improvement is significant for the relative orientation.

Fig 4 displays the relative error for the same quantities showed in Fig. 2 but for a variable synchronization error between the two cameras. Specifically, the measurements of the second camera are generated with a delay of Δt seconds. $\mathcal{A}$ is much more robust than $\mathcal{C}$. Its performance on the relative orientation remains very good, in contrast with the one of $\mathcal{C}$. Regarding the absolute scale, the relative error is smaller than 5% for $\Delta t \leq 0.02\ s$, in the case of $\mathcal{A}$ and smaller than 10% in the case of $\mathcal{C}$. Fig 5 displays the relative error for the same quantities showed in Fig. 2 but for a variable σ, i.e., the standard deviation adopted to randomly generate the acceleration of the two agents' trajectories. As expected, the precision on the absolute scale and on the speed improves by increasing σ, while the precision on the relative orientation is unaffected by the acceleration (see Property 1). On the other hand, in the case of $\mathcal{C}$, also the relative orientation is strongly affected by the acceleration. Note that the solution in [21], [22] is a generic closed-form solution, i.e., it is a heuristic procedure that does not meet necessarily the structural properties of the problems.

B. Real Experiments

We used a real dataset containing IMU and camera measurements. Specifically, we used two Intel aero RTF drones.

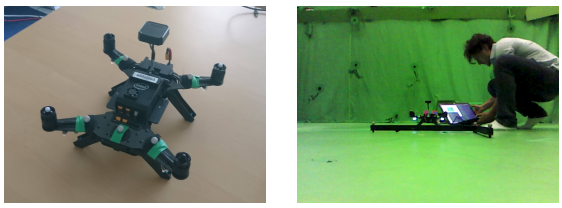

Fig. 6. One of the two drones adopted in our experiment. On the left it is equipped with the pattern that allows the motion capture system to detect the drones. On the right it is equipped with the two red tags that allows the detection of the drone by the camera of the other drone.

1) Experimental setup: The experiments were performed in a room equipped with a motion-capture system. This allowed us to compare the estimations of the relative position between the two drones, relative speed and relative orientation against ground truth. The drones moved indoor at low altitude (0:2m). The orientation of each drone was kept almost constant in order to allow them to observe each other during the entire experiment. The two drones moved along loops of radius $\simeq 1\ m$. Their acceleration was centripetal and with almost constant magnitude ($\simeq 1\ ms^{-2}$). The relative acceleration significantly changed, depending on the position of each drone in the loop. Fig 6 shows one of the two drones. In particular, on the left it is equipped with the pattern that allows the motion capture system its detection. On the right, it is equipped with the two red tags that allows its detection from the on-board camera of the other drone.

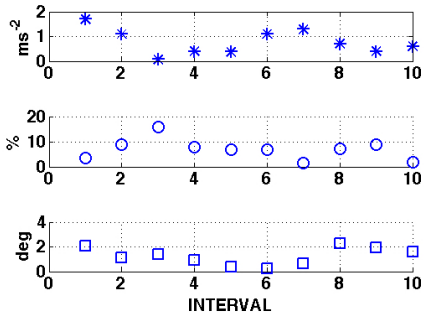

Fig. 7. Real experiment. From the bottom to the top the precision on the relative orientation, on the absolute scale and the magnitude of the relative acceleration (averaged on the corresponding time interval).

2) Results: The experiment lasted about one minute. We selected 10 intervals of 3 seconds.

Figure 7 shows the precision of the analytic solution. Specifically, the plot on the bottom displays the error on the relative orientation, the plot on the middle the error on the scale. The upper plot displays the magnitude of the relative acceleration obtained from the ground truth. The precision is excellent for the relative orientation. In particular, it never exceeds 2.3 deg and in few cases it is smaller than 1 deg. Regarding the absolute scale, the performance is worse. The relative error is always smaller than 15% with the exception of one case (16%). In two cases it is smaller than 2% (7^{th} and 10^{th} interval). From Figure 7, we remark that the error on the scale is correlated with the magnitude of the relative acceleration (where the error on the scale is high, the magnitude of the relative acceleration is small). This is consistent with the result stated by Property 1 and with the experiments performed with a single agent, which evidence how the precision on the scale needs a strong excitation [26]. On the other hand, both for the scale and for the relative orientation, the precision is definitely worse than the one obtained with simulations. Possible sources of the error could be: (i) Time delay between camera and IMU measurements; (ii) Time delay between the two drones (see Fig 4); (iii) Inaccurate camera-IMU extrinsic calibration.

VI. Conclusion

This paper analyzed the problem of visual inertial sensor fusion in the cooperative case. Specifically, the case of two agents was investigated. For this problem the paper provided the analytic solution. In particular, the problem was transformed in a simple Polynomial Equations System (PES). The main paper contribution was precisely the establishment of the equivalence between the cooperative visual inertial sensor fusion in the case of two agents (as defined in section II) and the aforementioned PES. This is the extension of the PES derived in [7] to the cooperative case. In that case (single agent), the PES contains only a single polynomial of second degree and the minimal cases have up to two solutions. In the case analyzed in this paper (two agents), the minimal cases contain three polynomials of second degree and the number of distinct solutions is up to eight.

The power of the analytic solution is twofold. From one side, it allows us to determine the state without the need of an initialization. From another side, it provides fundamental insights into all the theoretical aspects of the problem.

