Closed-Form Solution for Attitude and Speed Determination by Fusing Monocular Vision and Inertial Sensor Measurements

Agostino Martinelli

Closed-Form Solution for Attitude and Speed Determination by Fusing Monocular Vision and Inertial Sensor Measurements

Agostino Martinelli

***Abstract*—This paper considers the problem of data fusion when the adopted sensors are a monocular camera and inertial sensors (i.e. one tri-axial accelerometer and one tri-axial gyrometer). The investigation starts by performing an observability analysis to analytically derive all the observable modes, i.e. all the physical quantities that the information contained in the sensor data allows us to estimate. They are the position of the features in the camera frame, the vehicle speed in the same local frame and the absolute roll and pitch angles. The main contribution of the paper is a new algorithm to simultaneously estimate all the previous physical quantities. In particular, the algorithm is based on a closed-form solution which analytically expresses the vehicle speed and attitude in terms of the sensor measurements. In this algorithm the camera only needs to observe four times a single point feature in the environment. This allows performing the overall estimation in a very short time interval and without the need of any initialization or a priori knowledge. This is a key advantage since allows eliminating the drift on the scale factor and on the vehicle orientation. In addition, the algorithm can be easily extended in order to deal with biased inertial measurements and to deal with multiple features, in which case only three distinct camera poses are required (instead of four). Specifically, with three camera poses and two features, the vehicle speed and attitude together with the scale factor can be determined. The performance of the proposed approach is evaluated via Monte Carlo simulations and by using real data.**

***Index Terms*—Sensor Fusion, Inertial Sensors, Vision, Non linear Observability, Aerial Robotics**

I. Introduction

In recent years, vision and inertial sensing have received great attention by the mobile robotics community. These sensors require no external infrastructure and this is a key advantage for robots operating in unknown environments where GPS signals are shadowed. In addition, these sensors have very interesting complementarities and together provide rich information to build a system capable of vision-aided inertial navigation and mapping and a great effort has been done very recently in this direction (e.g. [16], [3], [1]). A special issue of the *International Journal of Robotics Research* has recently been devoted to the integration of vision and inertial sensors [6]. In [5], a tutorial introduction to the vision and inertial sensing is presented. This work provides a biological point of view and it illustrates how vision and inertial sensors have useful complementarities allowing them to cover the respective limitations and deficiencies. The majority of the approaches so far introduced, perform the fusion of vision and inertial sensors by filter-based algorithms. In [2], these sensors are used to perform egomotion estimation. The sensor fusion is obtained with an Extended Kalman Filter (EKF) and with an Unscented Kalman Filter (UKF). The approach proposed in [8] extends the previous one by also estimating the structure of the environment where the motion occurs. In particular, new landmarks are inserted on line into the estimated map. This approach has been validated by conducting experiments in a known environment where a ground truth was available. Also, in [18] an EKF has been adopted. In this case, the proposed algorithm estimates a state containing the robot speed, position and attitude, together with the inertial sensor biases and the location of the features of interest. In the framework of airbone SLAM, an EKF has been adopted in [10] to perform $3D-$SLAM by fusing inertial and vision measurements. It was remarked that any inconsistent attitude update severely affects any SLAM solution. The authors proposed to separate attitude update from position and velocity update. Alternatively, they proposed to use additional velocity observations, such as air velocity observation. Regarding the robot attitude, in [4] it has been noted that roll and pitch angles remain more consistent than the heading.

A fundamental issue to address when fusing vision and inertial measurements, is to understand which are the *observable modes*, i.e. the physical quantities that the information contained in the sensor data allows us to estimate. The next issue to address is to find a reliable and efficient method to estimate all the previous physical quantities. It is very reasonable to expect that the scale factor is an observable mode and can be obtained by a closed-form solution. Let us consider the trivial case where a robot, equipped with a bearing sensor (e.g. a camera) and an accelerometer, moves on a line (see fig 1). If the initial speed in A is known, by integrating the data from the accelerometer, it is possible to determine the robot speed during the subsequent time steps and then the distances $A-B$ and $B-C$ by integrating the speed. The lengths $A-F$ and $B-F$ are obtained by a simple triangulation by using the two angles β_A and β_B from the bearing sensor. Let us now assume that the initial speed v_A is unknown. In this case, all the previous segment lengths can be obtained in terms of v_A. In other words, we obtain the analytical expression of $A-F$ and $B-F$ in terms of the unknown v_A and all the sensor measurements performed while the robot navigates from A to B. By repeating the same computation with the bearing measurements in A and C, we have a further analytical expression for the segment $A-F$, in terms of the unknown v_A and the sensor measurements performed while the robot navigates from A to C. The two expressions for $A-F$ provide an equation in the unknown v_A. By solving this equation we finally obtain all the lengths in terms of the measurements performed by the accelerometer and the bearing sensor.

The previous example is very simple because of several unrealistic restrictions. First of all, the motion is constrained

A. Martinelli is with INRIA Rhone Alpes, Montbonnot, France e-mail: agostino.martinelli@ieee.org

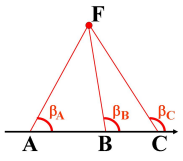

Fig. 1. A robot equipped with an accelerometer and a camera moves on a line. The camera performs three observations of the feature in F, repsectively from the points A, B and C.

on a line. Additionally, the accelerometer provides gravity-free and unbiased measurements. In this paper we will relax these restrictions[1] by considering the case of a robot equipped with IMU and bearing sensors. We want to know which are the observable modes, namely the physical quantities that we can estimate without any a priori knowledge (i.e. by only collecting the data from the previous sensors during a short time interval). For instance, are the scale factor, the robot speed and the robot orientation observable modes? If yes, is it possible to perform their estimation by a closed-form solution? Does the derived solution work independently of the vehicle trajectory?

An answer to the first question can be found by applying the method introduced in [13], where a non standard observability analysis, based on the new concept of continuous symmetry, has been introduced. The advantages of this non standard observability analysis is that, in contrast to previous approaches, it is able not only to check whether a given state is observable or not, but, in the negative case, it is also able to detect the quantities which are observable. In particular, by analyzing the continuous symmetries of a given system, it is possible to obtain a system of partial differential equations. The observable modes are all the independent solutions of this system of partial differential equations. In addition, this analysis can also be used to find necessary conditions on the vehicle trajectory in order to guarantee the observability of the observable modes.

