A Unified Framework for Constrained Visual-Inertial Navigation with Guaranteed Convergence

PhD Dissertation by

Francesco Di Corato

Corso di Dottorato in Automatica Robotica e Bioingegneria – Ciclo XXIV

Supervisors:
Dott. Lorenzo Pollini
Prof. Mario Innocenti

UNIVERSITÀ DI PISA

Università degli Studi di Pisa
Dipartimento di Ingegneria dell'Informazione

Printed in Italy.
Pisa, May 2013.

To my Family, Francesca and all the people who believed in me.

Francesco

Acknowledgments

I foremost want to thank my advisors, Lorenzo Pollini and Mario Innocenti, for having trusted me and having given me the opportunity to perform doctoral research. I thank them for the suggestions and guidance. Moreover I would like to thank Andrea Caiti for the honor of making me one of his collaborators. Not least, I want to thank my colleague and friend Vincenzo Calabrò for inspirational friendship, synergy and constructive discussions.

The presented work was partially supported by Northrop Grumman Italia s.p.a. and prof. Stefano Soatto, Vision Lab – University of California, Los Angeles. I whish to thank prof. Soatto for precious inspiration, guidance and support.

Thank you.

Pisa, May 12, 2013.

Abstract

This Thesis focuses on some challenging problems in applied Computer Vision: motion estimation of a vehicle by fusing measurements coming from a low-accuracy Inertial Measurement Unit (IMU) and a Stereo Vision System (SVS), and the robust motion estimation of an object moving in front of a camera by using probabilistic techniques.

In the first problem, a vehicle supposed moving in an unstructured environment is considered. The vehicle is equipped with a stereo vision system and an inertial measurements unit. For the purposes of the work, unstructured environment means that no prior knowledge is available about the scene being observed, nor about the motion. For the goal of sensor fusion, the work relies on the use of epipolar constraints as output maps in a loose-coupling approach of the measurements provided by the two sensor suites. This means that the state vector does not contain any information about the environment and associated keypoints being observed and its dimension is kept constant along the whole estimation task. The observability analysis is proposed in order to define the asymptotic convergence properties of the parameter estimates and the motion requirements for full observability of the system. It will be shown

that the existing techniques of visual-inertial navigation that rely on (features-based) visual constraints can be unified under such convergence properties. Simulations and experimental results are summarized that confirm the theoretical conclusions.

In the second problem, the motion estimation algorithm takes advantage from the knowledge of the geometry of the tracked object. Similar problems are encountered for example in the framework of autonomous formation flight and aerial refueling, relative localization with respect to known objects and/or patterns, and so on. The problem is challenged with respect to the classical literature, because it is assumed that the system does not know a priori the association between measurements and projections of the visible parts of the object and reformulates the problem (usually solved via algebraic techniques or iterative optimizations) into a stochastic nonlinear filtering framework. The system is designed to be robust with respect to outliers contamination in the data and object occlusions. The approach is demonstrated with the problem of hand palm pose estimation and motion tracking during reach-and-grasp operations and the related results are presented.

Contents

List of Figures

Chapter 1

Preface

1.1 Motivation and related works

Inertial navigation suffers from drifts due to several factors, in particular inertial sensor errors. Usually, aiding sensors like GPS, air data sensors or Doppler velocity loggers are used to provide corrections to the navigation system. A viable alternative to these systems is the adoption of a vision system that estimates motion of the camera, assumed rigidly attached to the body, given a stream of successive images and image features tracked over time. Navigation via fusion of visual and inertial data is perhaps the most straightforward *inspired-by-Nature* approach, having direct evidences in daily living.

This work follows a number of other attempts present in the literature to build a combined vision-inertial navigation system. The most relevant and recent works in the field of the vision-aided inertial navigation differ mainly by the approach used to integrate the visual measurements and the inertial mechanization equations. A family of solutions is based on the so called tight-coupling approach, see for example [16]: each collected image feature is added to the navigation filter state and cooperates to the estimation phase. In general this kind of approaches ensures the best accuracy but employs a prominent software infrastructure to manage features and estimation refinement systems, for example loop closure. On the other hand, approaches exist that rely on loose-coupling to perform motion estimation; some examples are [20, 4, 29, 27]. In the loosely-coupled structure, the vision system is usually used at an higher level, as relative pose estimator. In particular, in [20] the vision system acts as a visual odometer [31] with the IMU used as an attitude aid to correct the direction of integration of the visual odometry module. On the other extreme [4, 41, 28], the IMU is used as the main navigation sensor and a stereo camera pose estimation scheme serves for mitigation of drifts. The correction step is thus made by feeding to the filter the relative camera pose estimation between two successive time instants, in terms of angular parametrization and translation. In general, this approach is very simple to implement, however its reliability totally relies on the accuracy of the pose estimation algorithm, for which several techniques were developed for improving robustness, see for example [5, 24, 44, 43]. Actually, the *pseudo-measurement* of camera pose between successive frames is generally obtained via iterative nonlinear techniques and determining precisely how the image noise (assumed Gaussian with acceptable approximation) combines in the pose estimation is practically infeasible. In addition, the uncertainty (the noise) related to the pose estimation cannot be considered normally distributed, breaking the basic assumptions of the Kalman Filtering.

In this work, we focus on motion estimation of a vehicle by fusing measurements coming from an Inertial Measurement Unit (IMU) and a Stereo Vision System (SVS). Although these are the only sensors mentioned in the work, the approach can be easily extended to host auxiliary measurements coming from any other kind of sensor. The vehicle is supposed moving in an unstructured environment, meaning that no prior knowledge is available about the scene being observed, nor about the motion. For the goal of sensor fusion, we rely on a *implicit constraints-based* loose-coupling of the measurements provided by the two sensor suites, meaning that: i) epipolar constraints are constructed on tracked features and used as output maps; this formulation requires the use of implicit functions to define the system output. ii) The state vector does not contain any information about the environment and associated keypoints being observed and its dimension is kept constant along the whole estimation task.

1.2 Visual constraints

During the last decade, a certain number of works in the field of loosely-coupled visual-inertial navigation went beyond the principle of visual update based on the concept of *pose* and recast the problem into a *geometric* framework, showing that viewing a group of static features from multiple camera poses had the result to impose geometric constraints involving all the camera poses. In this framework, the vision module is taken at its *lowest* level, i.e. in terms of image features. Every feature is viewed as an individual entity (taking inspiration from the works on tight-coupling) and individually cooperates for the update step. However the basic idea of loose-coupling is kept, thus the state vector contains the motion parameters only. This led to an large reduction of the computational burden and of the estimator structure, moreover the Gaussian nature of output noise is not altered, since image features are employed. The works by Mourikis [27] and Diel [6] are the main two examples. The general idea is that each feature contributes for a constraint along one direction, leading to a fully constrained problem, when multiple features from different viewpoints are observed. Each single constraint is built on the image projections of the same point in space, corresponding to a couple of camera poses at two different time instants, and on the group transformation relating these two poses. In this work we restrict our attention to the case when only *opportunistic features* are observed, that is image projections of points which position in the space is unknown. The discussion will be dedicated to the *features-based* visual constraints case, in which every feature is taken per-se and has an output associated to it. With the above assumptions, a general formulation for the single constraint can be the following:

$$\phi\left(g_{\tau t},\, y_{\tau}^{i},\, y_{t}^{i}\right) = 0 \tag{1.1}$$

where we highlighted the group transformation $g_{\tau t} \in SE\left(3\right)$ relating two different poses of the viewer in two different time instants τ and $t > \tau$, and the i-the image feature at the corresponding times y_{τ}^{i}, y_{t}^{i}. The two principal geometric visual transformations, used for motion estimation purposes, the epipolar constraints (see [24] and Chapter 3) and the image-space projection of a point in space, via the projective operator (see [24, 27]), fall under this class. In the former case, the constraint is a native implicit function. In the latter case, the transformation is written in explicit form. However it can be rewritten in a pseudo-implicit form: given the position of a $3D$ point X_0^i in space, it does suffice to define the function as [40]

$$\phi\left(g_{0t},\, y_0^i,\, y_t^i\right) = y_t^i - \pi\left(g_{0t} y_0^i Z_0^i\right) = 0 \tag{1.2}$$

being π the perspective projection operator and $X_0^i = y_0^i Z_0^i$. The function ϕ has a certain number of properties that strictly depend on the class the adopted function belongs to. Obviously, by means of such properties, each class of functions has its peculiar direction along which the constraint acts. For instance the projective operator induces constraints that lie on the projective space $\mathbb{R}P^2$ [24], while epipolar constraints depend on transformations that lie on the *Essential Manifold*, and will constrain the motion of the viewer along the direction perpendicular to the epipolar plane. We will discuss further about this manifold in Chapter 3. We will omit further details, which were extensively analyzed in the literature, for example [24, 40]. Here we aim at characterizing this class of problems in terms of common convergence properties, regardless of the visual constraint adopted, provided that just opportunistic features are used (i.e. no landmarks are available). It will be shown that, based on such common properties, the two kinds of approaches can be *unified*.

Designing an optimal filter able to process implicit measurements falls in the realm of *Implicit Filtering*, that allows to use algebraic constraint equations as output maps; this idea was already explored in the framework of Vision-Only ego-motion estimation [39]. To the best of our knowledge, the work [6] falls in a class similar to the one described in this Thesis, except that monocular vision is used and simultaneous multi-frame constraints, in order to disambiguate the scale. The authors formulate a *their own* version of the epipolar constraints and employ a state covariance matrix approximation in order to deal with multiple groups of features together; filtering is done via Bayes' Least Squares. The works [34, 47] the epipolar constraints are fused with the kinematic model of an airplane and the filtering is made by employing the same Implicit Filtering technique as in[39] and the present Thesis. In all these three references no observability study is presented.

1.3 Contributions of the work

One contribution of the Thesis is the analytical characterization of the observability of the unknown motion variables, together with the biases of inertial sensors and the gravity, for the class of constraints-based loosely-coupled navigation problems like the one proposed. The observability properties are ensured under a condition defined *rich enough* motion, namely persistence of excitation (Section 4). The motion requirements for motion observability are made explicit and formalized. As already outlined, one intent of the work is to unify the existing approaches focused on loosely-coupled visual-inertial navigation, relying on visual constraints, under the same convergence properties, even if the specific problem of employing epipolar constraints is proposed throughout the work. Actually, these approaches can be generalized under the same category, even if they look at the problem apparently from different standpoints. Moreover the work faces the specific problem of navigation of ground vehicles from a practical point of view and, starting from the convergence properties, defines the countermeasures needed in order to let the navigation algorithm work even in the motion conditions typical of road vehicles.

Finally, in the last part of the work (Part II) a robust model-based pose estimation scheme is presented, able to estimate the relative motion – in terms of position, attitude and velocity – of a monocular vision system with respect to a tracked object of known geometry. It will be assumed that some markers are placed onto the object surface at known positions with respect to the object reference frame. Similar problems are encountered for example in the frameworks of autonomous formation flight and aerial refueling, relative localization with respect to known objects and/or patterns, and so on. The proposed algorithm reformulates the problem (usually solved via algebraic techniques or iterative optimizations) into a stochastic nonlinear filtering framework. It will be shown that it is robust with respect to outliers contamination of the visual data, marker disappearing and reappearing on the image plane and marker overlapping. The technique is able to recognize automatically less probable measurements, ban them from estimation and the estimation problem can still be solved even if a very low number of features (that would be non sufficient for standard algebraic algorithms) is observed. Moreover it is able to adaptively associate a given image measurement to a certain marker or to an outlier by using probabilistic techniques, thus it is totally self-contained and requires a very rough and fast detection phase, i.e. the prior association of a certain measurement is not needed to make the algorithm work.

1.4 Organization of the Thesis

The Thesis is organized as follows. Chapter 2 introduces the models employed for Ego-Motion estimation, in the framework of loosely-coupled visual-inertial navigation. Starting from the modeling of inertial navigation, a derivation of the approximated version of the mechanization equations, written in "local" form is proposed. Finally, the models suitable for estimation of Ego-Motion are derived. Chapter 3 addresses the formalization of the visual model employed, in the form of implicit constraints computed on tracked features and on the estimated system pose; the constraints, together with the local approximation of the inertial mechanization equations, define the full model used for estimation. Chapter 4 discusses the convergence properties of the proposed approach; the analytical characterization of the unobservable space in the class of constraints-based loosely-coupled problems is then addressed. Chapter 5 introduces the algorithms for fusing inertial measurements with visual constraints, in order to solve the Ego-Motion estimation problem. The chapter starts by introducing the iterative schemes for the optimal fusion of measurements, in the form of visual nonlinear equality constraints. Thus it will show how to incorporate the visual constraints in the state estimation problem, by using the Constrained Extended Kalman Filter algorithm. Experimental results performed outdoor are presented in Chapter 6. The second part of the work, exposed in Chapter 7, is dedicated to description of the robust pose estimation scheme. The approach is demonstrated with the problem of hand palm pose estimation and motion tracking during reach-and-grasp operations and the related results are presented.

Part I

Loosely Coupled Visual Motion Estimation

Chapter 2

Modeling Ego-Motion

This section introduces the models employed for Ego-Motion estimation, in the framework of loosely-coupled visual-inertial navigation. The first issue addressed is the modeling of inertial navigation, the reference frames adopted and the basic mechanization equations. The assumptions generally omitted in the literature of vision-aided navigation are made explicit, ending up with a derivation of the approximated version of the mechanization equations, written in "local" form. Finally, the models suitable for estimation of Ego-Motion are derived.

2.1 General framework for inertial navigation

The navigation equations are a set of nonlinear differential equations relating vehicle's Attitude, Velocity and Position to known/measured inertial quantities. In the general theory of inertial navigation, the equations are integrated given the measurements of inertial sensors, accelerometers (f_{ib}^{b}) and gyroscopes (ω_{ib}^{b}), which usually represent the inputs of the navigation system. The inertial mechanization state variables can be defined as the angular parametrization Θ of the Direction Cosine Matrix $R_b^n = R_b^n(\Theta)$, which rotates from body (b) to navigation (n) frames, the velocity vector $V^n = \left[\begin{array}{ccc} V_n & V_e & V_d \end{array} \right]^T$, expressed in navigation frame, and the position vector $r^e = \left[\begin{array}{ccc} \varphi & \lambda & h \end{array} \right]^T$, composed of the latitude, longitude and altitude of the navigation frame with respect to an Earth-fixed frame (e). Any navigation and Earth-fixed frames can be used. In this work we adopted the NED and ECEF reference frames [35]. Without loss of generality, we assume the body frame to be coincident with the IMU.

Following these assumptions, the continuous-time navigation equations resolved in the NED frame have the following form:

$$\begin{aligned}
\dot{\varphi} &= \frac{V_n}{R_m + h} && (2.1)\\
\dot{\lambda} &= \frac{V_e}{(R_n + h)\cos\varphi} && (2.2)\\
\dot{h} &= -V_d && (2.3)\\
\dot{V}^n &= R_b^n\, f_{ib}^b - (2\,\omega_{ie}^n + \omega_{en}^n) \wedge V^n + \gamma^n(\varphi) && (2.4)\\
\dot{R}_b^n &= R_b^n\,\left(\omega_{ib}^b - R_n^b\,\omega_{in}^n\right) \wedge && (2.5)
\end{aligned}$$

where ω_{in}^n is usually denoted as the *transport rate*, which can be computed as:

$$\omega_{in}^n = \omega_{ie}^n + \omega_{en}^n \qquad (2.6)$$

that is as the summation between the NED frame angular velocity (ω_{en}^{n}) and the Earth rotation rate (ω_{ie}^{n}), projected onto the axes of the navigation frame[1]. Those two terms are included into the navigation equations (2.4) also, to account for the Coriolis and centripetal acceleration effects. The term $\gamma^{n}(\varphi)$ denotes the local gravity acceleration, aligned with the vertical axis of the navigation frame, $\gamma^{n}(\varphi) = \begin{bmatrix} 0 & 0 & \gamma_{local}(\varphi) \end{bmatrix}^{\mathrm{T}}$. Note that the navigation equations depend on some local constants which are the Earth *WGS84* Datum constants, such as the local *Normal* (R_n) and *Meridian* (R_m) Earth radii of curvature, together with $\|\omega_{ie}^{n}\|$ and the local value of the gravitational acceleration, $\gamma_{local}(\varphi)$. Full derivation of the above equations and the detailed descriptions of the model local constants can be found in several textbooks and is omitted here (see, for instance, [35]).

2.2 Modeling *local* inertial navigation

The foregoing equations (2.1)-(2.5) are written in a *global* form, meaning that they are valid with a sufficient degree of accuracy everywhere on the Earth surface and for navigation tasks over long time periods, along several (hundreds or thousands) kilometers. On the opposite, the framework of vision-aided navigation, is usually assumed in the literature to be a *local* navigation problem (see for example [4, 16, 20]) meaning that the navigation task is performed

[1] The angular velocity ω_{en}^{n} can be defined as such velocity needed to make the navigation frame constantly aligned with the Geodetic *North-East-Down* configuration, while the body travels on the Earth surface

with respect to a reference position (usually the starting position of the vehicle) and the relative displacement with respect to the starting point, over the whole video stream, is *small enough*. This allows to make some approximations. Usually, however, details and drawbacks of such approximations are generally omitted in most visual-inertial navigation works, thus represent hidden assumptions that are not verified for later. Here we will make them explicit for completeness.

The first approximation arises when using low-cost inertial sensors, characterized by a significant level of noise in the measurements. This allows to neglect the Earth rotation rate from equations (2.4) and (2.5), as it can be understood by looking at Figures 2.1 and 2.2. The figures show an inertial data set collected during an outdoor experiment. The gyroscope and accelerometer streams were collected with the IMU in a static configuration, on the top of a car with engine on. The signals were detrended in order to isolate the noise component. As it can be noticed, the level of noise is far beyond the Earth-induced velocity effects: it is reasonable to think that this would have a comparable effects on the signal integration with the angular and velocity random walks induced by signal noise only. When the sensors bias come into play, the Earth-induced effect would be negligible.

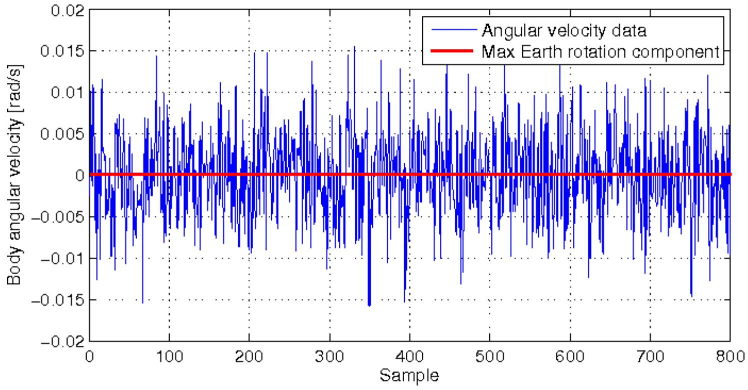


Figure 2.1: *Comparison between gyroscope output and Earth rotation rate. The shown Earth rotation correspond to the component with the maximum value of the rotation vector in the NED frame, computed at the reference latitude of* 43.720677 deg.

