Vision-Aided Inertial Navigation: Closed-Form Determination of Absolute Scale, Speed and Attitude

Agostino Martinelli, Chiara Troiani, Alessandro Renzaglia

Vision-Aided Inertial Navigation: Closed-Form Determination of Absolute Scale, Speed and Attitude

Agostino Martinelli, Chiara Troiani and Alessandro Renzaglia

***Abstract*— This paper investigates the problem of determining the speed and the attitude of a vehicle equipped with a monocular camera and inertial sensors. The vehicle moves in a 3D unknown environment. It is shown that, by collecting the visual and inertial measurements during a very short time interval, it is possible to determine the following physical quantities: the vehicle speed and attitude, the absolute distance of the point features observed by the camera during the considered time interval and the bias affecting the inertial measurements. In particular, this determination, is based on a closed form solution which analytically expresses the previous physical quantities in terms of the sensor measurements. This closed form determination allows performing the overall estimation in a very short time interval and without the need of any initialization or prior knowledge. This is a key advantage since allows eliminating the drift on the absolute scale and on the vehicle orientation. In addition, the paper provides the minimum number of distinct camera images which are needed to perform this determination. Specifically, if the magnitude of the gravity is unknown, at least four camera images are necessary while if it is a priori known, three camera images are necessary. The performance of the proposed approach is evaluated by using real data.**

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. [1], [3]). 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 [7] 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 [19] 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 airborne SLAM, an EKF has been adopted in [11] 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. To the best of our knowledge, no prior work has addressed the problem of determining the trajectory of a platform in closed form, by only using visual and inertial measurements.

Recent works investigate the observability properties of the vision-aided inertial navigation system [8], [10] and [17]. These works show that the absolute roll and pitch angles of the vehicle are observable modes while the yaw angle is unobservable. This result is consistent with the experimental results obtained in [4] which clearly show how the roll and pitch angles remain more consistent than the heading. In [9], the authors provide a theoretical investigation to analytically derive the motion conditions under which the vehicle state is observable. This analysis also includes the conditions under which the parameters describing the transformation camera-IMU are identifiable. On the other hand, a general theoretical investigation able to also derive the minimum number of camera images needed for the state determination still lacks.

In this paper, we focus our attention on the two issues:

- derivation of all the *observable modes*, i.e. the physical quantities that the information contained in the sensor data allows us to determine;
- derivation of closed form solutions to determine all the previous physical quantities.

It is very reasonable to expect that the absolute scale 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

This work was supported by the European Project FP7-ICT-2007-3.2.2 Cognitive Systems, Interaction, and Robotics, contract #231855 (sFLY). We also acknowledge the Autonomous System Lab at ETHZ in Zurich for providing us a $3D$ data set which includes a very reliable ground truth.

A. Martinelli, C. Troiani and A. Renzaglia are with INRIA Rhone Alpes, Montbonnot, France e-mail: agostino.martinelli@ieee.org, chiara.troiani@inria.fr, alessandro.renzaglia@inria.fr

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.

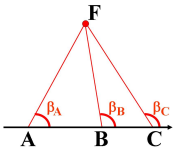

Fig. 1. A robot equipped with an accelerometer and a camera moves on a line. The camera performs three observations of the feature in F, repsectively from the points A, B and C.

The previous example is very simple because of several unrealistic restrictions. First of all, the motion is constrained on a line. Additionally, the accelerometer provides gravity-free and unbiased measurements.

In [15] we relaxed some of the previous restrictions. We considered the case of a robot equipped with IMU and bearing sensors. The motion of the vehicle was not constrained. However, only the case of one single feature was considered. In addition, we assumed unbiased inertial measurements.

In this paper we want to extend the results obtained in [15] by also considering the case of multiple features. To this regard, we will show that, when the number of available features is two, the precision on the estimated quantities increases significantly compared to the case of a single feature. Additionally, also the case when the accelerometers provide biased measurements will be considered. Finally, the experimental validation has been significantly improved with respect to the validation given in [15]. It now includes an experiment conducted in a flying machine arena equipped with a vicon motion capture system.