In this paper we consider a MAV equipped with a single camera and IMU sensors (one tri-axial accelerometer and one tri-axial gyrometer). We perform an observability analysis (based on the approach introduced in [13]) in order to understand which are the physical quantities that the information contained in these data allows us to estimate. In particular, this analysis allows us to analytically derive all the observable modes. Specifically, they are the position of the observed feature and the vehicle speed and attitude. Then, we derive a closed-form solution which analytically expresses the previous physical quantities in terms of the sensor measurements. This allows us to introduce a very simple and efficient algorithm to perform the estimation. This algorithm only requires to observe four times a single point feature. In this case of one single feature, the algorithm estimates eight independent quantities. They are: the three components of the speed in the camera frame, the three components of the $3D$ position of the observed feature in the same camera frame and the roll and pitch angles. In [14] and [15] we also derive necessary conditions on the vehicle trajectory in order to guarantee the observability of the vehicle speed and attitude.

[1]The case of biased measurements is actually dealt in [14] and [15].

Compared to the state-of-the-art the advantages of the proposed algorithm are:

1) It only requires a single point feature to work;
2) The estimation does not require any initialization;
3) The estimation can be performed at any moment by only collecting four consecutive observations;
4) It allows eliminating the drift on the scale factor and on the roll and pitch angles independently of the vehicle trajectory;
5) It can be easily extended in order to deal with the case of multiple features and biased inertial measurements ([14] and [15]).

Regarding the fourth advantage, we remark that other methods which try to bound the error by using a Bundle Adjustment approach (e.g. [7]), cannot fully eliminate the drift on the previous mentioned quantities. In particular, when the vehicle moves in large environments without closing any loop (i.e. without revisiting regions previously explored) the drift becomes a serious inconvenient.

A further contribution of this paper is a local decomposition of the system which separates the observable modes from the rest of the system. This allows us to implement a filter based approach to directly estimate the observable modes. In particular, we will use this decomposition to implement an Extended Kalman Filter (EKF).

The paper is organized as follows. Section II provides a mathematical description of the system. Starting from this description, in section III an observability analysis which accounts the system non linearities is provided. Then, in section IV we derive the closed-form solution and the algorithm to estimate the vehicle speed and attitude. In section VI we evaluate the performance of the proposed algorithm based on the closed-form solution by using synthetic and real data. Additionally, we compare two EKFs estimating respectively the entire vehicle state and the state which only contains the observable modes. Finally, conclusions are provided in section VII.

II. The Considered System

Let us consider an aerial vehicle equipped with a monocular camera and IMU sensors. The IMU consists of three orthogonal accelerometers and three orthogonal gyrometers. We assume that the transformations among the camera frame and the IMU frames are known (we can assume that the vehicle frame coincides with the camera frame). The IMU provides the vehicle angular speed and acceleration. Actually, regarding the acceleration, the one perceived by the accelerometer ($\boldsymbol{A}$) is not simply the vehicle acceleration ($\boldsymbol{A_v}$). It also contains the gravity acceleration ($\boldsymbol{A_g}$). In particular, we have $\boldsymbol{A} = \boldsymbol{A_v} - \boldsymbol{A_g}$ since, when the camera does not accelerate (i.e. $\boldsymbol{A_v}$ is zero) the accelerometer perceives an acceleration which is the same of an object accelerated upward in the absence of gravity.

We will use uppercase letters when the vectors are expressed in the local frame and lowercase letters when they are expressed in the global frame. Hence, regarding the gravity we have: $\boldsymbol{a_g} = [0,\ 0,\ -g]^T$, being $g \simeq 9.8\ ms^{-2}$.

We assume that the camera is observing a point feature during a given time interval. We fix a global frame attached to this feature. The vehicle and the feature are displayed in fig 2.

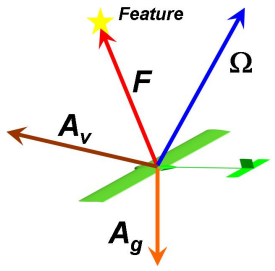

Fig. 2. The feature position ($\boldsymbol{F}$), the vehicle acceleration ($\boldsymbol{A_v}$) the vehicle angular speed ($\boldsymbol{\Omega}$) and the gravity acceleration ($\boldsymbol{A_g}$).

Finally, we will adopt a quaternion to represent the vehicle orientation. Indeed, even if this representation is redundant, it is very powerful since the dynamics can be expressed in a very easy and compact notation [11].

Our system is characterized by the state $[\boldsymbol{r},\ \boldsymbol{v},\ q]^T$ where $\boldsymbol{r} = [r_x,\ r_y,\ r_z]^T$ is the $3D$ vehicle position, $\boldsymbol{v}$ is its time derivative, i.e. the vehicle speed in the global frame ($\boldsymbol{v} \equiv \frac{d\boldsymbol{r}}{dt}$), $q = q_t + iq_x + jq_y + kq_z$ is a unitary quaternion (i.e. satisfying $q_t^2 + q_x^2 + q_y^2 + q_z^2 = 1$) and characterizes the vehicle orientation. The analytical expression of the dynamics and the camera observations can be easily provided by expressing all the $3D$ vectors as imaginary quaternions. In practice, given a $3D$ vector $\boldsymbol{w} = [w_x,\ w_y,\ w_z]^T$ we associate with it the imaginary quaternion $\hat{w} \equiv 0 + iw_x + jw_y + kw_z$. The dynamics of the state $[\hat{r},\ \hat{v},\ q]^T$ are:

$$\begin{cases} \dot{\hat{r}} = \hat{v} \\ \dot{\hat{v}} = q\hat{A}_v q^* = q\hat{A}q^* + \hat{a}_g \\ \dot{q} = \dfrac{1}{2} q\hat{\Omega} \end{cases} \tag{1}$$

being q^* the conjugate of q, $q^* = q_t - iq_x - jq_y - kq_z$. We now want to express the camera observations in terms of the same state ($[\hat{r},\ \hat{v},\ q]^T$). We remark that the camera provides the direction of the feature in the local frame. In other words, it provides the unit vector $\frac{\boldsymbol{F}}{|\boldsymbol{F}|}$ (see fig. 2). Hence, we can assume that the camera provides the two ratios $y_1 = \frac{F_x}{F_z}$ and $y_2 = \frac{F_y}{F_z}$, being $\boldsymbol{F} = [F_x,\ F_y,\ F_z]^T$. We need to express $\boldsymbol{F}$ in terms of $[\hat{r},\ \hat{v},\ q]^T$. We note that the position of the feature in the frame with the same orientation of the global frame but shifted in such a way that its origin coincides with the one of the local frame is $-\boldsymbol{r}$. Therefore, $\boldsymbol{F}$ is obtained by the quaternion product $\hat{F} = -q^* \hat{r} q$. The observation function provided by the camera is:

$$h_{cam}(\hat{r},\ \hat{v},\ q) = [y_1,\ y_2]^T = \left[\frac{(q^*\hat{r}q)_x}{(q^*\hat{r}q)_z},\ \frac{(q^*\hat{r}q)_y}{(q^*\hat{r}q)_z}\right]^T \tag{2}$$

where the pedices x, y and z indicate respectively the i, j and k component of the corresponding quaternion. We have also to consider the constraint $q^*q = 1$. This can be dealt as a further observation (system output):

$$h_{const}(\hat{r},\ \hat{v},\ q) = q^*q \tag{3}$$

III. Observability Properties

We want to investigate the observability properties of the system whose dynamics are given in (1) and whose observations are given in (2) and (3). The goal of this analysis is to understand how the information contained in the sensor data (from the IMU and the camera) is related to the state $[\hat{r},\ \hat{v},\ q]^T$, which defines our system. In particular, the question we wish to answer by performing this observability analysis is the following. By collecting the data provided by the IMU sensors and by the camera during a given time interval, which are the observable modes that can we estimate? An answer to this question for a general system can be found in [13] where it is shown that these observable modes are all the independent solutions of a system of partial differential equations defined on the space of all the states (in our case this space is a manifold belonging to $\Re^{10}$). In particular, in [13] it is derived a method able to determine these partial differential equations starting from the equations characterizing the system. When this method is used to find the observable modes for the system characterized by the equations (1), (2) and (3) the mentioned system of partial differential equations reduces to a single equation, which is:

$$\begin{aligned} -2r_y\frac{\partial\Lambda}{\partial r_x} + 2r_x\frac{\partial\Lambda}{\partial r_y} - 2v_y\frac{\partial\Lambda}{\partial v_x} + 2v_x\frac{\partial\Lambda}{\partial v_y} + \\ -q_z\frac{\partial\Lambda}{\partial q_t} - q_y\frac{\partial\Lambda}{\partial q_x} + q_x\frac{\partial\Lambda}{\partial q_y} + q_t\frac{\partial\Lambda}{\partial q_z} = 0 \end{aligned} \tag{4}$$

This is a linear partial differential equation. The number of independent solutions $\Lambda = \Lambda(r_x,\ r_y,\ r_z,\ v_x,\ v_y,\ v_z,\ q_t,\ q_x,\ q_y,\ q_z)$ is equal to the number of variables (i.e. 10) minus the number of equations (i.e. 1) [12]. Hence we have 9 independent solutions.

It is immediate to prove that the distance of the feature from the camera, i.e. $|r| \equiv \sqrt{r_x^2 + r_y^2 + r_z^2}$, is a solution of this equation (this can be checked by a simple substitution). This means that the distance of the feature is observable and it is one among the 9 independent solutions. On the other hand, since the camera provides the position of the feature in the local frame up to a scale factor, having the distance means that the feature position in the local frame is also observable. Therefore the three components of the feature position in the local frame are three independent solutions. By using quaternions we can say that three independent solutions are provided by the components of the imaginary quaternion $q^*\hat{r}q$. Furthermore, since the partial differential equation in (4) is invariant under the transformation $r \leftrightarrow v$, three other independent solutions are the components of the imaginary quaternion $q^*\hat{v}q$. Physically, this means that the camera speed in the local frame is also observable. It is possible to easily derive another independent solution. It is q^*q since it is directly

observed (see equation (3); it can be in any case verified that it satisfies (4)). The last two solutions are:

$$Q_r \equiv \frac{q_t q_x + q_y q_z}{1 - 2(q_x^2 + q_y^2)}; \qquad Q_p \equiv q_t q_y - q_z q_x \tag{5}$$

Also for these two solutions it is possible to find a physical meaning. They are related to the two angles: roll and pitch [11]. In particular, the first solution provides the roll angle which is $R = \arctan(2Q_r)$. The latter provides the pitch angle which is $P = \arcsin(2Q_p)$.

The results obtained with this observability analysis are fundamental. We know that by only considering a single point feature and by collecting the data simultaneously from the camera and the IMU during a given time interval we have all the necessary information to perform the estimation of the camera speed and the position of the feature in the local frame, and the roll and the pitch angles describing the orientation of the camera. Furthermore, this observability analysis tells us that the system does not contain any information related to the yaw angle. Finally, in [14] and [15] this observability analysis has been extended in order to detect special trajectories for which the previous estimation cannot be performed. On the other hand, the previous analysis does not provide a method to perform the estimation. This will be discussed in the following.

IV. The Closed-Form Solution for Vehicle Speed and Attitude Determination

We provide a closed form solution which directly expresses the observable modes in terms of the sensor measurements collected during a short time interval. We only consider the case of one feature. The extension to multiple features and to deal with the case of biased inertial measurements is straightforward and is available in [14] and [15]. According to the observability analysis in section III we know that the sensor data collected during a given time interval contain the information to estimate the vehicle speed and the position of the feature in the local frame. Hence, we start by expressing the dynamics and the observation in this frame. We have:

$$\begin{cases} \dot{\boldsymbol{F}} = M\boldsymbol{F} - \boldsymbol{V} \\ \dot{\boldsymbol{V}} = M\boldsymbol{V} + \boldsymbol{A} + \boldsymbol{A_g} \\ \dot{\boldsymbol{q}} = m\boldsymbol{q} \end{cases} \tag{6}$$

where $\boldsymbol{F}$ is the position of the feature in the local frame, $\boldsymbol{V}$ is the vehicle speed in the same frame, $\boldsymbol{A_g}$ is the gravity acceleration in the local frame, i.e. $\hat{A}_g = q^* \hat{a}_g q$, and $\boldsymbol{q}$ is the four vector whose components are the components of the quaternion q, i.e. $\boldsymbol{q} = [q_t,\ q_x,\ q_y,\ q_z]^T$. Finally:

$$m \equiv \frac{1}{2}\begin{bmatrix} 0 & -\Omega_x & -\Omega_y & -\Omega_z \\ \Omega_x & 0 & \Omega_z & -\Omega_y \\ \Omega_y & -\Omega_z & 0 & \Omega_x \\ \Omega_z & \Omega_y & -\Omega_x & 0 \end{bmatrix}$$

$$M \equiv \begin{bmatrix} 0 & \Omega_z & -\Omega_y \\ -\Omega_z & 0 & \Omega_x \\ \Omega_y & -\Omega_x & 0 \end{bmatrix}$$