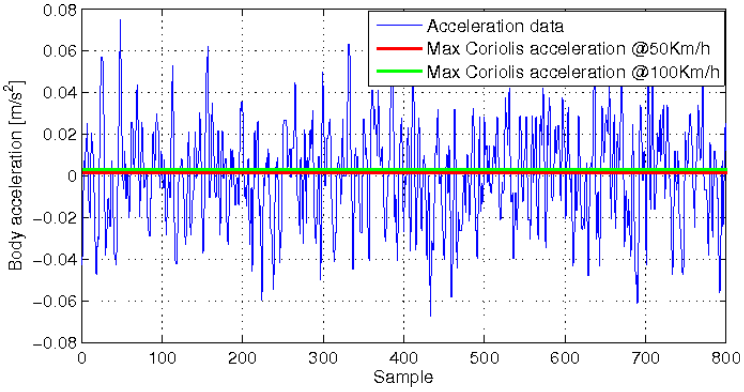


Figure 2.2: *Comparison between accelerometer output and the component with the maximum value of the Coriolis acceleration* $a_c^n = 2\omega_{ie}^n \wedge V^n$. *The Earth rotation vector was computed at the reference latitude of* 43.720677 deg. *The vehicle was supposed moving on the N-E plane, with equal velocity in the two directions.*

The most important assumption generally made is that the navigation frame (NED) is considered not to change its orientation with respect to the ECEF frame, during the whole navigation task. This means that the Earth is approximated as a flat surface in the neighborhood of the starting point. On the other side, this allows to neglect the term ω_{en}^{n} in equations (2.4) and (2.5), being approximately null. By a formal point of view, the assumption of flat Earth surface is equivalent to project a subspace of the *global* navigation equations, relative to the ECEF position, onto a *tangent space* to the manifold of Earth ellipsoid at a given point. This can be made by choosing a specific projection map $\xi : \mathbb{R}\mathrm{E}^2 \rightarrow T_{r_0}\mathbb{R}\mathrm{E}^2$ from the space of ellipsoidal coordinates $(\varphi, \lambda) \in \mathbb{R}\mathrm{E}^2$ to the space of local coordinates $(x^n, y^n) \in T_{r_0}\mathbb{R}\mathrm{E}^2$ in the tangent space. $T_{r_0}\mathbb{R}\mathrm{E}^2$ denotes the tangent space to the manifold represented by the Earth ellipsoid at the point (on the Earth surface)

$$\rho_0 = \begin{bmatrix} R_n \cos\varphi_0 \cos\lambda_0 \\ R_n \cos\varphi_0 \sin\lambda_0 \\ R_n \left(1 - \epsilon^2\right) \sin\varphi_0 \end{bmatrix} \tag{2.7}$$

corresponding to the coordinates $r_0 = \begin{bmatrix} \varphi_0 & \lambda_0 \end{bmatrix}^T$. In the previous equation, R_n denotes the radius of the curvature normal to the ellipsoid surface, at the tangent point ρ_0, while ϵ is the ellipsoid eccentricity [35], according to the *WGS84* model. Suppose now to put the NED reference frame on the Earth surface at the location ρ_0. Two convenient differential (unnormalized) directions on the space $T_{r_0}\mathbb{R}\mathrm{E}^2$ pointing respectively toward North and East can be easily

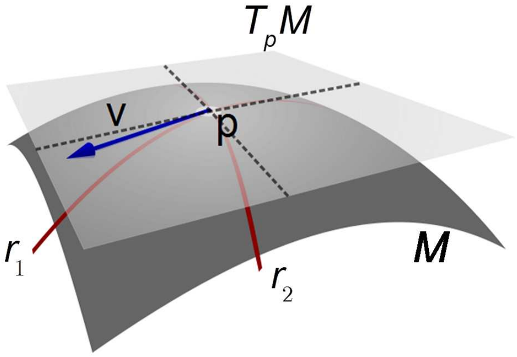


Figure 2.3: *Representation of the tangent space T_pM to the 2-dimensional manifold M at a given point p and the corresponding tangent vector $v \in T_pM$.*

found to be:

$$dX^n = R_m d\varphi \tag{2.8}$$

$$dY^n = R_n \cos\varphi_0 d\lambda \tag{2.9}$$

Thus, given a certain ECEF position $r^e = \begin{bmatrix} \varphi & \lambda & h \end{bmatrix}$ in the neighborhood of the point r_0, the position of the vehicle with respect to the *local* NED reference frame can be obtained as:

$$T^n = \begin{bmatrix} \xi\left(r^e\right) \\ -h \end{bmatrix} = \begin{bmatrix} R_m\left(\varphi - \varphi_0\right) \\ R_n \cos\varphi_0\left(\lambda - \lambda_0\right) \\ -h \end{bmatrix} \tag{2.10}$$

The point $r_0 = \begin{bmatrix} \varphi_0 & \lambda_0 \end{bmatrix}^T$ is usually defined as the position of the vehicle, in latitude and longitude, when the navigation task began its execution (at time t_0). Note the *minus* sign next to the vertical displacement, which is useful to express such coordinate with respect to the local reference system in the North-East-Down configuration. Taking the derivative of equation (2.10) with respect to time, evaluated locally at the point r_0, Equations (2.1)-(2.3) can be simply transformed in local coordinates as:

$$\dot{T}^n = V^n \tag{2.11}$$

Projecting the navigation equations onto a local tangent plane, has the additional advantage that the gravity field can be considered constant in modulus, in the neighborhood of the reference position r_0. This allows to drop the dependence from the current latitude φ, as in Equation (2.4), and to substitute the gravity acceleration term with the constant value $\gamma^n = \gamma^n\left(\varphi_0\right)$.

According to the previous assumptions, the inertial navigation model can be rewritten in local coordinates as:

$$\begin{cases} \dot{T}^n = V^n \\ \dot{V}^n = R_b^n \, f_{ib}^b + \gamma^n = a^n \\ \dot{R}_b^n = R_b^n \, \omega_{ib}^b \wedge \end{cases} \tag{2.12}$$

2.2.1 Models for Ego-Motion estimation

The previous section showed how to *localize* the navigation equations such that they can be used in approximated form in problems where the navigation task happens in a restricted area. This was necessary since the visual-inertial navigation problem is local by definition and it was convenient to recall the formal connection between the classical art of inertial navigation and particular navigation problems as the one this work deals with. Reducing the complexity of the navigation equations has the further advantage that the notation can be simplified, by dropping the subscripts/superscripts from equations, where the symbols are easy to disambiguate. Equation (2.12), in particular, is written in a common reference frame, exception made for the inertial measurements f_{ib}^{b}, ω_{ib}^{b}, which are referred to the body reference frame. For this reason it is convenient to introduce a more compact notation which will be largely adopted in the rest of the Thesis.

Notation. The remaining exposition relies on a simplified notation, very common in the Computer Vision and Robotics community [30]: the generic pose (rotation R_i^j and translation T_i^j) of the frame $\mathcal{I}$ with respect to the frame $\mathcal{J}$ is denoted with the group transformation $g_{ij} = \left\{R_i^j,\, T_i^j\right\} \in SE\,(3)$, which maps a vector expressed in the frame $\mathcal{I}$, into a vector expressed in the frame $\mathcal{J}$. The sole exception is made for the pose of the body frame with respect to the fixed navigation frame, $g_{bn} = \{R_b^n,\, T^n\}$, for which we drop the subscripts/superscripts, for cleaner notation, and it is denoted simply as $g = \{R,\, T\}$. The inverse transformation is indicated with the notation $g_{ij}^{-1} \triangleq \left\{{R_i^j}^T,\, -{R_i^j}^T T_i^j\right\} \in SE\,(3)$. The *action* of the group transformation g_{jk} on g_{ij}, usually denoted with the symbol $\circ$, to indicate function composition, is indicated with a simple product, i.e. $g_{ik} = g_{jk} g_{ij} \triangleq g_{jk} \circ g_{ij}$, being by definition: $g_{ik} \triangleq \left\{R_j^k R_i^j,\, R_j^k T_i^j + T_j^k\right\}$. Finally the action of *scaling* by a certain amount α is defined as: $\alpha g_i^j \triangleq \left\{R_i^j,\, \alpha T_i^j\right\}$. Finally, it is convenient to introduce the time dependence on such transformations and variables which are not constant, in general, and are assumed to vary over time. The transformations and variables for which the time index is dropped will be considered not to change over time.

With the introduced notation, the kinematic model (2.12) can be rewritten as:

$$\begin{cases} \dot{T}\left(t\right) = v\left(t\right) \\ \dot{v}\left(t\right) = a\left(t\right) \\ \dot{R}\left(t\right) = R\left(t\right)\Omega\left(t\right) \end{cases} \tag{2.13}$$

This model keeps the same information as the one in equation (2.12): $T\left(t\right), v\left(t\right)$ and $R\left(T\right)$ are respectively the position, linear velocity and rotation matrix of the *body* frame with respect to the local navigation frame. $\Omega\left(t\right) = \omega\left(t\right)\wedge$ is the skew symmetric matrix of the body angular velocity $\omega\left(t\right)$ expressed in the body frame. Finally $a\left(t\right)$ is the body acceleration expressed in the local navigation frame, which depends on the inertial acceleration sensed by the accelerometers and on the gravity. The *pose* variables $T\left(t\right)$ and $R\left(T\right)$ can be put together to define the group transformation $g\left(t\right) \triangleq \left\{R\left(t\right), T\left(t\right)\right\} \in SE\left(3\right)$.

The inertial measurements, linear accelerations and angular velocities, can be considered as *outputs* of the system, rather than inputs, like in the classical inertial navigation practice[2], see also [16]. The reason behind this is above all *philosophical*, in the sense that we look at the problem from a stochastic filtering point of view [15], thus treating the IMU outputs as *measurements* depending on system states, linear accelerations in NED frame and angular velocities in body frame. Without loss of generality, this is in line with some

[2] This, in turn, will simplify the proof of observability, as proven later by Lemma 1.

of the most recent works in visual-inertial navigation, see for example [16]. The output model relative to the inertial measurements, $y_b\left(t\right)$, can be thus written as:

$$y_b\left(t\right)=\left[\begin{array}{c}R^T\left(t\right)\left(a\left(t\right)+\gamma\right)\\ \omega\left(t\right)\end{array}\right] \tag{2.14}$$

In order to make the linear accelerations and angular velocities depend on the system states, it is necessary to augment the model of the system, with six more states, that is: three states for the system acceleration $a\left(t\right)$, resolved in the navigation frame and three states for the body angular velocity $\omega\left(t\right)$. Since we assume not to have a prior information regarding the nature of the system motion, the local accelerations and body angular velocities can be modeled as random walks. Moreover, we decided to follow the approach of [16, 18] (for example) and deal with the gravity by adding three more states to the state vector, corresponding to the gravity state variable. As it will be clear with the observability analysis, this choice is convenient in the case of non-knowledge of the initial system attitude or, equivalently, when dealing with non calibrated IMUs. The complete model can be written as:

$$\begin{cases} \dot{T}(t) = v(t) \\ \dot{v}(t) = a(t) \\ \dot{a}(t) = \eta_a(t) \\ \dot{R}(t) = R(t)\,\Omega(t) \\ \Omega(t) = \omega(t)\wedge \\ \dot{\omega}(t) = \eta_\omega(t) \\ \dot{\gamma}(t) = 0 \end{cases} \tag{2.15}$$

$$y_b(t) = \left[\begin{array}{c} R^T(t)\,(a(t) + \gamma(t)) + b_a + \nu_a(t) \\ \omega(t) + b_\omega + \nu_\omega(t) \end{array} \right] \tag{2.16}$$

In this case the gravity term was written dependent on time since it is part of the state space and thus admits a certain time evolution. Moreover, the output model $y_b(t)$ was written with the uncertainties affecting the inertial measurements: the variables b_a, b_ω model the slowly varying biases of the accelerometers and gyroscopes, respectively, and $\nu_a(t) \sim \mathcal{N}(0, R_a)$ and $\nu_\omega(t) \sim \mathcal{N}(0, R_\omega)$ are zero-mean white noises with constant variance, that model the noise in the measurements. In this case, the biases are assumed known (i.e. we are assuming a calibrated IMU). Alternatively, if the calibration parameters of the inertial sensors are unknown, we may choose to insert them into the estimation process. In such a case, we get an extended model with six more states, corresponding to the state variables of the inertial biases:

$$\begin{cases} \dot{T}(t) = v(t) \\ \dot{v}(t) = a(t) \\ \dot{a}(t) = \eta_a(t) \\ \dot{R}(t) = R(t)\,\Omega(t) \\ \Omega(t) = \omega(t) \wedge \\ \dot{\omega}(t) = \eta_\omega(t) \\ \dot{\gamma}(t) = 0 \\ \dot{b}_a(t) = 0 \\ \dot{b}_\omega(t) = 0 \end{cases} \tag{2.17}$$

$$y_b(t) = \begin{bmatrix} R^T(t)\left(a(t) + \gamma(t)\right) + b_a(t) + \nu_a(t) \\ \omega(t) + b_\omega(t) + \nu_\omega(t) \end{bmatrix} \tag{2.18}$$

Chapter 3

Visual measurements as motion constraints

This section introduces the visual output model employed for constraining the motion of the system, by using epipolar constraints as output maps. First, the visual constraints are treated as general nonlinear maps, a class in which other classical methods like projections on the image plane of $3D$ *points fall as well. The discussion will be dedicated to the "features-based" visual constraints case, in which every feature is taken per-se and has an output associated to it. Finally the output map used in this work is presented and formally characterized.*

3.1 Epipolar Constraints

Suppose a camera observes a $3D$ point $P^i = \begin{bmatrix} X^i & Y^i & Z^i \end{bmatrix}^T$ fixed in space from two distinct poses c_τ, c_t with respect to the fixed reference frame in which the coordinates of the point P^i are defined. We call $g_{c_\tau c_t} \in SE\left(3\right)$ the relative transformation between the two poses of the camera and P_τ^i and P_t^i the rays from the optical center of the camera to the point P^i in the two positions c_τ, c_t. These two points are related by a simple rigid motion relationship:

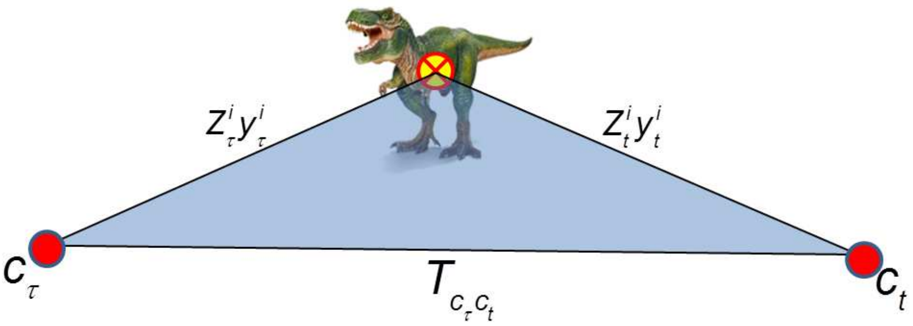


Figure 3.1: *Graphical interpretation of the epipolar constraint. The red circles indicate the position of the camera in two different instants.*

$$P_t^i = g_{c_\tau c_t} P_\tau^i \tag{3.1}$$

If we define the normalized coordinates of the points P_j^i as $y_j^i = \left[\begin{array}{ccc} X_j^i/Z_j^i & Y_j^i/Z_j^i & 1 \end{array}\right]^T$, $j = \tau,\, t$ we have:

$$Z_t^i y_t^i = g_{c_\tau c_t} Z_\tau^i y_\tau^i = R_{c_\tau c_t} Z_\tau^i y_\tau^i + T_{c_\tau c_t} \tag{3.2}$$

Via simple algebraic manipulation, it is possible to obtain (Longuet-Higgins [21]):

$$Z_t^i y_t^{iT}\left(T_{c_\tau c_t} \wedge y_t^i\right) = 0 = Z_\tau^i y_t^{iT} T_{c_\tau c_t} \wedge R_{c_\tau c_t} y_\tau^i,\ Z_\tau^i > 0 \tag{3.3}$$

Condition (3.3) has a direct interpretation: given the correspondences between two points, in normalized coordinates, y_τ^i and y_t^i, in two successive images (with time $t > \tau$), after the camera moved by a certain transformation $g_{c_\tau c_t} \triangleq \{R_{c_\tau c_t}, T_{c_\tau c_t}\}$, the three vectors y_τ^i, y_t^i and $T_{c_\tau c_t}$ are coplanar[1]. Points y_τ^i and y_t^i are expressed in the local coordinates of the camera at the time τ and t respectively. By expressing the features in a common reference frame, for example in the one corresponding to the last position, the condition highlighted above defines the well known formulation [21]:

$$y_t^{i^T} T_{c_\tau c_t} \wedge \left(R_{c_\tau c_t} y_\tau^i\right) = 0 \tag{3.4}$$

known as *epipolar constraint*. The above constraint can be written for every visible pair of points that share the same relative camera transformation, i.e.:

$$y_t^{i^T} T_{c_\tau c_t} \wedge \left(R_{c_\tau c_t} y_\tau^i\right) = 0, \forall i = 1, \ldots, N \tag{3.5}$$

From now on, the constraint in equation (3.4) will be written in the more compact notation, that is:

$$\phi\left(g_{c_\tau c_t}, y_\tau^i, y_t^i\right) = y_t^{i^T} T_{c_\tau c_t} \wedge \left(R_{c_\tau c_t} y_\tau^i\right) = 0 \tag{3.6}$$

The matrix $E = T_{c_\tau c_t} \wedge R_{c_\tau c_t} \subset \mathbb{R}^{3\times 3}$ is defined as *essential matrix* and it is a point of the *essential manifold* [21, 39], that is the particular set

$$\mathcal{E} \triangleq \{E = T \wedge R \,|\, T\wedge \in so(3), R \in SO(3)\} \tag{3.7}$$

[1] We recall that this happens regardless of the depth of the points, since equation (3.3) is zero $\forall Z_\tau^i$.

The complete characterization and the properties of the essential manifold can be found in the given references and will be omitted here for brevity. We are interested in exploiting the local parametrization of the essential manifold, in terms of rotation and translation, i.e. the motion recovery problem.

3.1.1 Iterative solution to motion recovery: Horn's method revisited

When eight or more independent constraints of the form (3.4) can be set, it is possible to characterize (up to a scale factor) the motion of the viewer [21]. Standard methods in the literature for recovering the motion parameters from a given set of point matches are well established and they are both algebraic (the eight-point algorithm [21, 24]) or iterative (see Horn [12] for instance). The last one, in particular, recovers the motion parameters by minimizing a certain norm of the set of epipolar constraints computed over the observed features. In this case, there is no closed solution, unlike the case of the eight-point algorithm, and an initial (weak) estimation of the motion parameters is required. At the same time, the whole set of visual features can be used in the optimization (unlikely in the the eight-point algorithm). However, the approaches are both developed for monocular vision and thus are affected by the scale ambiguity. The algebraic method fixes the gauge and returns an estimation of the translation vector which is normalized. The same does not generally happen for the iterative schemes, unless the scale normalization is enforced.

In case a calibrated stereo rig is available, we propose to revisit the idea by Horn and end up with an estimation scheme which is able to recover all the nine degrees of freedom of the motion parameters. Although this claim can be intuitive, a formal proof on disambiguation of the global scale can be found in Appendix A. When a $3D$ point P^i fixed in space is observed by the two cameras in a calibrated stereo configuration, from two distinct point of views, the $3D$ points seen by each camera are related by rigid motion relationships. For the left camera:

$$P^i_{l2} = g_{c_1c_2} P^i_{l1} \tag{3.8}$$

for the right one:

$$P^i_{r2} = g_{lr} g_{c_1c_2} P^i_{l1} \tag{3.9}$$

with obvious meaning of the symbols. It is assumed that $g_{c_1c_2} \in SE\left(3\right)$ describes the relative transformation between the two poses of the left camera. g_{lr} is the (constant and known) calibration of the stereo pair. When a group of N points is tracked on the left and right frames, $2N$ groups of constraints can be set up:

$$\begin{aligned} \phi\left(g_{c_1c_2},\, y^i_{l1},\, y^i_{l2}\right) &= 0 && (3.10)\\ \phi\left(g_{lr}g_{c_1c_2},\, y^i_{l1},\, y^i_{r2}\right) &= 0,\ \forall i = 1, \dots, N && (3.11)\end{aligned}$$

As a simplified example, we considered the case when the same number of features is observed in both the left and right images. However this usually does not happen, and will be taken into consideration in the definition of the visual measurement model. We propose to solve the following optimization problem, once some penalty function $\mathcal{L}$ of the constraints is chosen.

$$
\begin{aligned}
(T_{c_1c_2}, R_{c_1c_2}) = \min_{T_{c_1c_2}, R_{c_1c_2}} \Bigg(& \sum_{i=1}^{N} \mathcal{L}\left\{\phi\left(g_{c_1c_2},\, y_{l1}^{i},\, y_{l2}^{i}\right)\right\} + \dots \\
& \sum_{i=1}^{N} \mathcal{L}\left\{\phi\left(g_{lr} g_{c_1c_2},\, y_{l1}^{i},\, y_{r2}^{i}\right)\right\} \Bigg)
\end{aligned}
\tag{3.12}
$$

In general a squared 2-norm can be chosen or any robust version [13], in the case of outlier contamination. The optimization can then be solved via standard local gradient-based search methods.