The paper is organized as follows. Section II provides a mathematical description of the system. Then, in section III we derive the closed-form solution and the algorithm to estimate the vehicle speed and attitude. In section IV we evaluate the performance of the proposed algorithm based on the closed-form solution by using real data. Finally, conclusions are provided in section V.

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 gyroscopes. 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.

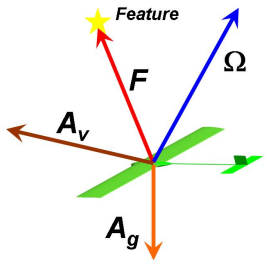

Fig. 2. The feature position ($\boldsymbol{F}$), the vehicle acceleration ($\boldsymbol{A_v}$) the vehicle angular speed ($\boldsymbol{\Omega}$) and the gravity acceleration ($\boldsymbol{A_g}$).

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 [12].

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:

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

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:

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

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):

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

A. The Case with Multiple Features

We consider the case when the camera observes N_f features, simultaneously. We fix the global frame on one of the features. Let us denote with $\boldsymbol{d_i}$ the $3D$ vector which contains the cartesian coordinates of the i^{th} feature ($i = 0,\ 1,\ ...,N_f-1$). We assume that the global frame is attached to the 0^{th} feature, i.e. $\boldsymbol{d_0} = [0\ 0\ 0]^T$. The new system is characterized by the state $[\hat{r},\ \hat{v},\ q,\ \hat{d}_1,\ ...,\ \hat{d}_{N_f-1}]^T$, whose dimension is $7 + 3N_f$. The dynamics of this state are given by (1) together with the equations:

$$\dot{\boldsymbol{d_i}} = [0\ 0\ 0]^T\ i = 1,\ ...,N_f - 1 \quad (4)$$

The position $\boldsymbol{F^i}$ of the i^{th} feature in the local frame is obtained by the quaternion product $\hat{F}^i = q^*(\hat{d}_i - \hat{r})q$. The corresponding observation function is:

$$h^i_{cam} = \left[\frac{(q^*(\hat{d}_i - \hat{r})q)_x}{(q^*(\hat{d}_i - \hat{r})q)_z},\ \frac{(q^*(\hat{d}_i - \hat{r})q)_y}{(q^*(\hat{d}_i - \hat{r})q)_z}\right]^T \quad (5)$$
$$i = 0, 1, ..., N_f - 1$$

which coincides with the observation in (2) when $i = 0$. Summarizing, the case of N_f features is described by the state $[\hat{r},\ \hat{v},\ q,\ \hat{d}_1,\ ...,\ \hat{d}_{N_f-1}]^T$, whose dynamics are given in (1) and (4) and the observations are given in (5) and (3).

B. The Case with Bias

We consider the case when the data provided by the IMU are affected by a bias. In other words, we assume that the measurements provided by the three accelerometers and the three gyroscopes are affected by an error which is not zero-mean. Let us denote with $\boldsymbol{b_A}$ and with $\boldsymbol{b_\Omega}$ the two 3D-vectors whose components are the mean values of the measurement errors from the accelerometers and the gyroscopes, respectively. The two vectors $\boldsymbol{b_A}$ and $\boldsymbol{b_\Omega}$ are time-dependent. However, during a short time interval, it is reasonable to consider them to be constant. Under these hypotheses, the dynamics in (1) become:

$$\begin{cases} \dot{\hat{r}} &= \hat{v} \\ \dot{\hat{v}} &= q\hat{A}_v q^* = q\hat{A}q^* + q\hat{b}_A q^* + \hat{a}_g \\ \dot{q} &= \frac{1}{2}q\hat{\Omega} + \frac{1}{2}q\hat{b}_\Omega \\ \dot{\boldsymbol{b}}_{\boldsymbol{A}} = \dot{\boldsymbol{b}}_{\boldsymbol{\Omega}} &= [0\ 0\ 0]^T \end{cases} \quad (6)$$

Note that the previous equations only hold for short time intervals. In the following, we will use these equations only when this hypothesis is satisfied (in particular, during time intervals allowing the camera to perform at most ten consecutive observations).