The validity of (6) can be checked by using $\hat{F} = -q^* \hat{r} q$, $\hat{V} = q^* \hat{v} q$ and by computing their time derivatives with (1). In the local frame, the observation in (2) is:

$$h_{cam} = [y_1,\ y_2]^T = \left[\frac{F_x}{F_z},\ \frac{F_y}{F_z}\right]^T \tag{7}$$

We remark that, because of the gravity, the first two equations in (6) cannot be separated from the equations describing the dynamics of the quaternion. Let us consider a given time interval, $[T_0,\ T_0 + T]$. Our goal is to estimate the observable modes at T_0 (i.e. $\boldsymbol{F_0},\ \boldsymbol{V_0},\ R_0,\ P_0$), by only using the data from the camera and the IMU during the interval $[T_0,\ T_0{+}T]$. We numerically integrate the equations in (6) by leaving symbolic the unknown components of the initial state. On the other hand, the components of $\boldsymbol{q}(T_0)$ are not observable since the yaw angle is not observable. We have the following fundamental property:

Property 1 *The position of the feature at any time, $\boldsymbol{F}(t)$, linearly depends on the initial feature position, $\boldsymbol{F_0}$, on the initial vehicle speed, $\boldsymbol{V_0}$, and on the three quantities: $\chi_\alpha \equiv 2g(q_{t0}q_{y0} - q_{x0}q_{z0})$, $\chi_\beta \equiv -2g(q_{t0}q_{x0} + q_{y0}q_{z0})$ and $\chi_\gamma \equiv 2g(q_{x0}^2 + q_{y0}^2) - g$. In other words:*

$$\boldsymbol{F}(t) = C_F(t)\boldsymbol{F_0} + C_V(t)\boldsymbol{V_0} + C_\chi(t)\boldsymbol{\chi_g} + \boldsymbol{C_B}(t) \tag{8}$$

where $\boldsymbol{\chi_g} \equiv [\chi_\alpha,\ \chi_\beta,\ \chi_\gamma]^T$ and $C_F(t)$, $C_V(t)$, $C_\chi(t)$ are 3×3 matrices and $\boldsymbol{C_B}(t)$ is a $3D-$vector. In addition, $C_F(t)$, $C_V(t)$ and $C_\chi(t)$ only depend on $\boldsymbol{\Omega}(\tau)$, $\tau \in [T_0,\ t]$.

Proof: Before integrating the second equation in (6) we consider the term $\boldsymbol{A_g}$, which depends on the quaternion. In particular, we separate in this term the time-dependent part from the part which is time-independent. Specifically, we introduce the quaternion $p(t)$ such that $q(t) = q_0 p(t)$. $\hat{A}_g(t) = q(t)^* \hat{a}_g q(t) = p(t)^* q_0^* \hat{a}_g q_0 p(t)$. Let us denote with $\boldsymbol{\chi_g}$ the $3D$ vector associated with the quaternion $q_0^* \hat{a}_g q_0$, i.e. $\hat{\chi}_g \equiv q_0^* \hat{a}_g q_0$. Note that $\boldsymbol{\chi_g}$ is the gravity vector in the local frame at the time T_0. By a direct computation we obtain: $\boldsymbol{\chi_g} = [\chi_\alpha,\ \chi_\beta,\ \chi_\gamma]^T$ and:

$$\boldsymbol{A_g}(t) = \Gamma(t)\boldsymbol{\chi_g}, \quad \Gamma(t) \equiv$$

$$\begin{bmatrix} p_t^2 + p_x^2 - p_y^2 - p_z^2 & 2p_t p_z + 2p_x p_y & -2p_t p_y + 2p_z p_x \\ -2p_t p_z + 2p_x p_y & p_t^2 + p_y^2 - p_x^2 - p_z^2 & 2p_t p_x + 2p_z p_y \\ 2p_t p_y + 2p_x p_z & -2p_t p_x + 2p_y p_z & p_t^2 + p_z^2 - p_x^2 - p_y^2 \end{bmatrix}$$

Note that $\Gamma(t)$ only depends on $p(t)$. $p(t)$ is obtained by integrating the equation $\dot{p} = \frac{1}{2} p \hat{\Omega}$, with $p(0) = 1$. Hence, $p(t)$ only depends on the values of the angular speed for $t > T_0$. As a result, the matrix $\Gamma(t)$ only depends on these values. In particular, $\Gamma(t)$ is independent of the initial state. The matrix $\Gamma(t)$ is the rotation matrix transforming vectors from the local frame at time T_0 into local frame at the time t. We integrate the second equation in (6), obtaining:

$$\boldsymbol{V_j} = (I_3 + M_j dt_j)\boldsymbol{V_{j-1}} + \boldsymbol{B_j} dt_j \tag{9}$$

where $\boldsymbol{B_j} = \boldsymbol{A_j} + \boldsymbol{A_g}_{\ j} = \boldsymbol{A_j} + \Gamma_j \boldsymbol{\chi_g}$.