3.2 Navigation error estimation via epipolar constraints

In this work, rather than in an algebraic way, the epipolar constraints are treated as the outputs of a suitable dynamical system, leading to the possibility to be used in the framework of stochastic filtering [15]. This section shows how to construct a visual measurement model for the kinematic model (2.17), starting from the definition (3.6).

With reference to real applications, image features are usually known up to a certain error ν_i which can be statistically modeled with a white zero-mean, normally distributed stochastic process, that is

$$\widetilde{y}^{i} = y^{i} + \nu_i,\ \ \nu_i \sim N\left(0, R_y\right) \tag{3.13}$$

Moreover, only an estimate of the relative camera motion $g_{c_\tau c_t}$ is usually available (e.g. by inertial mechanization),

$$\widehat{g}_{c_\tau c_t} = \delta g\, g_{c_\tau c_t} = \left\{\delta R R_{c_\tau c_t}, T_{c_\tau c_t} + \delta T\right\} \tag{3.14}$$

The term δg models the error between the true value and its estimation, and we assume it to be bounded. In the context of aided inertial navigation this is a realistic assumption, provided that successive measurements from the aiding sensors come frequently.

When the constraint (3.6) is applied to noisy quantities, a residual appears, in order to balance the equality in equation (3.6). We will indicate such situation with the notation:

$$\phi\left(\widehat{g}_{c_\tau c_t}, \widetilde{y}^i_\tau, \widetilde{y}^i_t\right) = \epsilon^i \neq 0 \tag{3.15}$$

Let's suppose that, at a given time τ, a group of features $y^i_{l\tau} \in \mathcal{Y}_{l\tau}$ was detected on the left image for the first time. At current time $t > \tau$, we select N features on the left image and M features on the right image, among the ones detected on the two images at the current time, according to the following rules:

$$\mathcal{Y}_{lt} \doteq \left\{ y^i_{lt} : y^i_{lt} \text{ is a track of } y^i_{l\tau} \in \mathcal{Y}_{l\tau},\ i = 1, \dots, N \right\} \tag{3.16}$$

$$\mathcal{Y}_{rt} \doteq \left\{ y^j_{rt} : y^j_{rt} \text{ is a track of } y^j_{l\tau} \in \mathcal{Y}_{l\tau},\ j = 1, \dots, M \right\} \tag{3.17}$$

Note that $\mathcal{Y}_{rt} \subseteq \mathcal{Y}_{lt} \subseteq \mathcal{Y}_{l\tau}$. Once the two above sets are defined, $N + M$ constraints like (3.15) can be written, N for the left camera and $M \leq N$ for the right one:

$$\begin{cases} \phi\left(\widehat{g}_{c_{l\tau} c_{lt}}, y^i_{l\tau}, \widetilde{y}^i_{lt}\right) = \epsilon^i_l \\ \phi\left(g_{lr} \widehat{g}_{c_{l\tau} c_{lt}}, y^j_{l\tau}, \widetilde{y}^j_{rt}\right) = \epsilon^j_r \end{cases} \tag{3.18}$$

Again, g_{lr} represents the (constant and known) calibration of the stereo pair. $\widehat{g}_{c_{l\tau} c_{lt}}$ is the estimation of the relative transformation between the positions of the left camera at time τ and t. The choice to consider only those features on the right image that have a correspondence with the left features being tracked is arbitrary and was made for convenience. The only condition to be respected is $\mathcal{Y}_{rt} \neq \emptyset$ which is needed to disambiguate the global scale, as it will be shown in Claim 4, Chapter 4, and in the Appendix A. Note that the normalized measurements $y^i_{l\tau}$ are written without the *tilde* hat: this is because we assume that the initial detection of a certain group of features *defines* the locations being tracked in the successive images [16]. This is equivalent to consider such location as a *reference* for the future detection of the features. It is thus considered *noiseless*, which, in turn, has the advantage of eliminating the need for including correlations between feature noise in different time steps [39] and to keep the loosely-coupled structure of the estimation[2]. Clearly, this assumption has the disadvantage that we loose confidence with the actual nature of the noise, resulting in an unavoidable bias in the estimation. Statistically, this approximation means to consider the reference features *on average* in the correct position on the image plane. One could be interested in quantifying statistically the error related to such choice. It is generally a reasonable assumption to consider the visual measurements statistically independent among them. In this case, the *standard uncertainty* can be used as an indicator about the uncertainty made on averaging (also known as standard error of the mean)[3]. When N features are observed, the standard uncertainty reduces to:

$$SU_{\bar{y}} = \frac{\bar{\sigma}}{\sqrt{N}} \tag{3.19}$$

[2] Otherwise we should add the noise components of the reference features in the state space, to comply with the causal estimation scheme.

[3] The standard uncertainty can be defined as the standard deviation of the estimate of the sample-mean's of a population [14].

where $\bar{\sigma}$ is the sample standard deviation [14], that is the sample-based estimate of the standard deviation of the observations set, usually computed as:

$$\bar{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_\tau^i - \bar{y}_\tau^i)^2}. \tag{3.20}$$

and $\bar{y}_\tau^i$ is the sample mean. This suggests that keeping the number of tracked features high would reduce the overall uncertainty in motion estimation. *De facto* this number can be around hundreds, like field tests suggest.

Minimizing the epipolar residual corresponding to one constraint like (3.18) results in constraining the motion of the camera (left or right) along one direction, which is normal to the plane defined by such constraint. Once $M + N$ independent constraints are given, in a stereo vision configuration, it is possible to constrain the motion of the stereo system along all the 6DOFs of the motion space (cf. Lemma 1 and Appendix A). In order to fuse the vision system with the inertial navigation system, it is necessary to define a common measurements vector. The epipolar constraints between the initial and current times can be related to the motion states (position and attitude) – equation (2.17) – as:

$$\left\{ \begin{array}{l} \phi\left(g\left(t\right)^{-1} g\left(\tau\right),\, y_{l\tau}^i,\, y_{lt}^i\right) = 0 \\ \phi\left(g_{lr} g\left(t\right)^{-1} g\left(\tau\right),\, y_{l\tau}^j,\, y_{rt}^j\right) = 0 \end{array} \right. \tag{3.21}$$

where $g\left(\cdot\right)$ are the *body* poses expressed in the navigation frame at current time and at the time τ, when the group of features being tracked was seen for the first time. Since we have only the estimation of the transformations $g\left(t\right)$ and $g\left(\tau\right)$ and only a measurement of image points y_{lt}^{i} and y_{rt}^{j}, equation (3.21) can be thus rewritten as:

$$\left\{\begin{array}{l} \phi\left(\widehat{g}\left(t\right)^{-1}\widehat{g}\left(\tau\right),\, y_{l\tau}^{i},\, \widetilde{y}_{lt}^{i}\right) = \epsilon_{l}^{i} \neq 0 \\ \phi\left(g_{lr}\widehat{g}\left(t\right)^{-1}\widehat{g}\left(\tau\right),\, y_{l\tau}^{j},\, \widetilde{y}_{rt}^{j}\right) = \epsilon_{r}^{j} \neq 0 \end{array}\right. \tag{3.22}$$

Equations (3.21) and (3.22) represent a simplified formulation for the epipolar residual since it makes the camera frame coincide with the IMU (body) frame. Usually, the reference frames associated to the IMU and the camera do not coincide and the epipolar residual equation should depend on their relative position. Since we assume that the constant transformation that maps the calibration parameters between the cameras and the IMU is known, we dropped such term from the residual formulation, to simplify the notation.

The rationale behind the proposed approach is to use the constraints (3.22) as a measure of the mismatch between estimated navigation state and the *actual* one, then to use this measure to correct filter state and improve navigation accuracy. To this end, a Constrained Extended Kalman Filter is proposed in the succeeding sections, which will make use of the foregoing visual measurement model.

Chapter 4

Observability Analysis

One of the contributions of this work is the full analytical characterization of the unobservable space of the class of constraints-based loosely-coupled problems where the one proposed falls. This result is in line with the most recent works [16, 18] of tightly-coupled visual-inertial navigation. The main contribution is summarized by the Lemma 1, which proof can be found in Appendix A.

A very good introduction to non linear observability can be found in [11]. In general, the role of observability analysis is to assess the possibility to uniquely disambiguate the initial conditions of the state movements of a dynamical system, by observing the outputs. Determining whether state movements that, starting from different initial conditions, return the same outputs exist or not is an essential problem in the estimation and filtering theory, since determines the possibility for an estimator to converge to the *actual* value. Observability of vision-only structure-from-motion is mainly due to [2], where the conditions for enforcing observability were made explicit. More recently, observability of motion from combined visual-inertial measurements has been assessed in the framework of tightly-coupled monocular navigation [16] and localization in a structured environment [18]. In the last case, the same results of the first work were proved, while solving the sole problem of sensor-to-sensor self calibration. The main results can be collected in some main claims.

Claim 1 (Chiuso et al. [2])
Observability can be enforced by fixing the direction of three non-collinear points on the image plane and one depth scale. This is equivalent to fix the global reference frame.

Fixing the global reference frame as in [2] is necessary in the approaches that rely on estimation of motion *and* structure to enforce the initial conditions $R(0) = I$ and $T(0) = 0$ in the filter. This avoids the structure to move freely along the unobservable direction, which otherwise would have destructive effects on the estimation of motion also.

Claim 2 (Jones and Soatto, [16])
Motion and structure are observable up to an arbitrary choice of the Euclidean reference frame, under a condition of general motion, provided that the global reference frame is fixed as in [2].

Claim 3 (Jones and Soatto, [16])
Scale, gravity and IMU-Camera calibration are observable as long as the motion is general and the global reference frame is fixed as in [2].

The concept of *general motion* is the same of persistence of excitation and it was identified with non-constant rotation along at least two axis and varying acceleration, in the case of monocular visual-inertial navigation.

One may wonder if the same conditions do apply to the problem proposed in this work or if some differences are noticeable. Moreover, if it is possible to unify the existing loosely-coupled approaches that fall in the class like the one proposed, under the same convergence properties. The reason behind the rest of the chapter is to give an answer to this question.

4.1 The observability Lemmas

The following results were obtained by considering the model (2.17), (2.18).

Claim 4 (Disambiguation of global scale)
The global scale α is observable, given the $M + N$ independent stereo epipolar constraints in Equation (3.21)*.*

We emphasize that the difference with respect to the cited works is that the disambiguation of the global scale is obtained by using stereo vision: as expected, the knowledge of the relative transformation between the left and right cameras is sufficient to recover the scale factor. The most important assumption is that there are enough common features between the left and right frames.

One contribution of the work is the extension of the observability results in the case of dealing with uncalibrated IMUs. Similar results were reached in [18] in the case of landmark-based and tightly-coupled navigation, while [16] assumed the use of a calibrated

IMU. Here we show that, in the case of constrained-based loosely-coupled navigation from visual-inertial measurements, the recovery of the sensitivity parameters is still feasible, under the conditions highlighted.

Claim 5 (Observability of the gyroscope biases)
The gyroscope biases b_ω are observable with any kind of motion in the combined vision-inertial configuration, provided they are added to the filter state with trivial dynamics (null time-derivative).

Claim 6 (Observability of the accelerometer biases)
The accelerometers biases b_a are observable provided that they are added to the filter state with trivial dynamics (null time-derivative) and the rotational motion is rich enough.

The following Lemma analytically defines the set of ambiguities of the system (2.17), (2.18).

Lemma 1
The system (2.17)-(2.18), *augmented with the (stereo) $N + M$ constraints* (3.21) *is locally observable up to the gauge transformation [26] $\bar{g} = \left\{\bar{R}, \bar{T}\right\}$, provided that the motion is rich enough. In particular, by arbitrarily choosing constant $\bar{R} \in SO(3)$ and $\bar{T} \in \mathbb{R}^3$, identical measurements are produced by:*

$$\begin{cases} \widetilde{R}(t) = \bar{R}R(t), \ \forall t \geq 0 \\ \widetilde{T}(t) = \bar{R}T(t) + \bar{T}, \ \forall t \geq 0 \\ \widetilde{V}(t) = \bar{R}V(t), \ \forall t \geq 0 \\ \widetilde{a}(t) = \bar{R}a(t), \ \forall t \geq 0 \\ \widetilde{\gamma} = \bar{R}\gamma \end{cases} \tag{4.1}$$

The variables in Equation (4.1) without the hat notation, are the *true* movements of the states of the system (2.17), i.e. the ones obtained starting from the actual initial conditions. The variables with the *tilde* hat are the state movements integrated starting from different initial conditions, obtained by selecting arbitrary values of the constant terms $\bar{R} \in SO(3)$ and $\bar{T} \in \mathbb{R}^3$. The Lemma 1 can be easily proven by substitution of the variables with the *tilde* hat in Equation (4.1) in the corresponding hat-free ones in Equations (2.18) and (3.21) and by showing that they produce the same measurements. In Appendix A a formal proof with the full derivation is given. The interpretation of the observability up to a *gauge ambiguity* is that the motion variables, together with the *local* direction of gravity, remain ambiguous up to a certain class of equivalence in $SE(3)$, which is intrinsic in the class of problems where only relative information are employed and no prior knowledge about the environment structure is available. Formally this implies the possibility to recover the sole equivalence class where the initial condition belongs, not the initial condition itself (see also [24]). This is in line with the most recent works [16, 18]. It is worth to notice that, unlike in structure-from-motion algorithms, the approach explained in this work automatically chooses a representative for the equivalence class which is coincident with the filter initial conditions. This is because, not involving structure in the estimation, all the elements of the reference structure are fixed, which allows to anchor the initial conditions.

The invariance results highlighted above are pretty general, and apply to the entire class of constrained-based problems where the global location of the observed points is not known a priori. Under such assumptions, the results do not depend on the visual constraint employed, provided that an approach based on opportunistic features is used, neither on the number of reference poses one considers into the estimation task. In any case, at least rotation and translation can be resolved up to an Euclidean group transformation, if no other assumptions are made. As a gentle discussion, we can think to work with multiple constraints as in [6] and consider for the visual update all the features that share the current time t, but have been collected in different times τ_i, thus have their own reference pose $g\left(\tau_i\right)$. The i-th constraint reads:

$$\phi\left(g\left(t\right)^{-1}g\left(\tau_i\right),\, y^i_{\tau_i},\, y^i_t\right)=0 \tag{4.2}$$

It is straightforward to show that infinitely many constant transformations $\bar{g}\in SE\left(3\right)$ (common among features) do exist, that are *in-between* $g\left(t\right)$ and $g\left(\tau_i\right)$, i.e. such that $\widetilde{g}\left(t\right)=\bar{g}g\left(t\right)$ and $\widetilde{g}\left(\tau_i\right)=\bar{g}g\left(\tau_i\right)$, to which the constraint is invariant, i.e.:

$$\phi\left(\widetilde{g}\left(t\right)^{-1}\widetilde{g}\left(\tau_i\right),\, y^i_{\tau_i},\, y^i_t\right)=\phi\left(g\left(t\right)^{-1}g\left(\tau_i\right),\, y^i_{\tau_i},\, y^i_t\right)=0,\ \forall\bar{g}\in SE\left(3\right) \tag{4.3}$$

This is in fact the classical gauge invariance, since $\bar{g}$ represents an arbitrary choice of the initial Euclidean reference frame.
The same problem happens when the constraints are written in form of image-space projection of $3D$ points in space, i.e.

$$\phi\left(\cdot\right)=y^i_t-\pi\left(g\left(t\right)^{-1}P^i_0\right)=0 \tag{4.4}$$

where P^i_0 is the global[1] position of the point which image-space projection at the current time is the measurement y^i_t. Excluding the

[1] That is resolved in the coordinates of the navigation frame.

case where some of the points P_0^i are landmarks (for which case the gauge ambiguity reduces to the identity transformation) in the cases where the global position of the points is estimated via multi-view optimization [44] (as in [27] for example), the gauge ambiguity is intrinsic in the estimated position of points, since it is always possible [26] to arbitrarily choose a constant transformations $\bar{g} \in SE(3)$ and another point $\widetilde{P}_0^i$ such that $P_0^i = \bar{g}\widetilde{P}_0^i$, for which:

$$y_t^i - \pi\left(g(t)^{-1} P_0^i\right) = y_t^i - \pi\left(g(t)^{-1} \bar{g}\widetilde{P}_0^i\right) \tag{4.5}$$

which is, again, the classical gauge invariance, once selecting $\widetilde{g}(t) = \bar{g}^{-1}g(t)$.

4.1.1 Dealing with gravity

An important issue is that the *best* estimation of the *local* gravity $\widetilde{\gamma}$ is different from the true one, γ; it is shown that this difference is related to the initial attitude alignment error, as the last equation in (4.1) suggests. All the possible *configurations* of the local gravity vector are rotated versions of the actual vector, which can be represented as the equivalence class

$$\left[\bar{R}\gamma\right]_{\bar{R}\in SO(3)} \tag{4.6}$$

We could be interested in estimating the initial misalignment of the body with respect to the actual vertical direction. To do so, we just need to extract the actual gravity vector γ from its estimation $\widetilde{\gamma}$ by *factoring-out* the rotational component. By performing a Singular Value Decomposition of the local gravity $\widetilde{\gamma}$, we get:

$$\widetilde{\gamma} = U\sigma v = U\gamma \tag{4.7}$$

The orthonormal matrix U is not $\bar{R}$ itself: the difference between the two is an arbitrary rotation about the axis of the vector γ, that is:

$$[U \exp(\gamma\wedge)\, \gamma]_{U,\exp(\cdot)\in SO(3)} \tag{4.8}$$

This is due to the fact that the direction of gravity spans only 2 degrees of freedom, being the vector γ invariant with respect to rotation along its direction (every rotation of the kind $\exp(\gamma\wedge)$).

This conclusion inspired us to modify the parametrization of the gravity in the state vector, by using two angles $\theta_\gamma, \psi_\gamma$ that span exactly the two degrees of freedom relative to the gravity, instead of using the three components of the gravity itself. It can be shown that the convergence properties are not influenced by such change of coordinates. However this can avoid numerical ill-conditioning due to the required constant norm of the gravity vector, by using approximated estimation methods such as Extended Kalman Filters.

4.1.2 Pushing the gauge recovery

As already specified, fixing the whole reference structure has the only effect to select one element in the equivalence class, thus enforcing the filter initial conditions. Even if in general this is the best thing we can do, some countermeasures can be taken in order to get closer to the best result possible, in terms of reducing the ambiguity. It should have become clear that the most important thing is the possibility to fix some known directions where the ambiguity happens, that is between the viewer and the navigation reference frames. One possibility which generally works is to recover the direction of the gravity in the body reference frame, for example dealing with calibrated IMUs. In this case a minimal realization is the model (2.17) once the states corresponding to the accelerometers bias and the gravity are removed from the system. Thus, the last equality in equation (4.1) would read $\gamma = \bar{R}\gamma$, thus forcing $\bar{R} = \exp(\gamma\wedge)$, that is an arbitrary rotation along the vertical direction. Obviously nothing can be done for the translation ambiguity, unless the initial position of the system is known.