III. Closed-Form Solutions to Perform the Estimation of All the Observable Modes

The observability properties of the system defined in section II have been derived in [16]. It has been shown that, the data delivered by the camera and the inertial sensors, contain the information to estimate the following quantities (called *the observable modes*): the position of the features in the local frame, the speed of the vehicle in the same local frame, the biases affecting both the accelerometers and the gyroscopes, the absolute roll and pitch angles. In addition, also the gravity can be estimated (this is in general not necessary since its magnitude is known with high accuracy). This means that all the previous physical quantities can be simultaneously estimated by only using the visual and inertial measurements without the need of any prior knowledge. In this section we provide closed form solutions which directly express these physical quantities in terms of the sensor measurements collected during a short time interval. We start by considering the case without bias.

A. The case without Bias

We express the dynamics and the features observation in the local frame. We have:

$$\begin{bmatrix} \dot{\boldsymbol{F}}^{\boldsymbol{i}} = M\boldsymbol{F}^{\boldsymbol{i}} - \boldsymbol{V} \\ \dot{\boldsymbol{V}} = M\boldsymbol{V} + \boldsymbol{A} + \boldsymbol{A_g} \\ \dot{\boldsymbol{q}} = m\boldsymbol{q} \end{bmatrix} \quad i = 0, 1, ..., N_f - 1 \quad (7)$$

where $\boldsymbol{F^i}$ is the position of the i^{th} feature in the local frame ($i = 0, 1, ..., N_f - 1$), $\boldsymbol{V}$ 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 (7) 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) for the i^{th} feature is:

$$h_{cam} = [y_1^i,\ y_2^i]^T = \begin{bmatrix} \frac{F_x^i}{F_z^i},\ \frac{F_y^i}{F_z^i} \end{bmatrix}^T \tag{8}$$

We remark that, because of the gravity, the first two equations in (7) cannot be separated from the equations describing the dynamics of the quaternion. Let us consider a given time interval, $[T_0,\ T_0 + T]$. Let us denote with R_0 and P_0 the roll and the pitch angles at the time T_0. In addition, let us denote with $\boldsymbol{F_0^i} \equiv \boldsymbol{F^i}(T_0)$ $(i = 0, 1, ..., N_f - 1)$ and $\boldsymbol{V_0} \equiv \boldsymbol{V}(T_0)$. Our goal is to estimate the observable modes at T_0 (i.e. $\boldsymbol{F_0^0},\ \boldsymbol{F_0^1},\ ...,\ \boldsymbol{F_0^{N_f-1}},\ \boldsymbol{V_0},\ R_0,\ P_0$), by only using the data from the camera and the IMU during the interval $[T_0,\ T_0 + T]$. In the following, we will denote with $\boldsymbol{\chi_g}$ the gravity vector in the local frame at time T_0. In other words, $\boldsymbol{\chi_g} \equiv \boldsymbol{A_g}(T_0)$. Note that, estimating $\boldsymbol{\chi_g}$ allows us to estimate the roll and pitch angles (R_0 and P_0). Indeed, from the definition of the roll and pitch angles it is possible to obtain:

$$\boldsymbol{\chi_g} = g[\sin P_0,\ -\sin R_0 \cos P_0,\ -\cos R_0 \cos P_0]^T \tag{9}$$