The previous expression for $\boldsymbol{V_j}$ provides the following expression in terms of the initial conditions:

$$\boldsymbol{V_j} = \Xi_j \left[\boldsymbol{V_0} + (t_j - T_0)\boldsymbol{\chi_g} + \sum_{k=1}^{j} \Xi_k^{-1} \boldsymbol{A_k} dt_k \right] \quad (10)$$

with $\Xi_j \equiv \prod_{k=1}^{j}(I_3 + M_k dt_k)$, which coincides with Γ_j since it is the rotation matrix transforming vectors from the local frame at time T_0 into local frame at the time t_j. The expression of $\boldsymbol{F_j}$ in terms of the initial conditions:

$$\boldsymbol{F_j} = \Xi_j \left(\boldsymbol{F_0} - \sum_{k=1}^{j} \Xi_k^{-1} \boldsymbol{V_k} dt_k \right) = \Xi_j \left[\boldsymbol{F_0} + \right. \quad (11)$$

$$\left. -(t_j - T_0)\boldsymbol{V_0} - \frac{(t_j - T_0)^2}{2}\boldsymbol{\chi_g} - \sum_{k=1}^{j}\sum_{k'=1}^{k} \Xi_{k'}^{-1} \boldsymbol{A_{k'}} dt_k dt_{k'} \right]$$

Hence, we have the expression in (8) with: $C_F(t_j) \equiv \Xi_j$, $C_V(t_j) \equiv (T_0 - t_j)\Xi_j$, $C_\chi(t_j) \equiv -\Xi_j \frac{(t_j - T_0)^2}{2}$, $\boldsymbol{C_B}(t_j) \equiv -\Xi_j \sum_{k=1}^{j}\sum_{k'=1}^{k} \Xi_{k'}^{-1} \boldsymbol{A_{k'}} dt_k dt_{k'}$ ■

We consider the components of $\boldsymbol{F}(t)$, i.e. $F_x(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0}, \boldsymbol{\chi_g})$ $F_y(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0}, \boldsymbol{\chi_g})$ and $F_z(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0}, \boldsymbol{\chi_g})$. By using (7) we obtain:

$$F_x(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) = y_1(t)\ F_z(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) \quad (12)$$

$$F_y(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) = y_2(t)\ F_z(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g})$$

i.e., each camera observation occurred at the time $t \in [T_0,\ T_0 + T]$ provides two equations in the nine unknowns (which are the components of $\boldsymbol{F_0}$, $\boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$). On the basis of property 1, the components of $\boldsymbol{F}(t)$ are linear on the unknowns. Hence, the equations in (12) are linear and, by having at least $n_{obs} = 5$ camera observations, we can easily obtain the initial state $[\boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}]^T$. In particular, when $n_{obs} \geq 5$, the components of $\boldsymbol{F_0}$, $\boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$ are obtained by computing the pseudoinverse of a $(2n_{obs} \times 9)$ matrix.

A. Exploiting Additional Information

On the basis of the observability analysis performed in section III, we know that, regarding the robot orientation, only the roll and pitch angles are observable modes. Hence, it must be possible to express the components of the vector $\boldsymbol{\chi_g}$ only in terms of these two angles. In appendix A we provide these expressions. These expressions contain additional information to estimate $[\boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}]^T$. Indeed, the components of $\boldsymbol{\chi_g}$ are three but they only depend on two quantities. An important consequence due to this additional information is that it is possible to estimate $[\boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}]^T$ even when the camera only performs $n_{obs} = 4$ observations. On the other hand, when more than four observations are available ($n_{obs} \geq 5$), the expressions in (20) can be adopted to improve the precision. We discuss the case of $n_{obs} = 4$ observations and we provide a procedure to perform the estimation. When $n_{obs} = 4$, the equations in (12) are eight. Hence, it is not possible to determine the components of $\boldsymbol{F_0}, \boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$ by a simple matrix inversion. On the other hand, the equations in (12) allow us to express eight among the nine unknowns in terms of one of them. Let us suppose to express the components of $\boldsymbol{F_0}$ and $\boldsymbol{V_0}$, and the first two components of $\boldsymbol{\chi_g}$ in terms of χ_γ. We have:

$$\boldsymbol{w} = \boldsymbol{c}\ \chi_\gamma + \boldsymbol{d} \quad (13)$$

where $\boldsymbol{w} = [F_{x0},\ F_{y0},\ F_{z0}, V_{x0},\ V_{y0},\ V_{z0}, \chi_\alpha,\ \chi_\beta]^T$ and $\boldsymbol{c}$ and $\boldsymbol{d}$ are two vectors whose components are obtained by using the eight linear equations provided by (12) where the expression of $\boldsymbol{F}(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g})$ is provided in the proof of property 1. $\boldsymbol{c}$ and $\boldsymbol{d}$ only depend on the vehicle angular speed and linear acceleration during the interval $[T_0,\ T_0 + T]$. We start by considering the last two equations in (13). They are:

$$\chi_\alpha = c_7\chi_\gamma + d_7; \qquad \chi_\beta = c_8\chi_\gamma + d_8$$

where $c_7,\ c_8$ are the 7^{th} and 8^{th} component of $\boldsymbol{c}$ and $d_7,\ d_8$ are the 7^{th} and 8^{th} component of $\boldsymbol{d}$. We know that the norm of the vector $\boldsymbol{\chi_g}$ is g. Hence, by using the previous two expressions we have:

$$|\boldsymbol{\chi_g}|^2 = (c_7\chi_\gamma + d_7)^2 + (c_8\chi_\gamma + d_8)^2 + \chi_\gamma^2 = g^2 \quad (14)$$

which is a second degree equation in χ_γ. Therefore, by solving this equation and by using (13), we immediately obtain $\boldsymbol{w}$, and so $\boldsymbol{F_0}$, $\boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$. Fig. 3 displays the steps of the procedure previously described. In this case of $n_{obs} = 4$ and one single feature the determination of the observable modes is not obtained by a simple matrix inversion or by the computation of a pseudoinverse matrix, as it happens in the other cases (see [14] and [15] where several conditions for the observability are derived, depending on the number of camera poses and features).

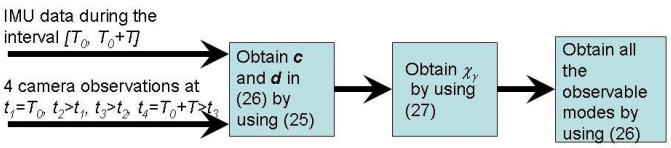

Fig. 3. The steps performed to estimate the observable modes when $n_{obs} = 4$ and in the case of one single feature.

In the case we have $n_{obs} \geq 5$, the values of $\boldsymbol{F_0}, \boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$ are obtained by using the $2n_{obs}(\geq 10)$ equations in (12) (it suffices to compute the pseudoinverse of a $(2n_{obs} \times 9)$ matrix). Then, the equations in (20) are used to obtain the roll and pitch angles. We have:

$$P = \arcsin\left(\frac{\chi_\alpha}{g}\right), \quad R = -\arcsin\left(\frac{\chi_\beta}{\sqrt{g^2 - \chi_\alpha^2}}\right) \quad (15)$$

The previous expressions only depend on χ_α and χ_β. In other words, by using them to estimate the roll and pitch angles, the information contained in χ_γ is not exploited. A possible way to exploit this information is to minimize the cost function:

$$c(R,\ P) = (g\sin P - \chi_\alpha)^2 + \tag{16}$$
$$+ (-g\sin R\cos P - \chi_\beta)^2 + (-g\cos R\cos P - \chi_\gamma)^2$$

where the initial values are set by using (15).