4.1.3 Motion requirements for observability

The formal definition of *rich enough* rotational motion is given in Appendix. To the extent of the proof of the above lemma, it means that $R(t)$ must not be constant and rotations should happen along at least two axes, while keeping the direction of the resulting angular velocity vector non constant. The proof shows that, given two non-constant angular velocities $\omega_i(t), \omega_j(t)$ along two independent (orthogonal) directions $\vec{i}$, $\vec{j}$ and such that $\omega_i(t) \neq \omega_j(t)$, every angular velocity of the form:

$$\omega(t) = \omega_i(t)\,\vec{i} + \omega_j(t)\,\vec{j} \tag{4.9}$$

ensures observability.

4.1.4 Reduced order observers for estimation with degenerate motions

Regular motions of ground vehicles do not generally meet the motion requirements for observability given by Lemma 1: in general, the vehicle accelerates for very short periods of time and most of the angular velocity is along the vertical axis (heading changes only). This was also the case of the experiments carried out. Unfortunately, the set of angular motions along (at most) one axis does not fall into the class of *rich enough* motions we identified. The observability study proposed above suggests that in these condition the system is non observable. In particular, the proof given in Appendix shows that the gravity and the accelerometers biases cannot be disambiguated; gyroscopes biases are observable in any case. Generally speaking, in order to push model observability, some countermeasures maybe set up: 1) use a reduced-order observer, i.e. remove the unobservable variables from the state or 2) saturate the filter along the unobservable components of the state space. The latter approach can be implemented by fixing the unobservable states to their initial conditions. In order to tell an EKF to keep a state almost fixed, it is sufficient to use a very small value in the corresponding entries of the state covariance matrix: this prevents the filter from moving freely along the unobservable directions of the system. In the section dedicated to the experimental results, an initialization procedure is presented which allows to successfully set-up a filter able to estimate the motion in the case of degenerate (more realistic) motion conditions.

Chapter 5

Motion estimation with state equality constraints

This section introduces the algorithms for fusing inertial measurements with visual constraints, in order to solve the Ego-Motion estimation problem. The chapter starts by introducing the iterative schemes for the optimal filtering with nonlinear equality constraints. Thus it will show how to incorporate the visual constraint with the state estimation problem, given the inertial measurements, by using the sub-optimal iterative schemes. Finally the Constrained Extended Kalman Filter algorithm adopted in this work is presented.

5.1 Projection method

Several (sub-)optimal algorithms were developed in the literature for dealing with linear and non linear equality-constrained state estimation. A complete and very useful treatise can be found in [42, 37].

As a gentle introduction, in the framework of optimal non linear filtering, the issue can be viewed as solving the following problem.

Problem 1 (Constrained optimal estimation)
Given a non linear discrete-time system with state equations

$$\begin{cases} x\left(t+1\right)=f\left(x\left(t\right)\right)+\nu_x\left(t\right) \\ y\left(t\right)=h\left(x\left(t\right)\right)+\eta_y\left(t\right) \end{cases} \tag{5.1}$$

with $\nu_x\left(t\right)\sim\mathcal{N}\left(0,Q\right),\ \eta_y\left(t\right)\sim\mathcal{N}\left(0,R\right)$, *determine:*

$$\begin{cases} \max\limits_{x} p\left(x\left(t\right)\left|Y^t\right.\right) \\ s.t.\ \ \phi\left(x\left(t\right)\right)=d \end{cases} \tag{5.2}$$

for a given function $\phi\left(x\left(t\right)\right)$ *and* d.

The function $p\left(\right)$ to be maximized is the posterior distribution of the state, given the set of outputs up to the current time, $Y^t=\left\{y_s,\ t_0\leq s\leq t\right\}$. It is known [15] that, being the system driven by a Gaussian noise, the Extended Kalman Filter is the minimum variance estimator which locally maximize the posterior distribution $p\left(x\left(t\right)\left|Y^t\right.\right)$. However, the maximization of the posterior alone will not, in general, satisfy the given constraint. Several solutions have already been proposed in the literature [37] to the given motivational problem. The most general solution falls in the class of the so called "Projection methods" , where the a-posteriori Kalman state estimation $\widehat{x}\left(t\right)^+$ (i.e. the one which maximizes the given posterior) is projected onto the constraint-space via classical constrained optimization methods. The solution is [42]:

$$\widehat{x}^p\left(t\right)^+=\widehat{x}\left(t\right)^+ + WH_\phi^T\left(H_\phi WH_\phi^T\right)^{-1}\left(d-\phi\left(\widehat{x}\left(t\right)^+\right)\right) \tag{5.3}$$

for every $W \succ 0$. H_ϕ is the Jacobian of the constraint function with respect to the state, computed on the unprojected estimation $\widehat{x}(t)^+$. The reduction in uncertainty in the state estimation can be easily found as:

$$P_x^p(t)^+ = P_x(t)^+ - WH_\phi^T \left(H_\phi W H_\phi^T\right)^{-1} H_\phi P_x(t)^+ \tag{5.4}$$

where $P_x(t)^+$ is the updated covariance matrix of the estimation error. It can be proven [38] that: 1) $\widehat{x}^p(t)^+$ is an unbiased state estimator for the foregoing system, given the constraint, for every symmetric positive definite matrix W. 2) By choosing $W = P_x(t)^+$, the projection method is equivalent to the maximum probability constrained estimation

$$\max_x p\left(x(t) \,\middle|\, Y^t, \phi(x(t)) = d\right) \tag{5.5}$$

provided that the projected estimation and its covariance are fed back into the estimation scheme [37]. This, in turn, corresponds to the most popular scheme known as Measurement Augmentation Kalman Filter [37], which treats the constraint as a *perfect measurement* included into the measurement vector.

5.2 Minimum variance estimation with stochastic constraints

The foregoing method is generally suitable when the constraints are given in the form of deterministic constraints, such that the projection actually lies exactly on the constraint space. When the constraints take the form of stochastic functions, the optimal state estimation is asked to *weakly* lie on the constraint space [3], i.e. to *minimize* an error with respect to the stochastic constraint. The solution to a similar problem (in the case of constraints in implicit form), was called *Implicit filtering* [39], in the framework of vision-only navigation on the Essential Manifold, which falls actually in the class of stochastic Measurement Augmentation Kalman Filter. This result can be included into the most general approach reviewed in the previous section, showing that it is just a particular solution. For this aim, the motivational problem 1 can be recast as:

Problem 2 (Stochastic-constrained estimation problem)
Given the system (5.1)*, determine:*

$$\begin{cases} \max\limits_x p\left(x\left(t\right)\left|Y^t\right.\right) \\ s.t.\ \phi\left(x\left(t\right),\widetilde{z}\left(t\right)\right) \textit{ is minimum} \end{cases} \tag{5.6}$$

given a certain random variable $\widetilde{z}\left(t\right) = z\left(t\right) + \nu_z\left(t\right)$, $\nu_z\left(t\right) \sim \mathcal{N}\left(0, R_z\right)$.

The problem is well posed if we assume that for noiseless variable z is $\phi\left(x\left(t\right),z\right) = 0$; in this case we will meet the property of the epipolar constraints also. Note that now we put $d = 0$. Under such condition, we have (dropping the time index for convenience):

$$0 = \phi\left(x, z\right) = \phi\left(x, \widetilde{z} - \nu_z\right) \tag{5.7}$$

A linearization of the constraint around small variations of the noise around its mean value leads to:

$$0 = \phi\left(x, \widetilde{z} - \nu_z\right) \approx \phi\left(x, \widetilde{z}\right) + \underbrace{\frac{\partial\phi}{\partial\nu_z}\nu_z}_{\widetilde{\nu}_\phi} \tag{5.8}$$

that is:

$$0 - \phi\left(x, \widetilde{z}\right) \approx \widetilde{\nu}_\phi \tag{5.9}$$

where the change in the sign (eq. (5.8)) was made for convenience, given the zero-mean characteristic of the noise. By fixing a certain value for the state x, equation (5.9) plays the role of an innovation term: the goal is to find such value that maximizes the posterior $p\left(x\middle|Y^t\right)$, while minimizing the innovation (5.9). A possible optimization objective is [15]:

$$\min_{\widehat{x}^p}\left\{\left(\widehat{x}^{p+} - \widehat{x}^{+}\right)^T W^{-1}\left(\widehat{x}^{p+} - \widehat{x}^{+}\right) + \left(-\phi\left(\widehat{x}^{p+}, \widetilde{z}\right)\right)^T \widetilde{R}_\phi^{-1}\left(-\phi\left(\widehat{x}^{p+}, \widetilde{z}\right)\right)\right\} \tag{5.10}$$

where $\widetilde{R}_\phi = \frac{\partial\phi}{\partial\nu_z} R_z \frac{\partial\phi}{\partial\nu_z}^T$ and $W \succ 0$. The projected estimation can be obtained in general form as:

$$\widehat{x}^{p+} = \left(W^{-1} + H_\phi^T \widetilde{R}_\phi^{-1} H_\phi\right)^{-1}\left(W^{-1}\widehat{x}^{+} + \widetilde{R}_\phi^{-1}\phi\left(\widehat{x}^{+}, \widetilde{z}\right)\right) \tag{5.11}$$

which, by using the Sherman-Morrison-Woodbury formulae, becomes:

$$\widehat{x}^{p+} = \widehat{x}^{+} - W H_\phi^T\left(H_\phi W H_\phi^T + \widetilde{R}_\phi\right)^{-1}\phi\left(\widehat{x}^{+}, \widetilde{z}\right) \tag{5.12}$$

which variance is:

$$P_x^{p+} = P_x^{+} - W H_\phi^T\left(H_\phi W H_\phi^T + \widetilde{R}_\phi\right)^{-1} H_\phi P_x^{+} \tag{5.13}$$

By using the same discussion as in [42] and putting $W = P_x^{+}$, equation (5.12) is the Minimum Variance Estimator of the state $x\left(t\right)$, which solves the problem 2 and (5.13) is its covariance. Again, provided that the projected estimation and its covariance are fed back into the estimation scheme, it can be shown that equations (5.12) and (5.13) with $W = P_x^+$ are equivalent to the solutions of:

$$\max_x p\left(x\left(t\right)\left|Y^t, \phi\left(x, \widetilde{z}\right)\right.\right) \tag{5.14}$$

which is solved by extending the system output vector with the constraints and by running the Extended Kalman Filter on the extended system.

5.3 Implementation

According to the motion and sensitivity parameters dynamics in equations (2.17) and (2.18), given the constraints (3.22), a nonlinear Kalman Filter was designed and tested, in order to solve the problem as in equation (5.14). In this case the function $\phi\left(x, \widetilde{z}\right)$ is replaced with the set of epipolar constraints (3.22). The aim of the filter is to estimate the state $x\left(t\right)$ of the system, consisting of: the navigation variables, $T\left(t\right), v\left(t\right), a\left(t\right)$, the angular parametrization of the rotation matrix $R\left(t\right)$, the body angular velocity, $\omega\left(t\right)$, the *local* gravity vector, $\gamma\left(t\right)$, and the inertial sensor biases, b_a, b_ω. In this work, rotation matrices were parametrized using the exponential map, computed via the Rodrigues' formulae [30]. The feature detection and tracking module adopted in this article uses stereo vision and the Scale Invariant Feature Transform (SIFT) algorithm [22, 32]. The algorithm starts by acquiring a stereo images pair and relies on the SIFT algorithm to detect, select, and match features from left and right images. The stereo matching of the features is performed by comparing the squared distance between the descriptors of each feature in the two images and selecting the couple with the lowest distance. Only those features on the right image that have a corresponding match on the current left image are considered valid. With the same distance-based approach it is possible to track the features which are present in the current and reference left images. For the purpose of numerical implementation, the kinematic equations (2.17) were time-discretized using the Euler integration method. The base sample time was chosen coincident with the sampling rate of the IMU i.e. $dt = 0.01s$, which is supposed to be the fastest sample time in the estimation loop. As a simple inspection of the visual measurement model (3.22) can tell, the visual constraints depend on the pose of the system at the current time $g\left(t\right)$ and on the reference pose $g_{ref} \doteq g\left(\tau\right)$, corresponding to the pose the system had some steps back in the past, when the group of features being tracked was seen for the first time. Although other methods are viable, we decided to augment the state vector of the estimator with the motion parameters (angular parametrization and translation) of the reference pose, see for example [28, 4], by assigning to the new state variable the trivial dynamics $g_{ref}\left(t+1\right) = g_{ref}\left(t\right)$, since it is not going to change over time. The resulting discrete time equations of the estimator are therefore:

$$\begin{cases} T\left(t+1\right)=T\left(t\right)+v\left(t\right)dt \\ v\left(t+1\right)=v\left(t\right)+a\left(t\right)dt \\ a\left(t+1\right)=a\left(t\right)+\eta_{a}\left(t\right)dt \\ R\left(t+1\right)=\exp\left(\Omega\left(t\right)dt\right)R\left(t\right) \\ \Omega\left(t\right)=\omega\left(t\right)\wedge \\ \omega\left(t+1\right)=\omega\left(t\right)+\eta_{\omega}\left(t\right)dt \\ \gamma\left(t+1\right)=\gamma\left(t\right)+\eta_{\gamma}\left(t\right)dt \\ b_{a}\left(t+1\right)=b_{a}\left(t\right)+\eta_{b_{a}}\left(t\right)dt \\ b_{\omega}\left(t+1\right)=b_{\omega}\left(t\right)+\eta_{b_{\omega}}\left(t\right)dt \\ g_{ref}\left(t+1\right)=g_{ref}\left(t\right) \\ y_{imu}\left(t\right)=\left[\begin{array}{c} R^{T}\left(t\right)\left(a\left(t\right)+\gamma\left(t\right)\right)+b_{a}\left(t\right)+\nu_{a}\left(t\right) \\ \omega\left(t\right)+b_{\omega}\left(t\right)+\nu_{\omega}\left(t\right) \end{array}\right] \\ \phi\left(g\left(t\right)^{-1}g_{ref}\left(t\right),\, y^{i}_{l\tau},\, y^{i}_{lt}\right)=0,\ i=1,\ldots,N\left(t\right) \\ \phi\left(g_{lr}g\left(t\right)^{-1}g_{ref}\left(t\right),\, y^{j}_{l\tau},\, y^{j}_{rt}\right)=0,\ j=1,\ldots,M\left(t\right) \end{cases} \tag{5.15}$$

Note that the number of tracked features on the left, $N\left(t\right)$, and right, $M\left(t\right)$, images incorporate the time index since we expect them to change over time.

The fundamental estimation scheme for optimally filtering motion, gravity and sensitivity parameters is based upon the stochastic Measurement Augmented Extended Kalman Filter (also named Implicit – or Essential – Filter by [39]) introduced before. Since the inertial and visual modules run with different frequencies, when no image measurements are available, the estimation cycle is performed as in standard EKF, via prediction of the state vector through the non linear system (2.17), by using the estimation of the state at the previous time step, and correction employing the new inertial measurements only. When a new pair of stereo images becomes available, the constrained correction step is performed. As anticipated, this is made by stacking the inertial measurements error vector with the epipolar constraints computed over the tracked features on the left and right images:

$$\delta y\left(t\right) = \left[\begin{array}{c} \widetilde{y}_{imu}\left(t\right) - \left[\begin{array}{c} \widehat{R}^{T}\left(t\right)\left(\widehat{a}\left(t\right)+\widehat{\gamma}\left(t\right)\right)+\widehat{b}_{a}\left(t\right) \\ \widehat{\omega}\left(t\right)+\widehat{b}_{\omega}\left(t\right) \end{array}\right] \\ \vdots \\ -\phi\left(\widehat{g}\left(t\right)^{-1}\widehat{g}_{ref}\left(t\right),\, y^{i}_{l\tau},\, \widetilde{y}^{i}_{lt}\right) \\ \vdots \\ -\phi\left(g_{lr}\widehat{g}\left(t\right)^{-1}\widehat{g}_{ref}\left(t\right),\, y^{j}_{l\tau},\, \widetilde{y}^{j}_{rt}\right) \\ \vdots \end{array}\right] \tag{5.16}$$

Then the constrained correction is performed by using the equations below:

$$S\left(t+1\right) = H\left(t\right)P\left(t+1\right)^{-}H\left(t\right)^{T}+\widetilde{R}\left(t\right) \tag{5.17}$$

$$K\left(t+1\right) = P\left(t+1\right)^{-}H\left(t\right)S\left(t+1\right)^{-1} \tag{5.18}$$

$$\widehat{x}\left(t+1\right)^{+} = \widehat{x}\left(t+1\right)^{-}+K\left(t+1\right)\delta y\left(t\right) \tag{5.19}$$

$$\Gamma\left(t+1\right) = I-K\left(t\right)H\left(t\right) \tag{5.20}$$

$$P\left(t+1\right)^{+} = \Gamma\left(t+1\right)P\left(t+1\right)^{-}\Gamma\left(t+1\right)^{T}+K\left(t+1\right)\widetilde{R}\left(t\right)K\left(t+1\right)^{T} \tag{5.21}$$

where $\widetilde{R}\left(t\right) = J\left(t\right)RJ\left(t\right)^{T}$ and R is the block diagonal matrix of noise covariances of the IMU and of the image features

$$R = \begin{bmatrix} R_{imu} & 0 \\ 0 & R_y \end{bmatrix} \tag{5.22}$$

The Jacobian matrices $H\left(t\right)$ and $J\left(t\right)$ are defined as:

$$H\left(t\right) = \begin{bmatrix} H_{imu}\left(t\right)^T & H_{\phi}\left(t\right)^T \end{bmatrix}^T \tag{5.23}$$

$$J\left(t\right) = \begin{bmatrix} I_{6\times 6} & 0 \\ 0 & J_{\phi}\left(t\right) \end{bmatrix} \tag{5.24}$$

$$\tag{5.25}$$

being respectively:

$$H_{imu}\left(t\right) = \begin{bmatrix} \left.\frac{\partial y_{imu}}{\partial x(t)}\right|_{\widehat{x}(t)} & 0 \end{bmatrix} \tag{5.26}$$

$$H_{\phi}\left(t\right) = \begin{bmatrix} \vdots \\ \begin{bmatrix} H^i_{\phi_l,t} & H^i_{\phi_l,ref} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} H^j_{\phi_r,t} & H^j_{\phi_r,ref} \end{bmatrix} \\ \vdots \end{bmatrix} \tag{5.27}$$

$$J_{\phi}\left(t\right) = \begin{bmatrix} \vdots \\ \left.\frac{\partial\phi\left(g(t)^{-1}g_{ref}(t),y^i_{l\tau},y^i_{lt}\right)}{\partial y^i_{lt}}\right|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^i_{l\tau},\widetilde{y}^i_{lt}} \\ \vdots \\ \left.\frac{\partial\phi\left(g_{lr}g(t)^{-1}g_{ref}(t),y^j_{l\tau},y^j_{rt}\right)}{\partial y^j_{rt}}\right|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^j_{l\tau},\widetilde{y}^j_{rt}} \\ \vdots \end{bmatrix} \tag{5.28}$$

with $i = 1, \ldots, N(t)$, $j = 1, \ldots, M(t)$, and

$$H^i_{\phi_l,t} = \frac{\partial \phi\left(g(t)^{-1} g_{ref}(t), y^i_{l\tau}, y^i_{lt}\right)}{\partial g(t)}\bigg|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^i_{l\tau},\widetilde{y}^i_{lt}} \tag{5.29}$$

$$H^i_{\phi_l,ref} = \frac{\partial \phi\left(g(t)^{-1} g_{ref}(t), y^i_{l\tau}, y^i_{lt}\right)}{\partial g_{ref}(t)}\bigg|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^i_{l\tau},\widetilde{y}^i_{lt}} \tag{5.30}$$

$$H^j_{\phi_r,t} = \frac{\partial \phi\left(g_{lr} g(t)^{-1} g_{ref}(t), y^j_{l\tau}, y^j_{rt}\right)}{\partial g(t)}\bigg|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^j_{l\tau},\widetilde{y}^j_{rt}} \tag{5.31}$$

$$H^j_{\phi_r,ref} = \frac{\partial \phi\left(g_{lr} g(t)^{-1} g_{ref}(t), y^j_{l\tau}, y^j_{rt}\right)}{\partial g_{ref}(t)}\bigg|_{\widehat{g}(t),\widehat{g}_{ref}(t),y^j_{l\tau},\widetilde{y}^j_{rt}} \tag{5.32}$$

Discussion: keeping the estimation coherent. As already discussed, the epipolar constraints depend on the system pose at two different time instants: one is the reference pose, the other is the one at the current time, when the update is performed. Every constraint brings new information about the current time, but provides no further information about the reference pose $g(\tau)$. This means that the uncertainty related to this pose is not affected by a new image acquisition and must be constant. Due to the linear approximations needed in computing the Kalman gain, it could happen that the state variables corresponding to the reference pose and their covariances are updated by the filter. This is undesired, because it is just an effect of the linearization (the system cannot *gain observability* during linearization). Avoiding the update of the reference pose and its covariance can be made by enforcing the null Kalman gain, that is by zeroing the rows of the Kalman gain corresponding to the state variable of the reference pose. This would ensure that the update step will not affect the reference pose and the related covariance matrix.