In this paper we mainly focused on the first issue. The PES was solved by using the method based on the Macaulay resultant matrices [23]. The PES has the following feature. The unknowns are the components of the relative state while the coefficients of the system and the vector with the constant terms only depend on the measurements from the two cameras and the inertial sensors delivered during a short time interval. As a result, the determination of the relative position (which includes the absolute scale), the relative speed and the relative orientation does not need any prior knowledge (initialization) and it is drift-free. Extensive simulations and real experiments clearly showed that the analytic solution is very powerful. In particular, we tested its robustness with respect to noisy measurements and with respect to a synchronization error between the cameras. We also tested the performance for various trajectories, in particular, characterized by different accelerations. We compared the performance of the analytic solution with the one of the closed-form solution introduced in [21], [22]. The analytic solution significantly outperforms the latter. The improvement is very significant for the relative orientation. The error on the relative position and speed is 20% smaller when estimated by the analytic solution while the error on the relative orientation is one order of magnitude smaller. In addition, the analytic solution is much more robust with respect to the noise, to an erroneous synchronization between the two cameras and with respect to the magnitude of the acceleration that characterizes the trajectories.

Regarding the second investigation, we provided two basic structural properties. These properties establish how the observability of the absolute scale and the relative orientation depend on the trajectories. Additionally, they provide fundamental tools to obtain an exhaustive analysis of all the problem singularities, degeneracies and minimal cases. This exhaustive analysis is currently under investigation. So far, we have obtained that the minimal case that provides 8 distinct solutions, can only occur in the case of a single camera and $n = 4$ camera images. With two synchronized cameras, we have at most two distinct solutions in the minimal case and, in general, a unique solution.

In this paper we did not present the results obtained by considering a bias in the inertial measurements. We obtained results very similar to the ones presented in [21], [22], where we analyzed the effect of a bias on the performance of the closed-form solution introduced in [21], [22].

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, vol. 26, pp. 577–589, 2007.

[2] C. Forster, M. Pizzoli, and D. Scaramuzza, Svo: Fast semi-direct monocular visual odometry, IEEE International Conference on Robotics and Automation, 2014, Hong Kong, China.

[3] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, On the complexity and consistency of ukf-based slam, IEEE International Conference on Robotics and Automation, 2009, Kobe, Japan.

[4] 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, pp. 407–430, 2011.

[5] M. Li and A. I. Mourikis, High-precision, consistent ekf-based visual-inertial odometry, The International Journal of Robotics Research, vol. 32, pp. 690–711, 2013.

[6] T. Liu and S. Shen, Spline-Based Initialization of Monocular Visual-Inertial State Estimators at High Altitude, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, Vancouver, CA.

[7] A. Martinelli, Closed-form solution of visual-inertial structure from motion, The International Journal of Computer Vision, vol. 106, pp. 138–152, 2014.

[8] J. Mustaniemi, J. Kannala, S. Sarkka, J. Matas and J. Heikkila, Inertial-Based Scale Estimation for Structure from Motion on Mobile Devices, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, Vancouver, CA.

[9] J. Hesch, D. Kottas, S. Bowman, and S. Roumeliotis, Consistency analysis and improvement of vision-aided inertial navigation, IEEE Transactions on Robotics, vol. 30, pp. 158–176, 2014.

[10] G. Huang, M. Kaess, and J. J. Leonard, Towards consistent visual-inertial navigation, IEEE International Conference on Robotics and Automation, 2015, Seattle, USA.

[11] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation, Robotics: Science and Systems, 2015, Rome, Italy.

[12] G. Huang, A. Mourikis, S. Roumeliotis, An observability-constrained sliding window filter for slam, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011, S. Francisco, USA.

[13] 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.

[14] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, Keyframe-based visual-inertial odometry using nonlinear optimization, The International Journal of Robotics Research, vol 34, pp. 314–334, 2015.

[15] T. Lupton and S. Sukkarieh, Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions, IEEE Transactions on Robotics, vol. 28, pp. 61–76, 2012.

[16] A. Mourikis and S. Roumeliotis, A multi-state constraint kalman filter for vision-aided inertial navigation, IEEE International Conference on Robotics and Automation, 2007, Rome, Italy.

[17] A. Mourikis and S. Roumeliotis, A dual-layer estimator architecture for long-term localization, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, Anchorage, USA.

[18] 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, vol 2, pp. 18–25, 2017.

[19] H. Guo Sartipi, R. DuToit, G. Georgiou, R. Li, J. O?Leary, E. Nerurkar, J. Hesch, S. Roumeliotis, Large-Scale Cooperative 3D Visual-Inertial Mapping in a Manhattan World, IEEE International Conference on Robotics and Automation, 2016, Stockholm, Sweden.

[20] A. Martinelli and A. Renzaglia, Cooperative Visual-Inertial Sensor Fusion: Fundamental Equations, The International Symposium on Multi-Robot and Multi-Agent Systems, 2017, Los Angeles, USA.

[21] A.Martinelli, Closed-form Solution to Cooperative Visual-Inertial Structure from Motion, 2018, arXiv:1802.08515 [cs.RO]

[22] A. Martinelli, A. Renzaglia and A. Oliva, Cooperative Visual-Inertial Sensor Fusion: Fundamental Equations and State determination in Closed Form, Autonomous Robots *conditionally accepted 2018*.

[23] S. Telen, B. Mourrain and M. Van Barel, Solving Polynomial Systems via a Stabilized Representation of Quotient Algebras, SIAM Journal on Matrix Analysis and Applications, vol 39, pp.1421–1447, 2018.

[24] G.L. Mariottini, S. Martini, and M. Egerstedt, A Switching Active Sensing Strategy to Maintain Observability for Vision-Based Formation Control, IEEE International Conference on Robotics and Automation, 2009, Kobe, Japan.

[25] M. Elkadi and B. Mourrain, Introduction à la résolution des systèmes polynomiaux, Mathématiques et Applications, Vol. 59, Springer, 2007.

[26] S. M. Weiss, Vision based navigation for micro helicopters, PhD Thesis, ETH Zurich, 2012.