V. Local Decomposition

In this section we want to separate the observable modes from the rest of the system. In control theory this is what it is called a local decomposition of the system [9]. Mathematically, we want to do the following operation. Once the observable modes are stuck in a common vector $\boldsymbol{S} \equiv [\boldsymbol{F},\ \boldsymbol{V}, Q_r,\ Q_p]^T$ we want to express its dynamics (i.e. $\dot{\boldsymbol{S}}$) and the observation function (i.e. the function h_{cam} defined in (2)) only in terms of the components of $\boldsymbol{S}$. This decomposition will allow us to implement an Extended Kalman Filter in order to perform the estimation of the observable modes.

The equations in (6) do not represent a local decomposition. First of all they do not contain the dynamics of the roll and pitch angles. Instead, they contain the dynamics of the quaternion, which is not observable. Additionally, also the second equation depends on the quaternion through the term $\boldsymbol{A_g}$. Instead of considering the two quantities defined in (5), we found easier to characterize the observable part of the vehicle orientation by the following two independent observable modes:

$$m_1 = \frac{q_x^2 + q_y^2}{q_t q_y - q_x q_z} \qquad m_2 = \frac{q_x^2 + q_y^2}{q_t q_x + q_y q_z} \tag{17}$$

which are both solutions of (4). The dynamics of these modes are:

$$\begin{cases} \dot{m}_1 = \Omega_x \dfrac{m_1}{m_2} + \dfrac{\Omega_y}{2}\left(m_1^2 + 1 - \dfrac{m_1^2}{m_2^2}\right) + \Omega_z \dfrac{m_1^2}{m_2} \\ \dot{m}_2 = \dfrac{\Omega_x}{2}\left(m_2^2 - \dfrac{m_2^2}{m_1^2} + 1\right) + \Omega_y \dfrac{m_2}{m_1} - \Omega_z \dfrac{m_2^2}{m_1} \end{cases} \tag{18}$$

In order to complete the system decomposition we have to express the quaternion q, which appears in the second equation of (6) through the term $\boldsymbol{A_g}$, only in terms of m_1 and m_2.

The expression of the quaternion in terms of the roll, pitch and yaw angles is given in A. We remind that $\boldsymbol{A_g}$ is the gravity in the local frame. On the other hand, rotating the gravity around the vertical axis does not affect the expression of $\boldsymbol{A_g}$. For this reason, we set the yaw angle equal to zero. Finally, we express the angles R and P in terms of m_1 and m_2:

$$R = atan\left(\frac{2m_1^2 m_2}{m_1^2 + m_2^2 - m_1^2 m_2^2}\right)$$
$$P = asin\left(\frac{2m_1 m_2^2}{m_1^2 + m_2^2 + m_1^2 m_2^2}\right) \tag{19}$$

The first two equations in (6) together with (18) and the expression of the quaternion in appendix A with $Y = 0$ provide the dynamics of the observable modes. Regarding the observation function given in (2), its expression in terms of $\boldsymbol{F}$ is given in (7). By discretizing the previous equations describing the dynamics of the observable modes and by computing the Jacobians (also of the observation in (7)) it is possible to implement the standard EKF equations to estimate the observable state: $[\boldsymbol{F},\ \boldsymbol{V},\ m_1,\ m_2]^T$.

VI. Performance Evaluation

We evaluate the performance of the strategy based on the form closed solution discussed in section IV. We show some results obtained by using both synthetic and real data. For more results the reader is addressed to [14] and [15]. Finally, we also compare the results achievable by estimating the non observable state $[\boldsymbol{r},\ \boldsymbol{v},\ \boldsymbol{q}]^T$ and by estimating the observable modes (i.e. the state $[\boldsymbol{F},\ \boldsymbol{V},\ m_1,\ m_2]^T$) by using an EKF.

A. Simulated Trajectories

We simulate an aerial vehicle moving along $3D$ trajectories. Each trajectory is generated by generating randomly the linear and angular acceleration at $100\ Hz$. In particular, at each time step, the three components of the linear and the angular acceleration are generated as zero-mean Gaussian independent variables with variance respectively equal to $\left(1\ \frac{m}{s^2}\right)^2$ and $\left(1\ \frac{deg}{s^2}\right)^2$. We adopt many different values for the initial vehicle speed and position. Starting from the accomplished trajectory, the true angular speed and the linear acceleration are computed at each time step of $0.01s$ (respectively, at the time step i, we denote them with $\boldsymbol{\Omega_i^{true}}$ and $\boldsymbol{A_i^{true}}$). Starting from them, the IMU sensors are simulated by generating randomly the angular speed and the linear acceleration at each step according to the following: $\boldsymbol{\Omega_i} = N\left(\boldsymbol{\Omega_i^{true}}, \sigma_\omega^2 I_3\right)$ and $\boldsymbol{A_i} = N\left(\boldsymbol{A_i^{true}}, \sigma_a^2 I_3\right)$ where $N(\boldsymbol{\mu},\ P)$ denotes the normal distribution with mean value $\boldsymbol{\mu}$ and covariance matrix P, I_3 the identity 3×3 matrix and σ_ω and σ_a are respectively set equal to $1\ deg\ s^{-1}$ and $0.1\ m\ s^{-2}$.

Regarding the camera, the provided readings are generated randomly in the following way. By knowing the true trajectory, the true bearing angles of the feature (at the origin) in the camera frame are computed. They are computed each $0.2s$. Then, the camera readings are generated by adding to the true values zero-mean Gaussian errors whose variance is equal to $(1\ deg)^2$ for all the readings.