To derive the closed form solution it is useful to first consider the special case where the vehicle does not rotate during the interval $[T_0,\ T_0 + T]$. In this case, the first two equations in (7) become:

$$\begin{bmatrix} \dot{\boldsymbol{F}}^{\boldsymbol{i}} = -\boldsymbol{V} \\ \dot{\boldsymbol{V}} = \boldsymbol{A} + \boldsymbol{\chi_g} \end{bmatrix} \quad i = 0, 1, ..., N_f - 1 \tag{10}$$

It is immediate to integrate the previous equations and obtain the position of the i^{th} feature in the local frame:

$$\boldsymbol{F^i}(t) = \boldsymbol{F_0^i} - \Delta t \boldsymbol{V_0} - \frac{\Delta t^2}{2} \boldsymbol{\chi_g} - \int_{T_0}^{t} \int_{T_0}^{t'} \boldsymbol{A}(\tau) d\tau dt' \tag{11}$$

where $\boldsymbol{A}(\tau)$ are provided by the accelerometers and $\Delta t \equiv t - T_0$.

Let us now consider a generic motion, namely when the vehicle is not constrained to move with a fixed orientation during the interval $[T_0,\ T_0 + T]$. Let us denote with $\Xi(t)$ the matrix which characterizes the rotation occurred during the interval $[T_0,\ t]$. By using the data from the gyroscopes during this time interval, it is possible to obtain $\Xi(t)$ (see appendix I). Hence, we obtain the extension of (11) to a generic motion. We have:

$$\boldsymbol{F^i}(t) = \Xi(t) \left(\boldsymbol{F_0^i} - \Delta t \boldsymbol{V_0} - \frac{\Delta t^2}{2} \boldsymbol{\chi_g} + \right. \tag{12}$$

$$\left. - \int_{T_0}^{t} \int_{T_0}^{t'} \Xi^{-1}(\tau) \boldsymbol{A}(\tau) d\tau dt' \right) \quad i = 0, 1, ..., N_f - 1$$

In [16] we obtained the same result by directly integrating the equations in (7).

We consider the components of $\boldsymbol{F^i}(t)$, i.e. $F_x^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g})$ $F_y^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g})$ and $F_z^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g})$. By using (8) we obtain:

$$\begin{gathered} F_x^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) = y_1^i(t)\ F_z^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) \\ F_y^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) = y_2^i(t)\ F_z^i(t;\ \boldsymbol{F_0},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}) \\ i = 0, 1, ..., N_f - 1 \end{gathered} \tag{13}$$

i.e., each camera observation occurred at the time $t \in [T_0,\ T_0 + T]$ provides $2N_f$ equations in the $3N_f + 6$ unknowns (which are the components of $\boldsymbol{F_0^i}$ $(i = 0, 1, ..., N_f - 1)$, $\boldsymbol{V_0}$ and $\boldsymbol{\chi_g}$). From the expression in (12), the components of $\boldsymbol{F}(t)$ are linear in the unknowns. Hence, the equations in (13) are linear.

Let us suppose that the camera performs observations from n_{obs} distinct poses. The number of equations provided by (13) is $2n_{obs}N_f$ while the number of unknowns is $3N_f + 6$. On the other hand, we do not know if the previous equations are all independent. To this regard, in [16] we proved the following fundamental result starting from an observability analysis:

Theorem 1 *In order to estimate the observable modes the camera must perform at least three observations (i.e. the observability requires to have at least three images taken from three distinct camera poses). When the magnitude of the gravitational acceleration (g) is unknown, the minimum number of camera images becomes four.*

Proof: The proof of this theorem is provided in [16] ■

From this theorem we know that the number of independent equations is always smaller than the number of unknowns for $n_{obs} \leq 3$. Let us discuss the determination for different values of n_{obs} and N_f.