5.4 Dealing with occlusions and new features

The loose-coupling approach has the advantage that dealing with occlusions, is straightforward, since it is delegated to the tracker. If a feature temporarily disappears from the field of view, the tracker will not find any correspondence in the database and simply the corresponding constraint will not be computed. Should the feature appear again, it would be used to compute the corresponding constraint again. Figure 5.1 shows a typical example where some features become occluded. The images were taken during a field experiment, where the sensor suites were mounted on the top of a car moving in a dynamic environment. Per each pair, the left frame is the key-frame taken at the reference pose, while the image on the right is the current left image. The green dots are the tracked point between the two images. The occluded features are just not used for visual correction. Once become visible again, they are used in the estimation filter.

A different approach must be used with features which exit from the field of view, due to the camera motion. Usually the average lifetime of a group of features being tracked, in terms of the number of frames in which each feature is visible, before being lost, stands within few tens of frames (usually 20-30 frames, based on field tests). This, however, depends on the motion of the camera. Obviously, when tracking/matching enough features is no more possible, due

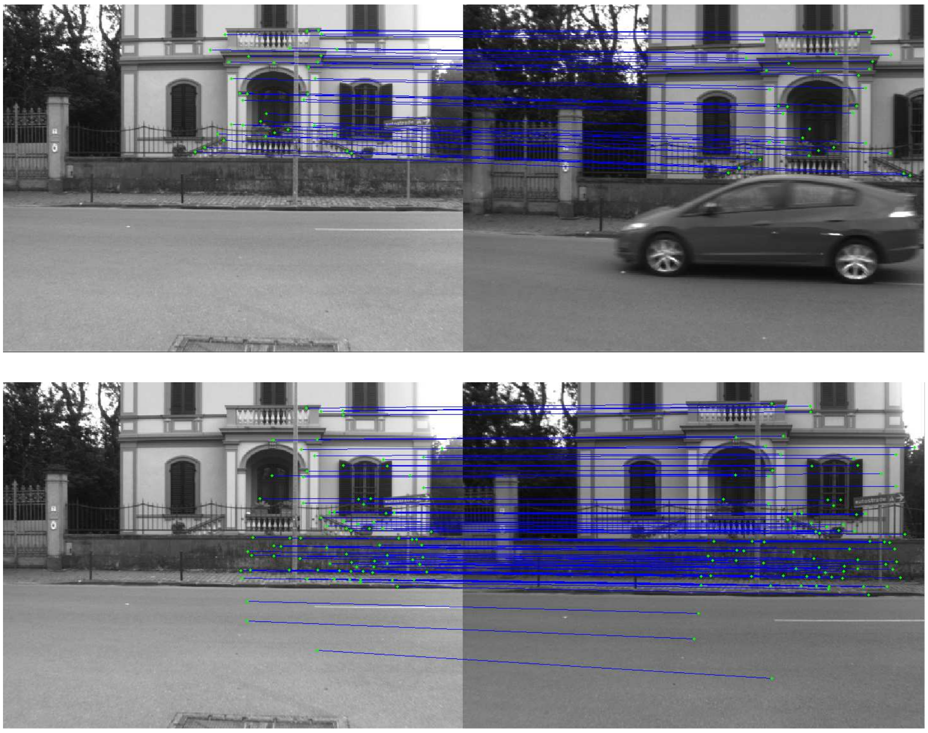


Figure 5.1: *An example on how the system deals with occlusions. The time flows from top to bottom.*

to camera motion, a new reference group must be selected that will be used to anchor the future epipolar constraints. Although other alternatives are viable [2], when a new group of reference features must be selected, at time τ: i) we first obtain the *best estimation* of the posterior $p\left(x\left(\tau\right)\middle|Y^{\tau},\phi_{i,j}\right)$ by performing a Kalman update step with the remaining tracked features. Then, ii) we proceed by marginalizing the posterior distribution with respect to the state corresponding to the refined body pose $\widehat{g}\left(\tau\right)^{+}$, namely $x_g\left(\tau\right)$, that is

$$p\left(x_g\left(\tau\right)\left|Y^{\tau},\phi_{i,j}\right.\right)=\int p\left(x\left(\tau\right)\left|Y^{\tau},\phi_{i,j}\right.\right)dx_g=\mathcal{N}\left(\widehat{g}\left(\tau\right)^{+},P^{+}_{\widehat{g}(\tau)\widehat{g}(\tau)}\right) \tag{5.33}$$

The vector $x_g\left(\tau\right)$ contains the components of the state vector relative to the position and attitude state variables (i.e. the pose variables). iii) We simply substitute the old reference pose, stored in the filter state, with this estimation of the body pose, $g_{ref}\leftarrow\widehat{g}\left(\tau\right)^{+}$. This state variable will evolve according to the null-time derivative dynamic:

$$g_{ref}\left(t+1\right)=g_{ref}\left(t\right) \tag{5.34}$$

At the same time, we *clone* the error covariance of the current body pose estimation $P^{+}_{\widehat{g}(\tau)\widehat{g}(\tau)}$ in the entries corresponding to the new state variable, i.e.

$$P^{+}_{\tau}\leftarrow\left[\begin{array}{cc} P^{+}_{\widehat{g}(\tau)\widehat{g}(\tau)} & P^{+}_{\widehat{g}(\tau)\widehat{x}(\tau)} \\ P^{+}_{\widehat{x}(\tau)\widehat{g}(\tau)} & P^{+}_{\widehat{x}(\tau)\widehat{x}(\tau)} \end{array}\right] \tag{5.35}$$

Finally, iv) we swap the group of reference features with the new acquired one, $\{y^i_{l\tau}\}$. The reference pose (and its relative covariance) is kept constant along the system evolution, according to the trivial dynamics, as long as the reference features $y^i_{l\tau}$ will be successfully tracked in the future images. When at time $\tau+h$ a new group of features must be acquired, we perform again a Kalman update step with the remaining tracked features, then we marginalize and replace the reference pose with $\widehat{g}\left(\tau+h\right)^{+}$, i.e. the new body pose estimation at the current time; finally we *clone* again the error covariance of the current body pose estimation in the entries corresponding to the new state variable. This is in line with a more general approach, known as *stochastic cloning* [28, 4] and allows to keep track of the

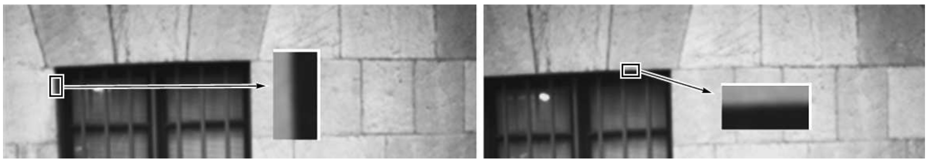


Figure 5.2: *Example of a possible matching ambiguity connected with the use of covariant feature detectors (e.g., SIFT).*

cross covariance of estimated navigation between the two time instants. Obviously, given the relative nature of the measurements employed, this technique does not prevent the filter from diverging over time from the actual state. The problem is in fact in the pose switching mechanism, as also specified in [2]. Before performing the pose substitution $g_{ref} \leftarrow \widehat{g}\left(\tau\right)^{+}$, the error between the actual state and the estimated one is a random variable with mean value $\widehat{x}\left(\tau\right)^{+}$ and covariance $P^{+}_{\widehat{x}(\tau)\widehat{x}(\tau)}$. When the pose and its covariance is stored in the filter memory, this error cannot be corrected anymore, unless more complex mechanisms are employed (for example loop-closure), and the state subspace corresponding to the system pose will move by an amount of $P^{+}_{\widehat{g}(\tau)\widehat{g}(\tau)}$. By assuming that at each pose switch, the error with respect to the reference pose is of the same amount, after m switches, the norm of pose estimation moved by an amount proportional to $m \left\| P^{+}_{\widehat{x}(\tau)\widehat{x}(\tau)} \right\|$, see also [2, 24].

5.5 Dealing with outliers

The desirable behavior of a feature detector is the capability of exactly tracking the same warped image regions as they are deformed along changes in view points. To do so, covariant feature detectors, such as SIFT, are designed to mod-out the effects of transformations belonging to some group. In SIFT, the features are canonized with respect to translation, rotation and scale, that is, the effects of these transformations are compensated and features become invariant to them. Canonization induces a certain amount of loss of information in the detected features, thus ambiguities could arise. Figure 5.2 shows an example where this information loss leads to a mismatch: the two highlighted image regions share almost the same appearance information even if they occur in distant areas; the only difference between them is that one is the rotated version of the other. As soon as they undergo the canonization process, involved in extracting invariant features, they can be considered being at the same scene point by the detector, leading to a mismatch. As a result, the whole set of features collected during the acquisition, matching and tracking phases may be affected by a certain amount of outliers. For the purposes of this work, we found it convenient to recast the definition of outliers in terms of the effect they have on the motion description. In particular, keypoint mismatching between the left/right frames (in a stereo vision configuration), keypoint mistracking between successive time instants and due to features belonging to moving objects on the scene will generate points which will move on the image plane in a different way with respect to the good points. A robust procedure is thus generally needed in order to guarantee a certain amount of insensitivity to the set of possible deviations from the nominal model assumptions. In this work outliers were partially rejected during the tracking phase by using a classical RANSAC approach based on epipolar geometry [10]. A maximum of 300 RANSAC steps per frame were sufficient to purge the set of image features from almost all outlying data. A further refinement can be done during estimation via simple statistics, for example by inspecting the likelihood of the $i-$th constraint, given the motion prediction. Good results were obtained by accepting for the update step those constraints which likelihood stands within a certain threshold ϕ_{th}:

$$\begin{aligned} \phi\left(\widehat{g}\left(t\right)^{-1}\widehat{g}_{ref}\left(t\right), y^{i}_{l\tau}, \widetilde{y}^{i}_{lt}\right)^{2} S^{i}_{\phi}\left(t+1\right)^{-1} &< \phi_{th} && (5.36) \\ \phi\left(g_{lr}\widehat{g}\left(t\right)^{-1}\widehat{g}_{ref}\left(t\right), y^{j}_{l\tau}, \widetilde{y}^{j}_{rt}\right)^{2} S^{j}_{\phi}\left(t+1\right)^{-1} &< \phi_{th} && (5.37) \end{aligned}$$

where the matrix $S^{i}_{\phi}\left(t+1\right)$ (the same being for $S^{j}_{\phi}\left(t+1\right)$) is the conditional variance of the $i-$th ($j-$th) constraint, defined as:

$$S^{i}_{\phi}\left(t+1\right) = H^{i}_{\phi}\left(t\right)P\left(t+1\right)^{-}H^{i}_{\phi}\left(t\right)^{T} + \widetilde{R}_{i}\left(t\right) \qquad (5.38)$$

5.6 Simulation Environment

A very large number of simulation experiments were carried out to assess the performance of the proposed approach both in ideal and degraded conditions, for example with partially known IMU-Camera and stereo pair transformations, and increasing level of noise in the inertial sensors, with different filter tuning, and along various trajectories, resulting in some thousands of tests carried out in different conditions. We present here a representative example able to generalize the results. The simulation was performed with the system moving as it was hand-held. Total travel was about 100 meters and 200 degrees (of heading). The visual features were corrupted with a zero-mean random noise, with a constant standard deviation (about 1.5 pixel of maximum error). Inertial sensor characteristics were selected to be those typical of low-cost MEMS Inertial Measurement Units, very similar to the one used in the real-world experiment. The EKF was initialized with maximum 2 degrees of attitude error, coherently to what is normally achieved with a gravity-based coarse alignment algorithm. The sensitivity parameters were all initialized with zero values. A total of 4000 (different) features where collected along the whole path and the average life-time of a reference group of features was around 40 frames. The total final error, for the presented simulation, was of about $0.5m$, corresponding to approximately 0.5% of error over path. Figures 5.3 and 5.4 show time histories of the navigation variables and estimation errors with respect to the known (simulated) values. Note that the initial misalignment between the actual attitude and the estimated one corresponds to the choice of the EKF initial conditions; this initial attitude error cannot be compensated by the filter itself due to the gauge ambiguity exposed before. The filter estimates $\widetilde{R}\,(t)$ that is offset by $\bar{R}$ (the initial attitude error in this experiment) from the actual $R\,(t)$. Analogously, the position estimation error increases proportionally to the actual position, due to such misalignment (cf. Lemma 1). The direction of the gravity is locally perturbed from the expected ($true$) vertical direction, compatibly with the initial angular misalignment.

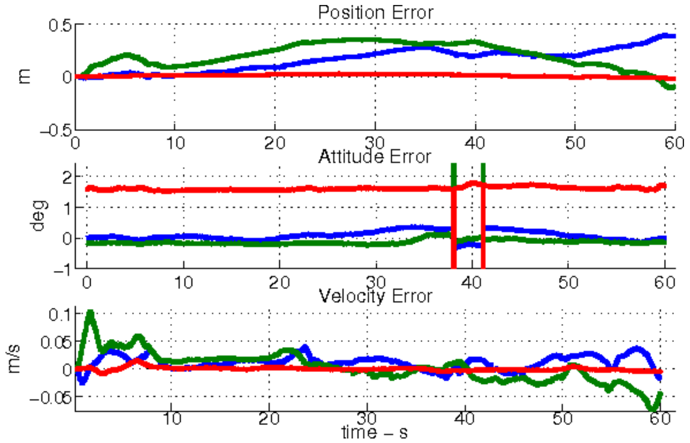
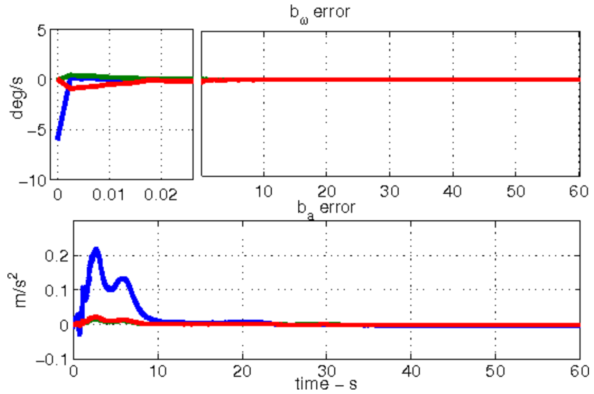
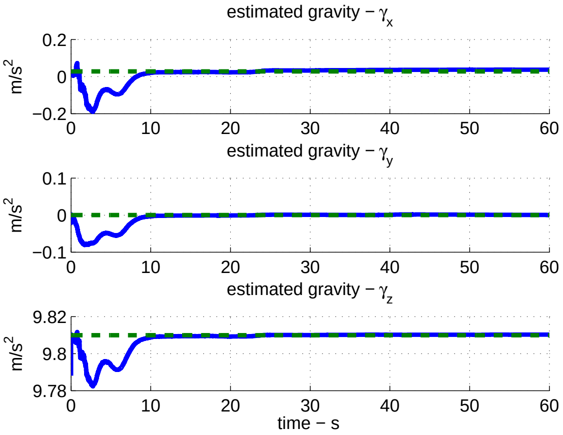


Figure 5.3: *Errors between true and estimated motion variables. Note the initial misalignment between the actual attitude and the estimated one, due to the choice of the EKF initial conditions (the almost fixed error in the attitude estimation). This misalignment cannot be compensated (as stated by the observability analysis) by the EKF. X-axis (roll), blue line; Y-axis (pitch), green line; Z-axis (yaw), red line. The spikes in the attitude error plots are due to the* $-\pi$*,* $+\pi$ *change.*



(a)



(b)

Figure 5.4: *Estimation of gravity and inertial biases. (a) Errors between true and estimated inertial sensors biases. (b) Estimation of local gravity. Note that the direction of the gravity is locally perturbed from the expected (true) vertical direction, compatibly with the initial angular misalignment: as a matter of fact, the continuous line is the estimated local gravity, whereas the dashed line is the true gravity vector, rotated by using the known initial attitude error.*

Chapter 6

Experimental results

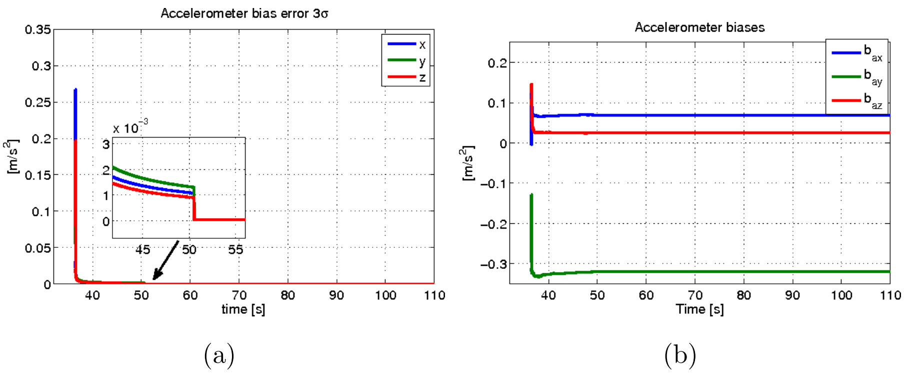
The proposed algorithm was tested using an experimental setup in a real dynamic environment. This section shows some sample experiments performed outdoor. The first test was performed in the University of Pisa, Faculty of Engineering parking lot, using a wheeled ground vehicle. A snapshot of about 80 seconds, where recognition of the actual traveled path was easier, was extracted from a longer recording. Total travel was of few hundreds of meters. A longer experiment was carried out in a typical situation of a car moving in a dynamic environment of city streets, with the presence of cars and people moving. In this case, the driver was asked to drive as naturally as possible.

Data were collected by using a stereo camera pair (Point Grey Flea2) and a low-cost low-accuracy IMU, and processed off-line to obtain an estimation of the motion variables. The IMU used for the Test 1 was the Crossbow mNav 100CA, while Test 2 was performed by using a newer Analog Devices IMU (ADIS16355). The two IMUs had comparable level of noise, while the second one ensured a better in-run bias stability, after an initial transient due to temperature stabilization. The tests were performed after this transient expired. Images were taken asynchronously from IMU measurements, at an average rate of approximately $26Hz$, while the IMU recorded the inertial measurements at a rate of $100Hz$. Visual and inertial data were successively synchronized using a common time stamp. Images resolution was 516×388 pixels. The SIFT detector [22] was used to detect and track features on the left and right images. The experimental results presented here were obtained after a coarse filter tuning.