B. Performance of the Closed-Form Solution

1) Simulation Results: For all the simulations we adopt the closed-form solution introduced in section IV to estimate the distance of the feature ($d \equiv \sqrt{r_x^2 + r_y^2 + r_z^2} = \sqrt{F_x^2 + F_y^2 + F_z^2}$), the speed of the camera ($s \equiv \sqrt{v_x^2 + v_y^2 + v_z^2} = \sqrt{V_x^2 + V_y^2 + V_z^2}$) and the roll and the pitch angles ($R \equiv \arctan(2Q_r)$ and $P \equiv \arcsin(2Q_p)$). Specifically, in all the simulations the values of the estimated $d,\ s,\ R,\ P$ are compared with the ground truth values. The results are obtained by running 100 simulations and the errors provided are averaged on them. Regarding d and s the precision is expressed in terms of the relative error in $\%$ (for instance, if the true value of d is d^t it is computed the difference $\Delta d = |d^t - d|$ and then it is provided $100\frac{\Delta d}{d^t}$).

Regarding the roll and pitch angles it is provided the absolute error in deg. We obtain the following values: $\frac{\Delta d}{d} = 3.2\%$ $\frac{\Delta s}{s} = 2.8\%$ $\Delta R = 0.18\ deg$ $\Delta P = 0.22\ deg$. More results are available in [14] and [15] where the case of biased inertial measurements and the case of multiple features with different camera poses are considered. In particular, significant improvement is found by considering two features.

2) Performance Evaluation with Real Data: We adopted the data set provided in [17]. This is an excellent test bed since it also provides a thorough ground truth. The only drawback for our purposes is that our strategy works also in $3D$ while this data set regards experiments carried out in $2D$.

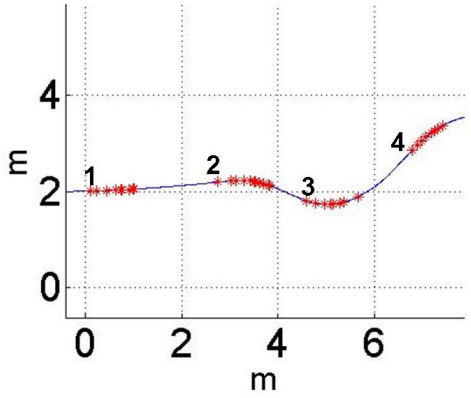

Fig. 4. A piece of the $2D$ robot trajectory as recovered by the ground truth data file in the session Bicocca-2009-02-26a of [17]. The four groups of stars represent the points where the same point feature has been extracted.

We adopted the data provided in the session *Bicocca-2009-02-26a*. The robot trajectory in the ground truth data file is provided at around $50\ Hz$. This allowed us to get a reliable ground truth for the vehicle speed. The data provided by the IMU are also available. These data are delivered at around $130\ Hz$. Finally, by using the provided vision data files, we were able to extract several point features. In fig 4 we display a piece of the robot trajectory (as provided by the ground truth data file). In particular, all the points in blue represent the robot positions. In the figure four groups of points are also displayed by using red stars. Each group of star marks represent the true robot positions where the same point feature has been extracted from the vision data file. Unfortunately, through the provided ground truth data set, we do not have the actual position of our extracted four point features. For this reason, we cannot evaluate the performance of our strategy in evaluating the distance of these features. On the other hand, by having the true robot speed as previously mentioned, we evaluate the accuracy of the proposed strategy in estimating the robot speed. For the four groups of points we obtained by using the proposed approach the following four initial speeds (in ms^{-1}): 0.49, 0.65, 0.57, 0.63, while the true values are: 0.466, 0.638, 0.585, 0.661. An additional real experiment is available in [14] and [15].

C. Comparison of the performances of the two EKFs

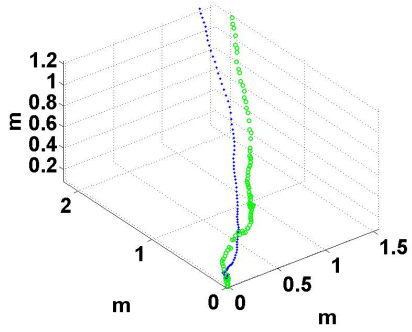

Fig. 5. Results from the EKF which estimates the entire non-observable state. The $3D$ trajectory is displayed. The blue dots indicate the ground truth while the green circles the estimated trajectory.

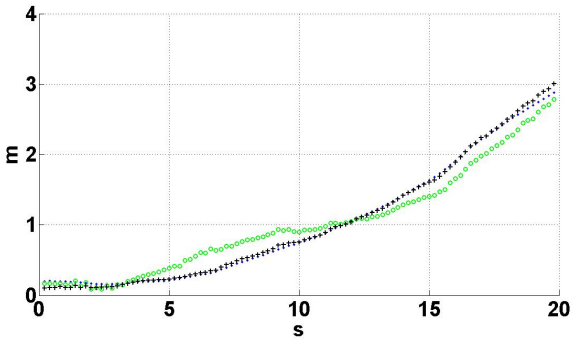

Fig. 6. The true distances from the origin (blue dots), the value of the distances obtained from the values of x, y and z shown in figure 5 (green circles) and the distances estimated by implementing an EKF which directly estimates the observable modes (black crosses).

We also perform the estimation by using two distinct EKFs. The former estimates the non-observable state $[\boldsymbol{r},\ \boldsymbol{v},\ \boldsymbol{q}]^T$. The latter estimates the observable state $[\boldsymbol{F},\ \boldsymbol{V},\ m_1,\ m_2]^T$. Figures 5 and 6 show the typical results we obtained from them. In fig 5 the $3D$ trajectory is displayed. The blue dots indicate the ground truth while the green circles the trajectory estimated by the first EKF. As expected, the estimation is affected by a drift. Fig 6 displays the distance of the vehicle from the origin vs time. Blue dots indicate the true values, green circles the values obtained from the values of x, y and z shown in figure 5 and black crosses indicate the distances estimated by implementing the second EKF.

We performed many simulations obtaining similar results. We conclude with the following remarks:

1) The estimation of the non-observable state ($[\boldsymbol{r},\ \boldsymbol{v},\ \boldsymbol{q}]^T$) is affected by a drift;
2) The observable modes obtained from the estimates of the previous non-observable state seem to be not affected by a drift;
3) The observable modes directly estimated by an EKF are not drift-affected;
4) Directly estimating the observable modes is much more

convenient than obtaining them from the non-observable state in terms of precision.