1) $n_{obs} = 3$: When $N_f = 1$ the number of equations is 6 and the number of unknowns is 9. Hence the estimation cannot be performed. When $N_f \geq 2$ the number of equations is larger or equal to the number of unknowns. On the other hand, according to theorem 1, the vector $\boldsymbol{\chi_g}$ cannot be determined, since its magnitude is not observable. In other words, the equations in (13) are not independent. In III-A.3 we will show that, by using the knowledge of the gravity (i.e. the magnitude of the vector $\boldsymbol{\chi_g}$), it is possible to determine the unknowns by solving a second order polynomial equation. Hence, in this case, two solutions are determined. Note that there are special situations, whose probability of occurrence is zero, where the determination cannot be carry out. For instance, in the case $n_{obs} = 3,\ N_f = 2$, if one of the three camera poses is aligned along with the two features, the determination cannot be performed. Another special case is when the three camera poses and the two features belong to the same plane.

2) $n_{obs} \geq 4$: In this case the equations in (13) are in general independent. On the other hand, when $n_{obs} = 4$ and $N_f = 1$, the number of equations is 8, which is less than the number of unknowns 9. As in the case $n_{obs} = 3$, it is possible to determine the unknowns by solving a second order polynomial equation. Hence, also in this case, two solutions are determined (see section III-A.3 and [16] for further details).

When $n_{obs} \geq 5$ and/or $N_f \geq 2$ the determination can be done by the computation of a pseudoinverse. Hence, a single solution can be obtained. Then, the knowledge of the magnitude of the gravitational acceleration, can be used to improve the precision (see section III-A.3).

3) Exploiting the knowledge of g: The linear system in (13) will be denoted with:

$$\Gamma \boldsymbol{x} = \boldsymbol{\beta} \tag{14}$$

where the vector $\boldsymbol{x}$ contains all the unknowns, i.e. $\boldsymbol{x} \equiv [\boldsymbol{F_0^0},\ \boldsymbol{F_0^1},\ ...,\ \boldsymbol{F_0^{N_f-1}},\ \boldsymbol{V_0},\ \boldsymbol{\chi_g}]^T$. Γ and $\boldsymbol{\beta}$ are respectively a $(2n_{obs}N_f \times 3N_f + 6)$ matrix and a $(2n_{obs}N_f \times 1)$ vector and are obtained as follows. For a camera observation occurred at time t, each feature contributes with two rows to the matrix Γ and with two entries to the vector $\boldsymbol{\beta}$. It is possible in general to compute the pseudoinverse (or the inverse) of the matrix Γ in the following cases:

1) when $n_{obs} \geq 4$ and $N_f \geq 2$;
2) when $n_{obs} \geq 5$ and $N_f = 1$.

When the rank of Γ is one less than the number of its columns, the nullspace of Γ has dimension one. This is in general the case when $n_{obs} = 3,\ N_f \geq 2$ or when $n_{obs} = 4,\ N_f = 1$. In this case, the system in (14) has an infinite number of solutions. By denoting with $\boldsymbol{\nu}$ the unit vector belonging to the nullspace of Γ, with $\boldsymbol{x}_p$ one among the solutions of (14), any solution of (14) is

$$\boldsymbol{x} = \boldsymbol{x}_p + \lambda \boldsymbol{\nu}$$

where λ is a real number. On the other hand, by knowing the magnitude of the gravitational acceleration, it is possible to determine two values of λ. This is obtained by enforcing the constraint that the vector $\boldsymbol{s}_\lambda$ constituted by the last three entries of the solution $\boldsymbol{x}_p + \lambda \boldsymbol{\nu}$ is a vector with norm equal to g. In other words:

$$|\boldsymbol{s}_\lambda|^2 = g^2 \tag{15}$$

which is a second order polynomial equation in λ. Hence, in this case two solutions are determined.

Finally, when Γ is full rank, the knowledge of the magnitude of the gravitational acceleration can be exploited by minimizing the cost function:

$$c(\boldsymbol{x}) = |\Gamma \boldsymbol{x} - \boldsymbol{\beta}|^2 \tag{16}$$

under the constraint $|\boldsymbol{\chi_g}| = g$. This minimization problem can be solved by using the method of Lagrange multipliers.