6.1 Algorithm Initialization

Initialization of the estimation filter deserves a particular discussion, in order to better understand the points raised in section 4. As already specified, regular motions of ground vehicles do not generally meet the motion requirements for observability given by Lemma 1. In particular the gravity and the accelerometers biases cannot be disambiguated. Given the non observability issues, in order to force the observability of motion variables, we decided to saturate the state corresponding to the gravity to the value $\gamma = \begin{bmatrix} 0 & 0 & 9.81 \end{bmatrix}^T$, and to partially recover the initial misalignment by performing a gravity-based coarse alignment of the IMU, by means of the sensed accelerations in the static configuration. The filter was thus initialized with the resulting value of the attitude and the corresponding covariance was initialized with a very small value (typically $6 \div 10$

orders of magnitude smaller than the covariance of the noise). After coarse alignment, the filter started with the system standing still for the first few seconds, in order to partially estimate the accelerometer biases. When the estimation of the biases reached a locally-optimal value, i.e. the corresponding covariance of the estimation settled to a local equilibrium (as Figure 6.1(a) shows), the corresponding error covariance was saturated to a very small value and the filter was left free to evolve normally. We called this procedure *partial auto-calibration*, which performances are summarized in Figures 6.1 and 6.2. These figures refer to the first test proposed, that is the ground vehicle in the parking lot. Figure 6.1(a) shows the covariance reduction during the partial auto-calibration procedure of the accelerometers biases in static conditions. At time $50s$, the bias estimation (as Figure 6.1(b) shows) reaches a local minimum: this implies that the variation of the covariance is small enough (Figure 6.1(a)). Once such equilibrium is reached, the entries in the covariance matrix corresponding to the biases can be saturated to a very small value and kept constant along filter evolution. This prevents the filter from updating the bias estimation, which, otherwise, would result in an unconstrained walk along a subset of the unobservable subspace.



(a) (b)

Figure 6.1: *Partial auto-calibration of the accelerometers biases. (a) The 3σ bounds for the estimated errors of acceleration bias, during the partial auto-calibration procedure. The shown values are 3 times the square root of the corresponding diagonal entries of the state covariance matrix. (b) Estimation results of the acceleration biases.*

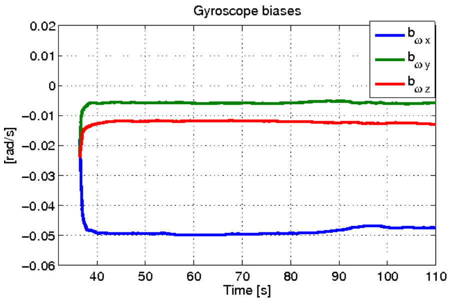


Figure 6.2: *Auto-calibration of the gyroscope biases.*

6.2 Field Tests

6.2.1 Ground vehicle in parking lot

The system was kept still in the starting position, for the first 50 seconds, then motion started. Total travel was approximately 60m and 200 degrees (of heading). Some moving objects (mainly people and tree leaves) were visible during experiment and appeared on the image plane. Moreover a dramatic illumination change and high image contrast happened after the vehicle turned 180 deg of heading to return toward the starting direction of motion. Figure 6.3 shows some snapshots collected along the path.

Figure 6.3: *Some example images from the video sequence recorded during experiment in the parking lot.*

The proposed method was compared with two pure vision-only navigation approaches. The first one is a raw stereo visual odometry, obtained by composing all the relative poses between couples of frames. In the second case, the stereo visual odometry was refined via multiframes Sparse Bundle Adjustment [20]. The same technique of Konolige [20] was used, except that the IMU was not used to correct the attitude. For both cases, the *absolute* position estimation drift was partially reduced by under-sampling the images[1]. In the second case only, the Sparse Bundle Adjustment was used before switching to the next keyframe, employing all the collected features and all the estimated relative camera poses.

Figures 6.4 and 6.5 show the estimated path for the proposed algorithm and the vision-only navigation approaches, projected on the X-Y (North-East) and X-Z (North-Down) planes. The smoothing effect performed by the Kalman filter on the noisy visual measurements is noticeable throughout the entire time range of the experiment, compared with the vision-only navigation results. Unfortunately no ground truth was available during the motion; however it is known the initial position and the approximated position of some visited areas on the map. Based on these approximated knowledge, it was computed that the error over path was approximately 1%.

[1] an average of 1 image every 20 was considered for estimating the relative pose.

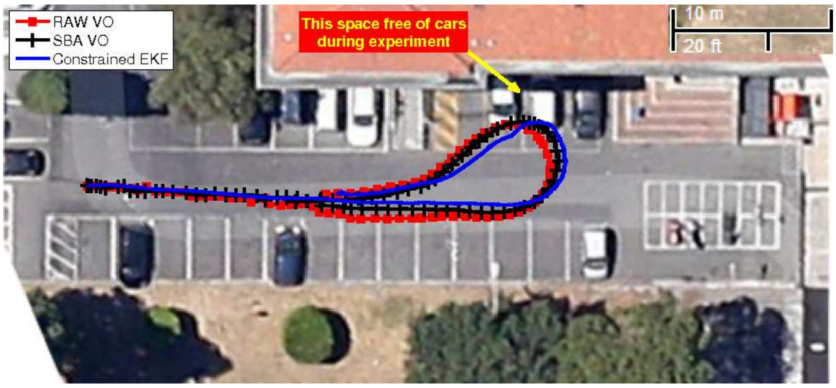


Figure 6.4: *Comparison of the reconstructed path by using the proposed approach (Constrained EKF – blue line), raw visual odometry (RAW VO – red-square line) and visual odometry with multi-frames sparse bundle adjustment refinement (SBA VO – black-cross line).*

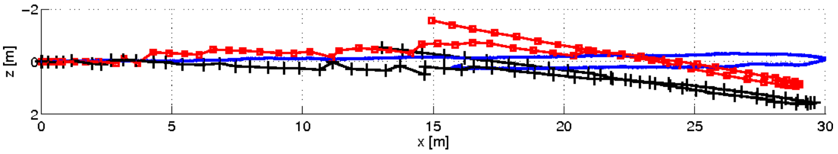


Figure 6.5: *Comparison of the drift on the vertical direction by using the proposed approach (Constrained EKF – blue line), raw visual odometry (RAW VO – red-square line) and visual odometry with multi-frames sparse bundle adjustment refinement (SBA VO – black-cross line).*

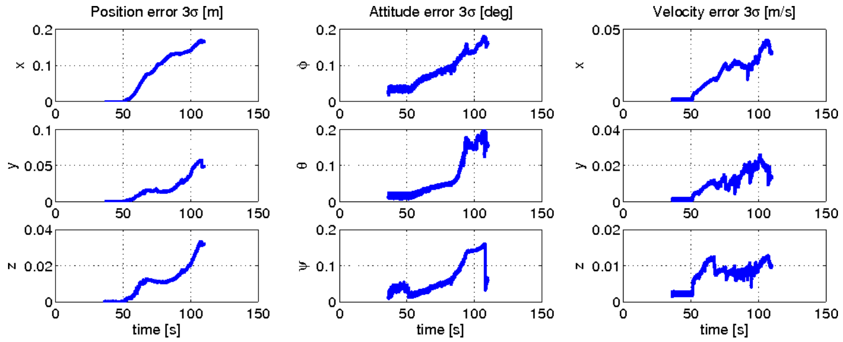


Figure 6.6: *3σ bounds for the estimated errors of navigation variables. The shown values are 3 times the square root of the corresponding diagonal entries of the state covariance matrix.*

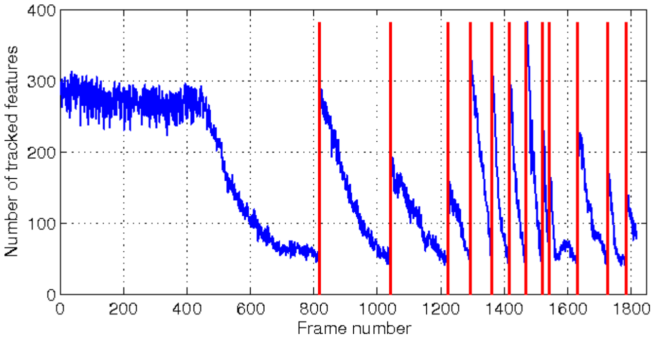


Figure 6.7: *Number of tracked features on the right image over time. The red lines indicate the images corresponding to a new group of reference features used to anchor the epipolar constraints along the path.*

6.2.2 Challenging the test: car driving in city

During this test, the acquisition system was mounted on the top of a car moving in a typical dynamical environment of city streets.

The system was kept still (with the engine on) in the starting position, for the first 30 seconds, then motion started. Total travel was approximately 1.1Km and more than 300 degrees (of heading). The maximum average velocity reached during the test was approximately $30 Km/h$. No loop closure technique was employed to refine the estimation. Several moving objects, such as people, cars, tree leaves, were visible during experiment and appeared on the image plane. Moreover, the streets were not very regular and some significantly rough areas were encountered during motion. Figure 6.8 shows the sensors suite mounted on the top of the car used for the experiments. Figures 6.9 to 6.11 show some snapshots collected along the path.

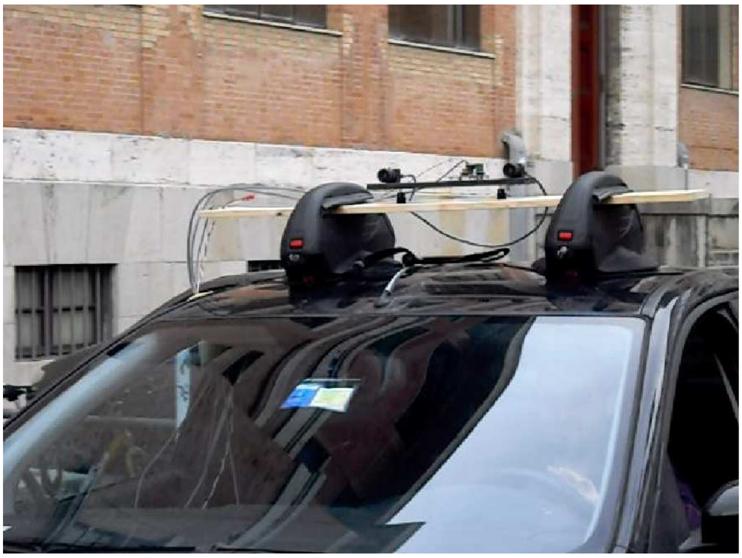


Figure 6.8: *Navigation sensors suite mounted on the top of the car used for city experiments.*

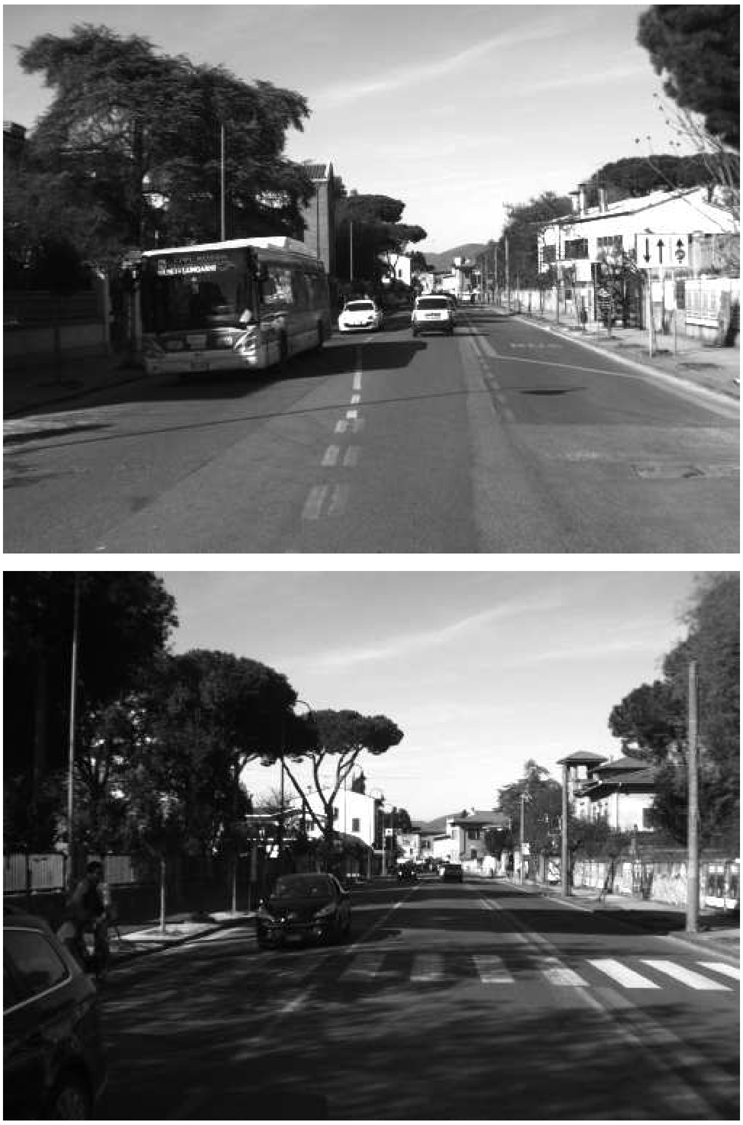


Figure 6.9: *Some example images from the video sequence recorded during the car driving in city experiment.*

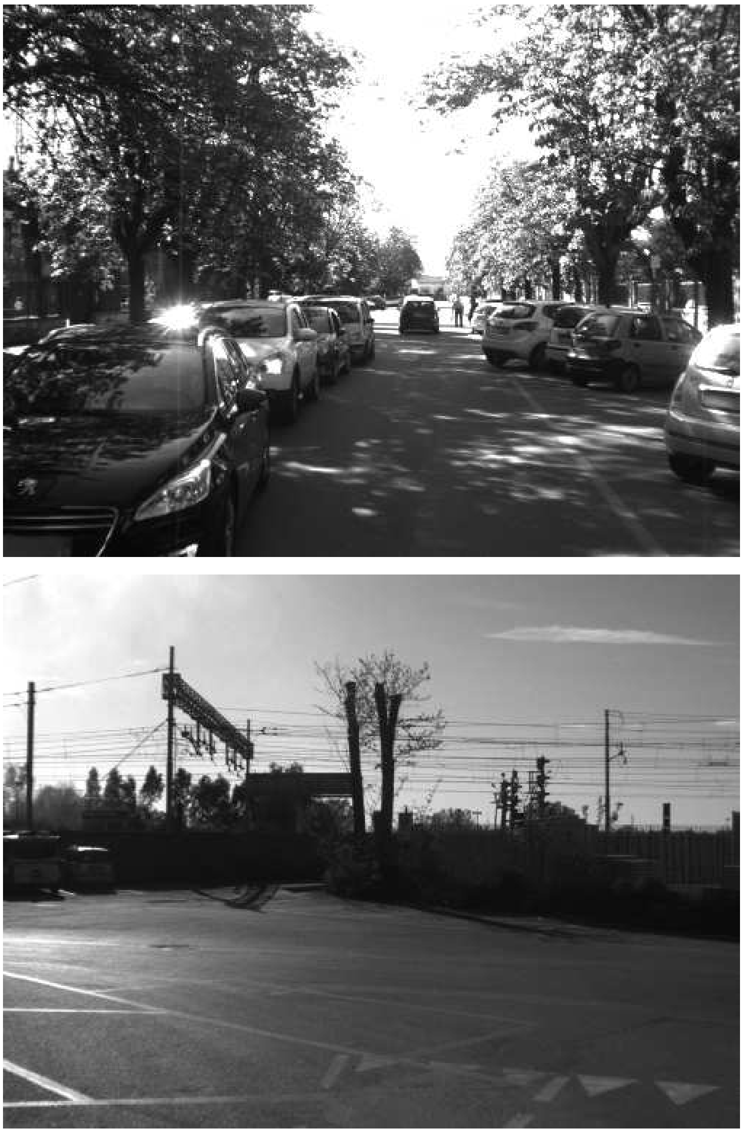


Figure 6.10: *Some example images from the video sequence recorded during the car driving in city experiment.*

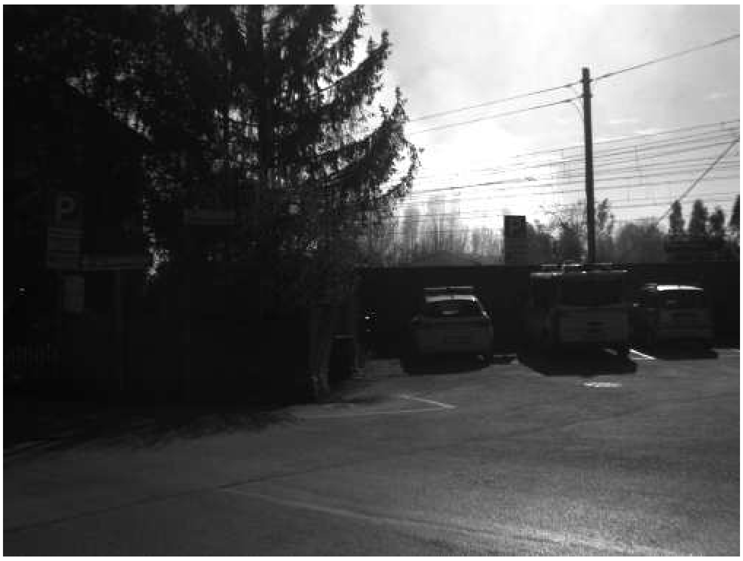


Figure 6.11: *Change of the viewpoint with respect to the last image of Figure 6.10 (previous page). Note the drastic illumination change and contrast due to the presence of the sunlight against the camera.*

The method was compared with the ground truth given by the GPS. Figure 6.14 shows the $2D$ final estimation error on the North-East plane, with respect to the final position measured by the GPS. The total error was obtained by computing the norm of the error between the GPS final position P^{gps}_{end} and the one estimated by the filter, $\widehat{P}^{ekf}_{end}$. The error over path was thus obtained via:

$$\epsilon_{path} = \frac{\left\| P^{gps}_{end} - \widehat{P}^{ekf}_{end} \right\|}{total\ length} \approx 2\% \tag{6.1}$$

The percentage error in the final position can also be computed, that is:

$$\epsilon_{\%} = \left| \frac{\|P^{gps}_{end}\| - \left\| \widehat{P}^{ekf}_{end} \right\|}{\|P^{gps}_{end}\|} \right| \approx 4\% \tag{6.2}$$

Figure 6.12 shows the comparison between the velocity estimated by the filter and the one computed employing GPS measurements. The latter did not measured velocity by itself, thus the pseudo-measurement of velocity was obtained via numerical differentiation of the measured North-East positions. Figure 6.13 shows the estimated attitude of the system during the motion. Both Figure 6.12 and 6.13 are snapshot taken from a longer sequence, where the variability of the motion was more appreciable, in particular during turns and in the middle-final part of the motion. Figure 6.15 shows the estimated trajectory and the ground truth projected onto a map of the visited area.