VII. Conclusions

In this paper we considered the problem of data fusion when the adopted sensors are a monocular camera and inertial sensors (i.e. one tri-axial accelerometer and one tri-axial gyrometer). An observability analysis which accounts the system non linearities allowed us to analytically detect all the observable modes, i.e. all the physical quantities that the information contained in the sensor data allows us to estimate. They are the position of the features in the camera frame, the vehicle speed in the same local frame and the absolute roll and pitch angles. The main contribution of the paper is a new algorithm to simultaneously estimate all the previous physical quantities. In particular, the algorithm is based on a closed-form solution which analytically expresses the vehicle speed and attitude in terms of the sensor measurements. The algorithm only needs to observe four times a single point feature. Compared to the state-of-the-art the advantages are:

1) It only requires a single point feature to work;
2) The estimation does not require any initialization;
3) The estimation can be performed at any moment by only collecting four consecutive observations;
4) It allows eliminating the drift on the scale factor and on the roll and pitch angles independently of the vehicle trajectory;
5) It can be easily extended in order to deal with the case of multiple features and biased inertial measurements ([14] and [15]).

Regarding the extension to deal with biased inertial measurements and multiple features, we show in [14] and [15] that three camera poses and two features allow determining the vehicle speed and attitude together with the scale factor. In order to also determine the bias affecting the inertial measurements, one additional camera observation is required. Note that the algorithm is in general linear, i.e. the determination of the previous quantities only requires a matrix inversion (or the computation of a pseudoinverse). Only in few cases it is necessary to also solve a quadratic polynomial equation (e.g. in the case of one feature and four camera poses (see also [15] where several conditions for the observability are derived, depending on the number of camera poses and features)).

Acknowledgment

This work was supported by the European Project FP7-ICT-2007-3.2.2 Cognitive Systems, Interaction, and Robotics under the contract #231855 (sFLY). We also acknowledge the Rawseeds project for the data sets provided on line [17].

Appendix A
Analytical expression of χ_α, χ_β and χ_γ in terms of the roll and pitch angles

Let us consider the unit quaternion: $q_t + q_x i + q_y j + q_z k$. By denoting with R, P and Y respectively the roll, pitch and yaw angles, we have [11]:

$$q_t = \cos\frac{R}{2}\cos\frac{P}{2}\cos\frac{Y}{2} + \sin\frac{R}{2}\sin\frac{P}{2}\sin\frac{Y}{2}$$

$$q_x = \sin\frac{R}{2}\cos\frac{P}{2}\cos\frac{Y}{2} - \cos\frac{R}{2}\sin\frac{P}{2}\sin\frac{Y}{2}$$

$$q_y = \cos\frac{R}{2}\sin\frac{P}{2}\cos\frac{Y}{2} + \sin\frac{R}{2}\cos\frac{P}{2}\sin\frac{Y}{2}$$

$$q_z = \cos\frac{R}{2}\cos\frac{P}{2}\sin\frac{Y}{2} - \sin\frac{R}{2}\sin\frac{P}{2}\cos\frac{Y}{2}$$

We use the previous expressions to obtain $\chi_\alpha = 2g(q_t q_y - q_x q_z)$, $\chi_\beta = -2g(q_t q_x + q_y q_z)$ and $\chi_\gamma = 2g(q_x^2 + q_y^2) - g$ in terms of the roll, pitch and yaw angles. They only depend on the roll and pitch angles. By a direct substitution we obtain:

$$\chi_\alpha = g\sin P,\ \ \chi_\beta = -g\sin R\cos P,\ \ \chi_\gamma = -g\cos R\cos P \tag{20}$$

References

[1] Ahrens, S.; Levine, D.; Andrews, G.; How, J.P., Vision-based guidance and control of a hovering vehicle in unknown, gps-denied environments, IEEE International Conference on Robotics and Automation (ICRA 2009), Kobe, Japan, May, 2009.

[2] 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 2007 26: 577-589

[3] Blosch, M., Weiss, S., Scaramuzza, D., and Siegwart, R. (2010), Vision Based MAV Navigation in Unknown and Unstructured Environments, IEEE International Conference on Robotics and Automation (ICRA 2010), Anchorage, Alaska, May, 2010.

[4] Bryson, M. and Sukkarieh, S., Building a Robust Implementation of Bearing-only Inertial SLAM for a UAV, JFR, 2007, 24, 113-143

[5] P. Corke, J. Lobo, and J. Dias, An Introduction to Inertial and Visual Sensing, International Journal of Robotics Research 2007 26: 519-535

[6] J. Dias, M. Vinzce, P. Corke, and J. Lobo, Editorial: Special Issue: 2nd Workshop on Integration of Vision and Inertial Sensors, The International Journal of Robotics Research, June 2007; vol. 26, 6: pp. 515-517.

[7] Chris Engels, Henrik Stewnius, David Nistr. Bundle Adjustment Rules. Photogrammetric Computer Vision (PCV), September 2006.

[8] P. Gemeiner, P. Einramhof, and M. Vincze, Simultaneous Motion and Structure Estimation by Fusion of Inertial and Vision Data, The International Journal of Robotics Research 2007 26: 591-605

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

[10] Kim, J. and Sukkarieh, S. Real-time implementation of airborne inertial-SLAM, Robotics and Autonomous Systems, 2007, 55, 62–71

[11] Quaternions and rotation Sequences: a Primer with Applications to Orbits, Aerospace, and Virtual Reality. Kuipers, Jack B., Princeton University Press copyright 1999.

[12] F. John, Partial Differential Equations, Springer-Verlag, 1982.

[13] A. Martinelli, State Estimation Based on the Concept of Continuous Symmetry and Observability Analysis: the Case of Calibration, Accepted for publication on Transaction on Robotics.

[14] A. Martinelli, Deriving and Estimating All the Observable Modes when Fusing Monocular Vision and Inertial Measurements: A Closed Form Solution. Transaction on Robotics (under review).

[15] A. Martinelli, Closed-Form Solutions for Attitude, Speed, Absolute Scale and Bias Determination by Fusing Vision and Inertial Measurements, INRIA technical report.

[16] A.I. Mourikis, N. Trawny, S.I. Roumeliotis, A. Johnson, A. Ansar, and L. Matthies, "Vision-Aided Inertial Navigation for Spacecraft Entry, Descent, and Landing", Transactions on Robotics, 25(2), pp. 264–280

[17] The Rawseeds Project http://www.rawseeds.org/home/

[18] M. Veth, and J. Raquet, Fusing low-cost image and inertial sensors for passive navigation, Journal of the Institute of Navigation, vol. 54(1), 2007