B. The case with Bias

We derive a closed-form solution only when the accelerometers are affected by a bias, i.e. we will consider the case $\boldsymbol{b_A} \neq [0\ 0\ 0]^T$ and $\boldsymbol{b_\Omega} = [0\ 0\ 0]^T$. The case of having a bias also on the gyroscopes will be considered in a future work.

The expression in (12) can be easily extended to deal with this case by the substitution: $\boldsymbol{A}(\tau) \rightarrow \boldsymbol{A}(\tau) + \boldsymbol{b_A}$. We obtain:

$$\boldsymbol{F^i}(t) = \Xi(t) \left(\boldsymbol{F_0^i} - \Delta t \boldsymbol{V_0} - \frac{\Delta t^2}{2} \boldsymbol{\chi_g} + \right. \tag{17}$$

$$\left. - \int_{T_0}^{t} \int_{T_0}^{t'} \Xi^{-1}(\tau) d\tau dt' \boldsymbol{b_A} - \int_{T_0}^{t} \int_{T_0}^{t'} \Xi^{-1}(\tau) \boldsymbol{A}(\tau) d\tau dt' \right)$$

$$i = 0, 1, ..., N_f - 1$$

By proceeding as in the case without bias we obtain the analogous of equations (13). The new equations also depend on the vector $\boldsymbol{b_A}$.

IV. Performance Evaluation

We evaluated the performance of the proposed algorithm by using a $3D$ data set. These data have been provided by the autonomous system laboratory at ETHZ in Zurich. The data are provided together with a reliable ground-truth, which has been obtained by performing the experiments at the ETH Zurich Flying Machine Arena [14], which is equipped with a Vicon motion capture system. The visual and inertial data are obtained with a monochrome USB-camera gathering 752×480 images at $15Hz$ and a Crossbow VG400CC-200 IMU providing the data at $75\ Hz$. The camera field of view is $150\ deg$. The calibration of the camera was obtained by using the omnidirectional camera toolkit by Scaramuzza [18]. Finally, the extrinsic calibration between the camera and the IMU has been obtained by using the strategy introduced in [13].

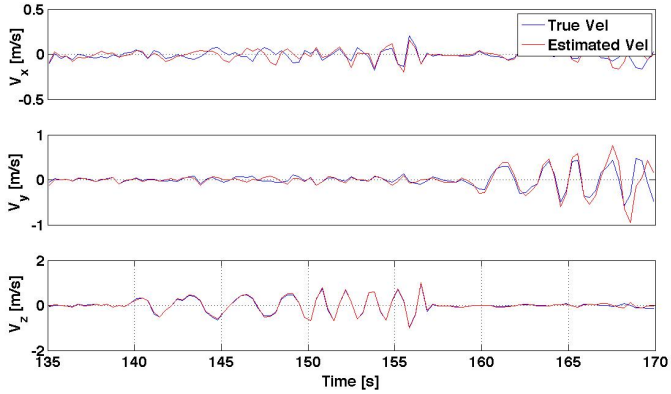

Fig. 3. The three components of the vehicle speed.

Figures 3 and 4 show the results regarding the estimated speed, roll and pitch angles, respectively. In all those figures, the blue lines are the ground truth while the red lines are the estimated values.

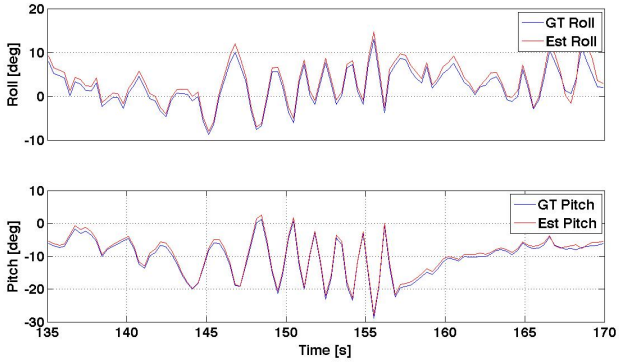

Fig. 4. Roll (up) and pitch (down) angles in the experiment.