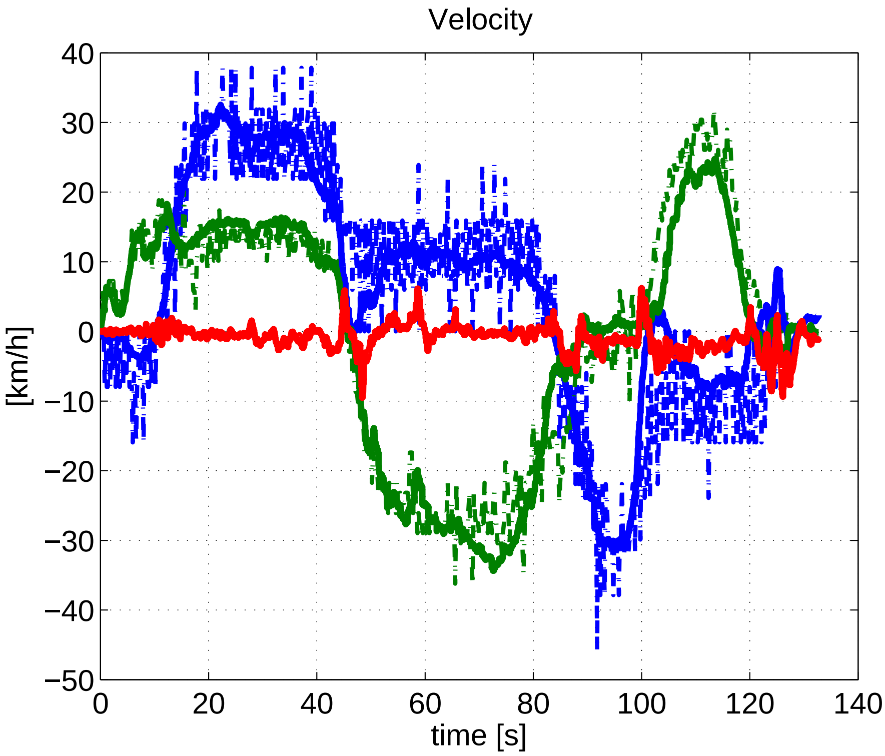


Figure 6.12: *Comparison of estimated velocity and the velocity measured by GPS. The GPS adopted did not compute the vehicle velocity, thus it was obtained via numerical differentiation of the measured North-East positions. The figure shows a snapshot taken from a longer sequence, where the variability of velocity was appreciable. Solid lines - estimated velocity (blue - V_{north}, green - V_{east}, red - V_{down}); dashed line - GPS measured velocity (blue - V_{north}, green - V_{east}).*

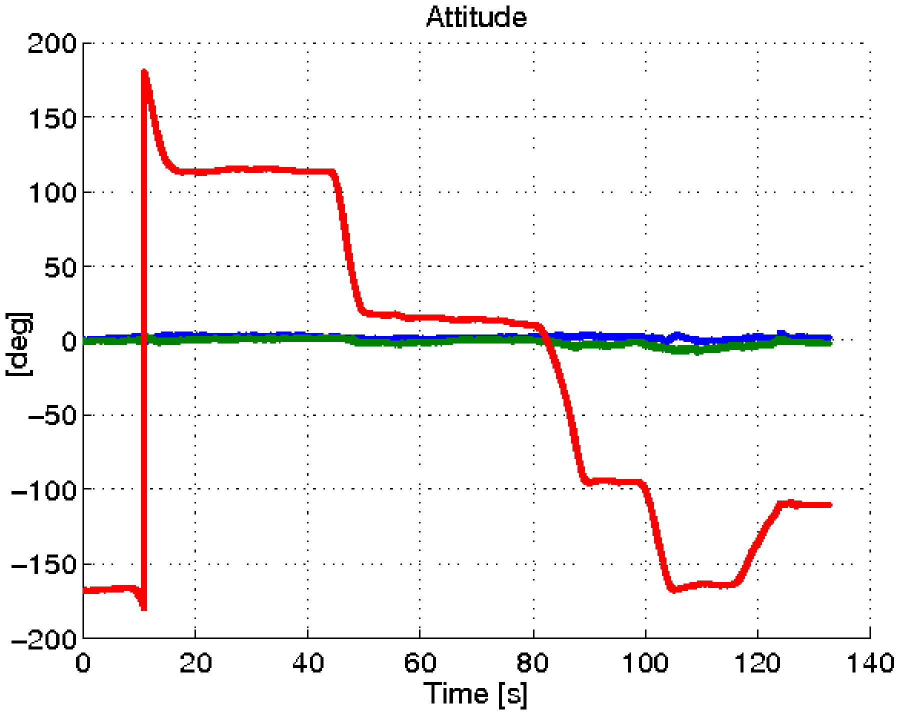


Figure 6.13: *Estimated attitude of the vehicle. The figure shows a snapshot taken from a longer sequence, where the variability of angular motion was appreciable. Blue - roll; green - pitch; red - yaw.*

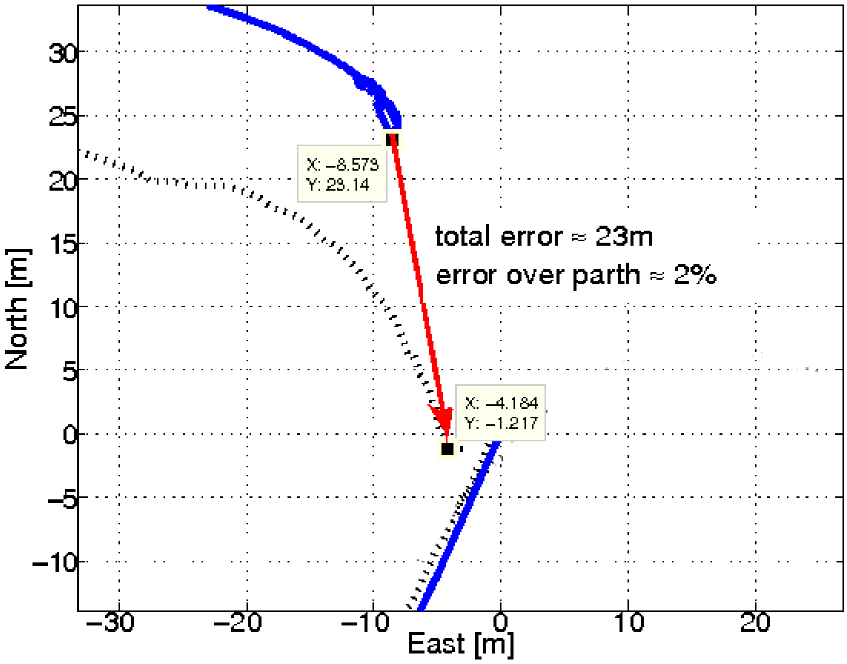


Figure 6.14: *Evaluation of the final error estimation in the North-East plane, with respect to the GPS measured path.*

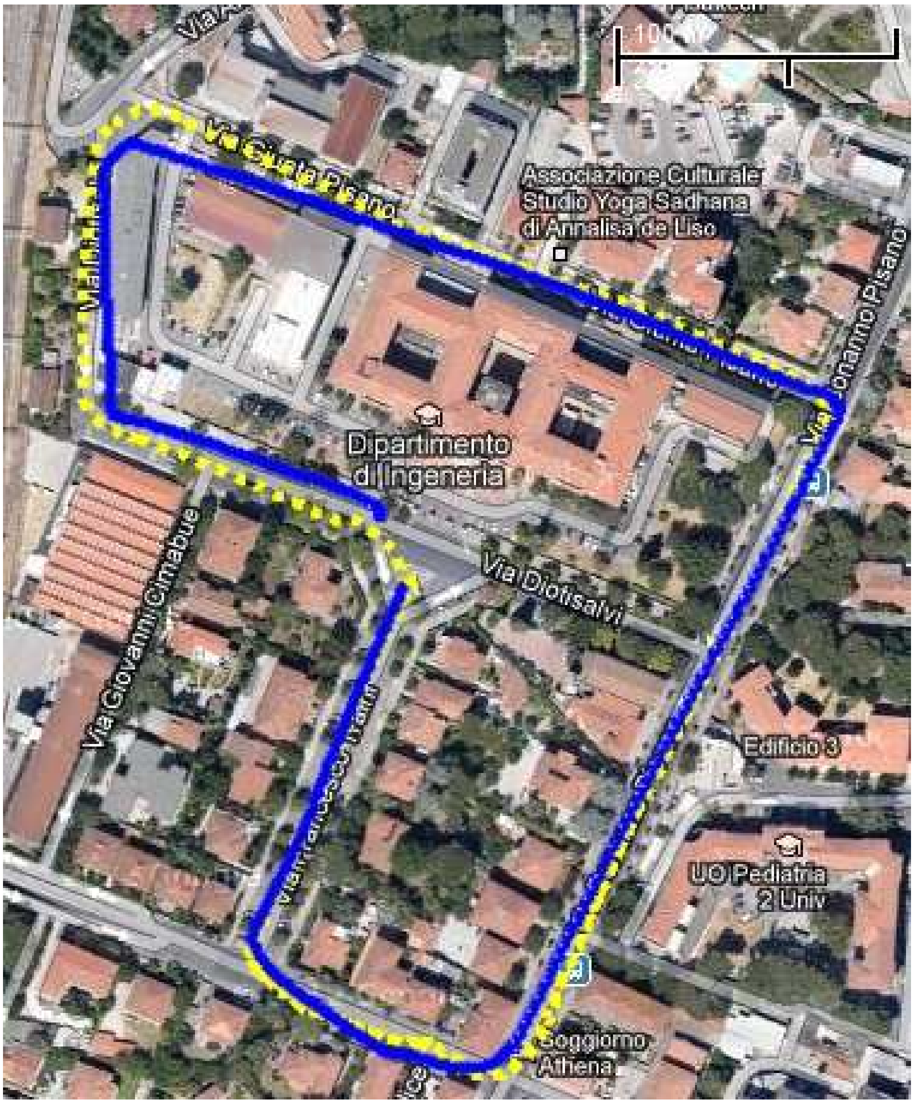


Figure 6.15: *Projection of the estimated trajectory and ground truth on the map of the visited area.*

Conclusions

We have proposed a constraints-based loose-coupling approach for the vision aided inertial navigation problem, which makes use of stereo vision and the epipolar geometry to constrain the motion of the system and correct errors in navigation. The full analytical characterization of the unobservable space of the class of constraints-based loosely-coupled problems where the one proposed falls was presented. It was shown that the existing techniques of visual-inertial navigation that rely on (features-based) visual constraints can be unified under the convergence properties highlighted. We have analyzed the conditions in which the algorithm can operate in an ideal manner, that is the motion conditions that ensure observability. The algorithms for fusing inertial measurements with visual constraints were presented, in order to solve the Ego-Motion estimation problem, showing how to define, implement the algorithm and make it work. We faced the specific problem of navigation of ground vehicles from a practical point of view and, starting from the convergence properties, we proposed the countermeasures needed in order to let the navigation algorithm work even in the motion conditions typical of road vehicles, which do not meet the requirements for observability.

The algorithm was tested both in simulation and with real data in unstructured dynamic environments, demonstrating the theoretical results.

Part II

Robust Pose Estimation

Acknowledgments

The part of the following work concerning the experimental evaluation was developed in collaboration with the PRISMA Lab, Università Federico II, Napoli and the Laboratory of Biomedical Robotics and Biomicrosystems, Università Campus Bio-Medico, Roma. F. Di Corato wishes to thank Francesca Cordella (PRISMA Lab) and Loredana Zollo (Campus Bio-Medico).

Chapter 7

Robust model-based pose estimation with unknown measurements association

This Chapter describes the robust pose estimation scheme able to estimate the relative motion – in terms of position, attitude and velocity – of a monocular vision system with respect to a tracked object of known geometry. The proposed algorithm reformulates the problem (usually solved via algebraic techniques or iterative optimizations) into a stochastic nonlinear filtering framework. It will be shown that it is robust with respect to outliers contamination of the visual data, marker disappearing and reappearing on the image plane and marker overlapping. Moreover it is able to adaptively associate a given image measurement to a certain marker or to an outlier by using probabilistic techniques, thus it is totally self-contained and requires a very rough and fast detection phase. The approach is demonstrated with *the problem of hand palm pose estimation and motion tracking during reach-and-grasp operations and the related results are presented.*

7.1 Motivation

One of the core problems in the field of applied Computer Vision is the estimation of the pose of the vision system with respect to the observed scene. Pose estimation is the concluding step in a sequence of different phases: i) detection of significant features in the scene from camera images; ii) tracking of the same set of features between successive frames; iii) estimation of the motion of the visual system between such frames, employing the collected features (i.e. pose estimation). A huge number of techniques are available in the literature, and differ each other mainly for the (most challenging) issue of ensuring a certain level of robustness to the presence of noise and outliers in the data.

A very particular subclass of pose estimation techniques are relative to the problems where the motion estimation is performed with respect to objects of known shape or geometry, for example in the framework of autonomous formation flight[33], autonomous aerial refueling [33, 45, 25, 1], relative localization with respect to known objects and/or patterns and so on. In these cases, the problem is reformulated taking advantage of the known geometry of the object being tracked. Here we concentrate on the monocular case and to the case where some markers are placed onto object surface at specified positions or, analogously, some regions of interest of discriminative appearance at known positions can be extracted. In this case, the

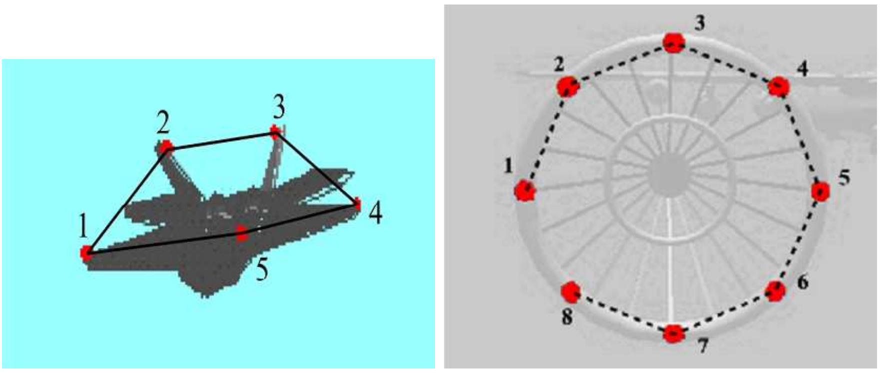


Figure 7.1: *Some synthetic examples of marker positioning for autonomous aerial refueling and formation flight [33].*

pose estimation problem is solved by recognizing the regions of interest in the image and by minimizing a given cost function, with respect to relative rotation and translation, of a certain residual built upon such measured projections. Existing and most used approaches mainly rely on iterative optimization techniques, see for instance [23, 9, 7], and generally do not provide information on relative velocity. In some particular cases, the pose estimation is made by employing active markers (mainly LEDs, remotely switched by the controller of the vision system) which improve association and outlier rejection. One examples is the VisNav system[8], used in the framework of autonomous aerial refueling and formation flight (see for example [45]).

In the most general approach, passive markers are used. The techniques of this class usually require that all the markers on the object surface are visible along the whole video stream and do not reach degenerate configurations (for example marker overlapping or occlu-

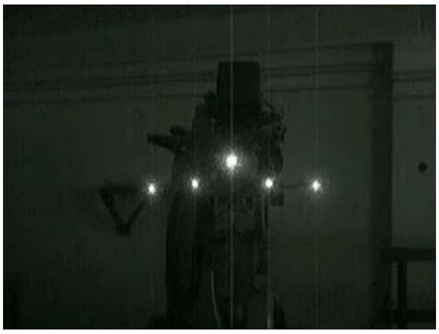


Figure 7.2: *A real example of camera tracking IR LEDs placed on the back of airplane model for pose estimation in autonomous formation flight [33].*

sions). Finally, they usually require that the association between physical markers and image measurements is known. A robust variant of the algorithm originally presented in [23] (LHM) was proposed in [33], which employs an Integer Linear Programming optimal algorithm for marker labeling, which revealed robust with respect to marker disappearing and occlusions. In all these cases, however, the velocity information is generally not available as the output of the optimization problem.

This chapter proposes a novel and robust algorithm for the monocular pose estimation of an object with known geometry. The problem is reformulated into a stochastic nonlinear filtering framework and it is robust with respect to outliers contamination of the visual data, marker disappearing and reappearing on the image plane and marker overlapping. With the term "robust" we mean that the technique is able to recognize less probable measurements and the estimation problem can still be solved even if a very low number of features (that would be non sufficient for standard algebraic algorithms) is observed. Moreover it is able to adaptively associate a given image measurement to a certain marker or to an outlier by using probabilistic techniques, thus it is totally self-contained and requires a very rough and fast detection phase, i.e. the prior association of a certain measurement is not needed to make the algorithm work.

7.2 Motion model

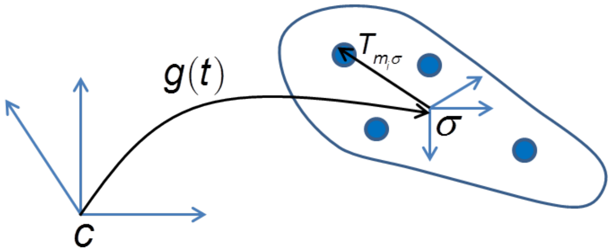


Figure 7.3: *Schematic formalization of the pose estimation problem relative to a known geometry. The Region of Interest are represented as circular markers placed onto object surface.*

Suppose to have a rigid body moving in front of a camera. Actually, we are interested in the relative motion of the system with respect to the camera, thus the camera can be assumed fixed in space or moving with respect to another reference frame, which does not change the terms of the problem. The motion of the frame σ, rigidly fixed to the body, is modeled with respect to the camera frame c according to the following continuous-time kinematic model:

$$\begin{cases} \dot{T}(t) = v(t) \\ \dot{v}(t) = \eta_v(t) \\ \dot{R}(t) = R(t)\,\Omega(t) \\ \Omega(t) = \eta_\omega(t) \wedge \end{cases} \tag{7.1}$$

where $\Omega(t)$ is the skew symmetric matrix of the body angular velocity $\eta_\omega(t)$ expressed in the coordinates of the body σ (being $\wedge$ the cross-product operator), $T(t), v(t)$ and $R(T)$ are respectively the position, linear velocity and rotation matrix of σ with respect to the frame c. Finally $\eta_v(t)$ and $\eta_\omega(t)$ are zero-mean white noises with constant variance, modeling the body linear accelerations and angular velocities as random walks. This choice is justified by the fact that we assumed not to have any prior information regarding the nature of the body motion. The variables $T(t)$ and $R(t)$ define the group transformation $g(t) \triangleq \{R(t), T(t)\} \in SE(3)$, which fully describe the 6 Degrees of Freedom localization problem of the body σ with respect to the defined reference frame.

Suppose that some markers have been placed on the body surface; the $3D$ positions of these markers with respect to the body frame is known (Figure 7.3). Assume now to measure the (noisy) projections of these markers on the image plane of the camera. The motion of the body with respect to the fixed frame c can thus be described by means of the motions of these features on the image plane. Let $T_{m_i\sigma} \in \mathbb{R}^3$ be the known position of the $i-$th marker, expressed in the coordinates of the body frame σ: its projection on the image plane can be written as:

$$y_i(t) = \pi\left(g(t)\, T_{m_i\sigma}\right) + \nu_i(t) \tag{7.2}$$

where $\pi() : \mathbb{R}^3 \to \mathbb{RP}^2$ denotes the projective operator, according to the pinhole model, and $\mathbb{RP}^2$ represents the projective space, see [24]. Moreover, $\nu_i(t)$ is a zero-mean white noise with variance R_i, assumed constant among features. When N markers are placed onto the body surface, the measurement equations can thus be written as:

$$y(t) = \begin{bmatrix} \pi\left(g(t)\, T_{m_1\sigma}\right) + \nu_1(t) \\ \pi\left(g(t)\, T_{m_2\sigma}\right) + \nu_2(t) \\ \vdots \\ \pi\left(g(t)\, T_{m_N\sigma}\right) + \nu_N(t) \end{bmatrix} \tag{7.3}$$

As in all pose estimation algorithms, the rationale is to use the measurements (7.3) as a measure of the *pose* – $g(t) \triangleq \{R(t), T(t)\} \in SE(3)$ – of the body with respect to the camera, then to use this measure to estimate the relative motion variables – position, orientation and velocity – between the body and the camera. The challenge in the proposed approach is twofold: above all the marker measurements are assumed to come in a random way, such that the association between a measurement and a physical marker cannot be made a priori; moreover we want the algorithm to be robust with respect to the presence of outliers, occlusions and markers entering and exiting from the field of view.