V. Conclusions

In this paper we considered the problem of determining the speed and the attitude of a vehicle equipped with a monocular camera and inertial sensors (i.e. three accelerometers and three gyroscopes). The vehicle moves in a 3D unknown environment. It has been shown that, by collecting the visual and inertial measurements during a very short time interval, it is possible to determine the following physical quantities: the vehicle speed and attitude, the absolute distance of the point features observed by the camera during the considered time interval and the bias affecting the inertial measurements. In particular, this determination, is based on a closed form solution which analytically expresses the previous physical quantities in terms of the sensor measurements. This closed form determination 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 absolute scale and on the vehicle orientation. Real experiments validate the proposed approach.

Appendix I
Expression of the rotation matrix Ξ by integrating the angular speed

Let us consider a vehicle and let us refer to a frame attached to this vehicle. When the vehicle performs a motion during the infinitesimal interval of time $[t_j,\ t_j + \delta t]$, the rotation matrix which transforms vectors in the reference before this motion and the reference after this motion is: $I_3 + M_j \delta t$, where I_3 is the 3×3 identity matrix and M_j is the skew-symmetric defined in section III at the time t_j.

Now, let us suppose that the vehicle moves during the interval of time $[t_i,\ t_f]$. In order to compute the rotation matrix which transforms vectors in the reference before this motion and the reference after this motion, we divide the motion in many (N) steps. For each step, the expression of the rotation matrix is the one previously provided. Then, it suffices to compute the product of all these matrices. We obtain:

$$\Xi = \prod_{k=1}^{N} (I_3 + M_k \delta t_k) \tag{18}$$

where $t_1 = t_i$ and $t_N = t_f$.

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.

[7] 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

[8] E. Jones, A. Vedaldi, and S. Soatto, "Inertial Structure From Motion with Autocalibration," in Proc. IEEE Int'l Conf. Computer Vision Workshop on Dynamical Vision, Rio de Janeiro, Brazil, Oct. 2007.

[9] E. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach", The International Journal of Robotics Research, published on-line: January 17, 2011

[10] J. Kelly and G. Sukhatme, Visual-inertial simultaneous localization, mapping and sensor-to-sensor self-calibration, Int. Journal of Robotics Research, Oct. 2010

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

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

[13] J. Lobo and J. Dias, Relative pose calibration between visual and inertial sensors, International Journal of Robotics Research, 26(6):2007, 561-575.

[14] S. Lupashin, A. Schollig, M. Sherback and R. D'Andrea, A simple learning strategy for high-speed quadrocopter multi-flips, IEEE International Conference on Robotics and Automation, Anchorage, 2010

[15] A. Martinelli, Closed-Form Solution for Attitude and Speed Determination by Fusing Monocular Vision and Inertial Sensor Measurements, International Conference on Robotics and Automation, ICRA 2011, Shanghai, China

[16] A. Martinelli, Vision and IMU Data Fusion: Closed-Form Solutions for Attitude, Speed, Absolute Scale and Bias Determination, Transaction on Robotics (under review). INRIA tech rep, http://hal.archives-ouvertes.fr/inria-00569083.

[17] A.I. Mourikis and S.I. Roumeliotis, "A Multi-State Constrained Kalman filter for Vision-aided Inertial Navigation", In Proc. 2007 IEEE International Conference on Robotics and Automation (ICRA'07), Rome, Italy, Apr. 10-14, pp. 3565-3572

[18] D. Scaramuzza, A. Martinelli and R. Siegwart, A toolbox for easy calibrating omnidirectional cameras, IEEE International Conference on Inteligent Robots and Systems, 2006

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