7.3 Robust pose estimation

7.3.1 Features detection

At this step of the problem, we assume that the parts of the image the algorithm uses as measurements are distinctive enough, such that a simple and fast feature detector can be employed at this stage. For the purposes of the work, a very raw and simple detector based on Regions Of Interests (ROIs) is sufficient. ROIs provide a complementary description of image structures in terms of regions, recommended when the information to be extracted from images belong to a certain and known class, which can be expressed in terms of colored regions, patterns, and so on. This is the case of the proposed problem also. The detector proceeds iteratively, by looking for those regions of connected pixels, which histogram is as close as possible (in the sense of Bhattacharyya similarity coefficient) to the reference histogram. The center of the detected area is then considered the location of the feature on the image plane. A robust ROI extraction and tracking has proved to be difficult, in the literature, and some detection ambiguities could raise. For example, the detector could fail in some regions due to local illumination changes or shadows. The result is the inability to detect some visible ROIs or the possibility that the features set may be contaminated by outliers. Finally some ROIs may disappear from the field of view due to body movements. For these reasons, the tracking phase in passive marker-based visual systems may be problematic and lead to an extremely tricky detection phase. Since a model of the body and some information about its shape are available, it is convenient to reformulate the tracking problem into a stochastic optimization problem, embedded into the estimation task.

7.3.2 Filtering motion and pose

According to the motion parameter dynamics in Equation (7.1), given the image-space measurements (7.3), a non-linear stochastic estimation scheme, can be implemented to estimate the state $x(t)$ of the system, consisting of the motion variables, $T(t), v(t)$ and the angular parametrization of the rotation matrix $R(t)$. For the purpose of real-time implementation, the kinematic equations (7.1) can be time-discretized using the Euler integration method. The resulting discrete time equations of the estimator are:

$$\begin{cases} T\left(t+1\right) = T\left(t\right) + v\left(t\right)dt \\ v\left(t+1\right) = v\left(t\right) + \eta_v\left(t\right)dt \\ R\left(t+1\right) = R\left(t\right)e^{\Omega\left(t\right)dt} \\ \Omega\left(t\right) = \eta_\omega\left(t\right)\wedge \\ y_i\left(t\right) = \pi\left(g\left(t\right)T_{m_i\sigma}\right) + \nu_i\left(t\right),\ i \in \mathcal{V}\left(t\right) \subseteq \left\{1,2,...,N\right\} \end{cases} \tag{7.4}$$

Note the exponential approximation used to numerically integrate the rotation matrix. The set $\mathcal{V}\left(t\right)$ denotes the group of those markers that are visible at the current time (modulo the clutters). The set is time dependent, since the physical markers may move out of the field of view or become occluded. Given the non linearity of the model with respect to the state and the orientation noise terms, a certain number of estimation schemes can be implemented, from the ones taking inspiration from the Kalman Filter (EKF, UKF, ...), to the particle filters and so forth. Although arguably other choices can be made, we are not interested in estimating the *whole* conditional density function of the state, as in particle evolution schemes, but only the *point estimate* of the state, since we expect a unimodal posterior density of the motion variables. All the deviations from the nominal model assumptions (i.e. *the tails* of the posterior) are considered to be due to clutters, and are ignored in the estimation process. This fact, together with the Gaussian nature of the model and measurements noises, motivated us to limit the discussion to nonlinear Kalman Filtering.

7.3.3 Dealing with occlusions and outliers: the association problem

When using simple detection algorithms like the one described in Section 7.3.1, render the task of associating a-priori a projection to a physical marker or deciding whether a given measurement is an outlier or a valid marker projection difficult. For this reason, we consider the outputs given by the detection algorithm, corresponding to the image at the time t, as a random sequence of M_t measurements $\mathbf{y}_t = \{y_1(t), y_2(t), ..., y_{M_t}(t)\}$ of ROI candidates. In general the condition $M_t \neq N$ holds, which means that the sequence $\mathbf{y}_t$ does contain projections of visible markers and outliers. For example, a possible situation could be the following:

$$\begin{gathered} y_1(t) \rightarrow \text{marker 3} \\ y_2(t) \rightarrow \text{outlier} \\ y_3(t) \rightarrow \text{marker 5} \\ y_4(t) \rightarrow \text{marker 1} \\ \vdots \\ y_{M_t}(t) \rightarrow \text{marker h},\ h \leq N \end{gathered} \tag{7.5}$$

The randomness of the measurement sequences is a fundamental issue in this framework, since it implies some important consequences: i) the associations between measurement h and marker j or with a clutter cannot be decided a priori and has to be estimated; ii) each sequence of measurements for each frame can be considered conditionally independent from every other sequence in the past; iii) once the current sequence of associations has been defined, it can be considered conditionally independent from the past history of associations as well. A direct consequence is that predicting the order in which markers and clutters are detected, for each image, can be very tricky. Because of the above hypotheses, we propose a solution to the filtering problem, while ensuring robustness, by employing probabilistic techniques. To this end, we use a latent variable $a_i\left(t\right)$, for each measurement $y_i\left(t\right) \in \mathbf{y}_t$, to model the measurement-to-marker or measurement-to-clutter association:

$$a_i\left(t\right) = \begin{cases} 0, \text{ if } y_i\left(t\right) \text{ is a clutter} \\ j, \text{ if } y_i\left(t\right) \text{ is the projection of marker } j \end{cases} \tag{7.6}$$

Introducing the latent variable is the same as considering the non linear model (7.3), in compact form $y\left(t\right) = h\left(x\left(t\right)\right)$, as a conditional measurement model over the variable $a_i\left(t\right)$. In fact, it is possible to condition the output function over a certain value of the latent variable: i.e. $y_i\left(t\right) = h\left(x\left(t\right)\left|a_i\left(t\right) = j \neq 0\right.\right)$, with the meaning of selecting the rows corresponding to the projection of the marker j from the function $h\left(x\left(t\right)\right)$. If $a_i\left(t\right) = 0$ the output model reduces to $y_i = \nu_o$, $\nu_o \sim \mathcal{N}\left(\bar{\nu}_o, \Sigma_o\right)$. It is desired to find the most probable value of the variable $a_i\left(t\right)$, $\forall i = 1, \ldots, M_t$, that is for every measurement collected at the current time step. The association problem can be recast as maximizing the belief that the current measurement $y_i\left(t\right) \in \mathbf{y}_t$ is either the projection of a visible marker or a clutter. Formalizing, the aim is to find the maximum of the posterior distribution:

$$p\left(a_i\left(t\right)\left|y_i\left(t\right), \mathbf{y}_{0:t-1}\right.\right) \propto p\left(y_i\left(t\right)\left|a_i\left(t\right), \mathbf{y}_{0:t-1}\right.\right) p\left(a_i\left(t\right)\right) \tag{7.7}$$

given the current measurement $y_i(t)$ and the whole history of the measurements up to the previous step. The previous equation was obtained via application of Bayes' rule. The prior $p(a_i(t))$ is assumed to be independent from the previous measurements and it is determined by the a priori knowledge of clutter and marker association event probabilities. Since no specific prior is generally available, one possible choice is to consider the probability of detecting the marker j as the same of detecting the marker $h \neq j$, i.e. by considering a uniform distribution for the marker association. Thus, one way to determine such probabilities is to infer an a priori probability of the clutter event, $p(a_i(t) = 0)$, and to equally split the complementary probability $1 - p(a_i(t) = 0)$ among the N markers association events, that is:

$$p(a_i(t) = j) = \frac{1 - p(a_i(t) = 0)}{N}, \ j = 1, \ldots, N \qquad (7.8)$$

The prior $p(a_i(t) = 0)$ is a tunable parameter and depends on the expected number of outliers with respect to the total number of measurements at each frame, i.e. on the relative frequency $\frac{N_o}{M_t}$. Other choices exist in the literature to solve the problem of estimating the foregoing prior, for example Expectation-Maximization, as in [43]. However, the explained approach showed acceptable results in experiments with real datasets and was adopted in this work. In the following the time index will be dropped for simplicity, when its disambiguation is straightforward.

The density $p\left(y_i \middle| a_i, \mathbf{y}_{0:t-1}\right)$ in equation (7.7) is the likelihood that the current measurement is associated to a given marker or to a clutter. This distribution can be obtained via marginalization of a proper joint density:

$$\begin{aligned} p\left(y_i \middle| a_i, \mathbf{y}_{0:t-1}\right) &= \int p\left(y_i \middle| x, a_i, \mathbf{y}_{t-1}\right) p\left(x \middle| a_i, \mathbf{y}_{t-1}\right) dx \qquad (7.9) \\ &= \int p\left(y_i \middle| x, a_i, \mathbf{y}_{t-1}\right) p\left(x \middle| \mathbf{y}_{t-1}\right) dx \qquad (7.10) \end{aligned}$$

where the last equality is obvious since the prediction of the motion parameters of the body does not depend on the value of the association for the current measurements set. Fixing a certain guess for the association, $a_i(t) = j$, $j \neq 0$, Equation (7.10) is the Kalman Filter likelihood of the measurement $y_i(t)$, given the prediction of the marker j, i.e. given the conditioning of the measurement model over that value of the latent variable. Thus, given the predicted state $\widehat{x}^-(t)$ and its covariance $P^-_{xx}(t)$, computed by employing the nonlinear state model, its transformation through the conditioned measurement function can be obtained, as in the classical Kalman Filtering. The mean and covariance of the predicted measurement are calculated as:

$$\begin{aligned} \widehat{y_j^-} &= h\left(\widehat{x}^-(t) \middle| a_i = j\right) \qquad (7.11) \\ P^-_{yy,j} &= H_j(t) P^-_{xx}(t) H_j^T(t) + R_j \qquad (7.12) \end{aligned}$$

where $\widehat{y_j^-}$ is the predicted projection of the marker j and $P^-_{yy,j}$ its covariance, and $H_j(t)$ is the Jacobian of the function $h\left(\widehat{x}^-(t) \middle| a_i = j\right)$ computed around the predicted state variable.

The probability of the association $a_i = j$ can be thus computed as:

$$p\left(a_i = j \middle| y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_j^-}, P^-_{yy,j}\right) p\left(a_i = j\right) \qquad (7.13)$$

being $\mathcal{N}\left(\right)$ the multivariate normal distribution of proper mean value and covariance. The set of possible associations is discrete, thus the (discrete) value of the association posterior distribution can be computed by inspecting all the possible values of the associations, that is:

$$\begin{cases} p\left(a_i = 0 \middle| y_i, \mathbf{y}_{0:t-1}\right) \propto \frac{1}{RES_u \times RES_v} p\left(a_i = 0\right) \\ p\left(a_i = 1 \middle| y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_1}^{-}, P^{-}_{yy,1}\right) p\left(a_i = 1\right) \\ \vdots \\ p\left(a_i = N \middle| y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_N}^{-}, P^{-}_{yy,N}\right) p\left(a_i = N\right) \end{cases} \tag{7.14}$$

In the case of clutter association ($a_i = 0$) the likelihood function has been set equal to $1/\left(RES_u \times RES_v\right)$, where $RES_u \times RES_v$ is the image resolution, meaning that a clutter can happen everywhere in the image. This choice is usually considered valid in approaches similar to ours, for example [36]. Selecting the maximum probability among the ones in Equation (7.14), will give the most probable value of the variable $a_i\left(t\right)$, corresponding to the measurement $y_i\left(t\right)$. The association problem is solved by repeating the above procedure for all the measurements in the set $\mathbf{y}_t$. In the following we determine the conditions that must hold to perform a the Kalman correction step, given the solution to the association problem, and how to perform such correction.

When the equation (7.14) is applied to the entire measurement set, the sequence of probabilities can be normalized and put into a matrix which we call *Feasible Association Matrix*, of the form:

$$\mathcal{F}_{M_t} = \begin{bmatrix} \pi_{10} & \pi_{11} & \cdots & \pi_{1N} \\ \pi_{20} & \pi_{21} & \cdots & \pi_{2N} \\ \vdots & & \ddots & \vdots \\ \pi_{M_t0} & \pi_{M_t1} & \cdots & \pi_{M_tN} \end{bmatrix} \tag{7.15}$$

where $\pi_{ij} = \frac{p\left(a_i=j \middle| y_i, \mathbf{y}_{0:t-1}\right)}{\sum_j p\left(a_i=j \middle| y_i, \mathbf{y}_{0:t-1}\right)}$, with the property $\pi_{ij} > 0$ and $\sum_{j=0}^{N} \pi_{ij} = 1$.

Each row in the previous matrix contains the belief for each measurement to be an outlier or the projection of each expected marker. We can introduce some definitions.

Definition 1 (Strictly Dominant Feasible Association Matrix) *The feasible association matrix $\mathcal{F}_{M_t} = \left[\ \pi_{ij}\ \right]$, is strictly dominant if for each $i = 1, \ldots, M_t$ exists one j^* such that:*

$$\pi_{ij^*} > \sum_{j \neq j^*} \pi_{ij} \tag{7.16}$$

The foregoing condition defines a feasible association matrix for which every measurement is *univocally* assigned (let's say with a probability of more than the 50%) to an outlier or to a marker.

Definition 2 (Non-degenerate Feasible Association Matrix) *The Feasible Association Matrix $\mathcal{F}_{M_t} = \left[\ \pi_{ij}\ \right]$, is non-degenerate if it is strictly dominant and*

$$\nexists j^* \neq 0 \left| \pi_{hj^*} > \sum_{j \neq j^*} \pi_{hj},\ \pi_{ij^*} > \sum_{j \neq j^*} \pi_{ij},\ \forall h \neq i \right. \tag{7.17}$$

The condition of non-degenerateness states that, while it is expected that more measurements can be classified as outliers ($j^* = 0$), two (or more) different measurements cannot be assigned to the same marker. These two definitions are useful since when the property of non-degenerateness is met, the Kalman Filter correction can be performed employing the (estimated) visible markers and their associated image projections, factoring out the measurements classified as outliers. However this cannot always happen and some ambiguities could raise, which are the cases when the feasible association matrix is degenerate, i.e. when the condition (7.17) is not met. While the condition (7.16) alone usually holds[1], as experimental tests revealed, the multiple association case is very common and some countermeasures need to be taken. The next paragraph is dedicated to this problem.

7.3.4 Solving multiple associations

It is possible that the situations where two (or more) different measurements can be assigned to the same marker arise. This is the case when, for example, two markers projections are very close each other or when the same marker is split into two (or more) different projections due, for instance, to illumination artifacts. The association optimization treats the measurements in a serial fashion and the association problem for one measurement does not take into account the associations already solved. Thus, in the degenerate situations highlighted, the solution to the association problem becomes am-

[1] Otherwise it should be hopefully possible to extract the subset of strictly dominant rows from the matrix and work with them.

biguous. Instead of changing the association algorithm into a more complex one, we propose a fast and easy a posteriori algorithm which showed very good results in tests with real data sets. In the case of multiple associations, define the set $\mathcal{H} = \{h\}$ of all the indexes such that:

$$\exists j^* \neq 0 \, \Big| \, \pi_{hj^*} > \sum_{j \neq j^*} \pi_{hj}, \; \forall h \in \mathcal{H} \tag{7.18}$$

We propose to disambiguate the association by simply taking the maximum among all the π_{hj^*}, $h \in \mathcal{H}$ and associate the marker j^* to the measurement which probability π_{hj^*} has the maximum value. For all the remaining $h \in \mathcal{H}$ we force the association to an outlier: $\pi_{h0} = 1$ and $\pi_{hj} = 0, \forall j = 1, \ldots, N$.

7.3.5 Extended models for articulated bodies

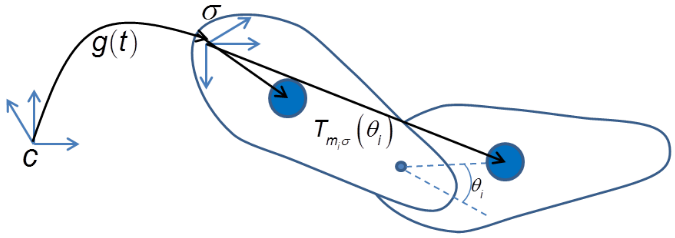


Figure 7.4: *Schematic representation of an articulated body. The Region of Interest are represented as circular markers placed onto object surface.*

The robust pose estimation algorithm explained so far can be easily adapted to the case of articulated bodies without big effort. In this case, the body is intended as the interconnection of a certain number of rigid bodies, which can move with respect to each other via the actuation of proper joint variables. The Figure 7.4 shows an example of articulated body, composed by two rigid bodies which relative motion is actuated via a rotational joint. The variable θ_i models the relative angular displacement of each rigid body with respect to the one which comes before. Classical examples of articulated bodies are robotic manipulators, the human body, the human hand and so forth. In this case, the position of the i-th ROI with respect to the body frame σ depends on the value of the joint variables θ_i and can be determined via classical direct kinematics, parametrized via Denavit-Hartenberg parameters [30]. If a measurement of the joint parameters is not available (for example robot manipulators employ encoders to measure joint displacements), it is necessary to extend the state space with the joint variables too and a possible model becomes:

$$\begin{cases} \dot{T}\left(t\right) = v\left(t\right) \\ \dot{v}\left(t\right) = \eta_v\left(t\right) \\ \dot{R}\left(t\right) = R\left(t\right)\Omega\left(t\right) \\ \dot{\theta}_i\left(t\right) = \eta_{\theta_i}\left(t\right) \\ \Omega\left(t\right) = \eta_\omega\left(t\right)\wedge \end{cases} \tag{7.19}$$

The output function is still the projection of the ROIs, but now their positions with respect to the body frame are functions of the joint variables $\Theta = \begin{bmatrix} \theta_1 & \dots & \theta_M \end{bmatrix}^T$, obtained via direct kinematic:

$$y_i\left(t\right) = \pi\left(g\left(t\right)T_{m_i\sigma}\left(\Theta\right)\right) + \nu_i\left(t\right) \tag{7.20}$$

In this case, the estimation loop is designed to estimate the pose of the body reference frame with respect to the camera frame *and* the pose of the articulated body itself, that is the value of the joint variables. Such a model, together with the robust estimation scheme presented above, can be useful, for example, in the pose estimation of the human hand by employing low cost sensors, instead of more expensive and cumbersome infrastructures like multi-cameras systems or active markers.

7.4 Applications

The proposed algorithm has been experimentally tested in the case of pose estimation of the human wrist during angular movements and grasp operations. The person who performed the test was seated in front of a table with the right hand placed on it. In the starting configuration of the hand, all the fingers were fully extended and the wrist was in a neutral position. The subject was asked to perform some wrist movements and to reach and grasp an object located on the table. In both the experiments, the subject paid a special attention at not rotating the arm.

7.4.1 Experimental Setup

In order to have information about the wrist pose (position, orientation and velocity) during motion, six colored markers were positioned on the human hand, as shown in (Figure 7.5). The markers were made of blue paper of diameter 1.2 cm. The protocol for positioning markers on the hand was chosen in order to minimize artifacts, due for example to skin movements or marker occlusion, to have the center of the markers approximately aligned with the CoR of the

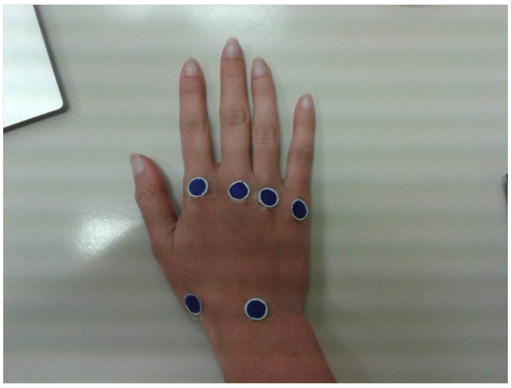


Figure 7.5: *Protocol used for markers positioning.*

corresponding joints and to easily evaluate the anatomy parameters of the wrist (relative positions of the markers), which determination strongly affects the accuracy of the estimation. In this implementation, the hand palm is considered a non-articulated body and the body reference frame is positioned on the marker placed on the wrist. Markers detection is made via a very raw blob detection algorithm, scanning the image row-wise and extracting the regions of interest characterized by a certain minimum number of blue pixels. The visual system used was the Asus Xtion Prolive visual sensor suite, which is a motion sensing device consisting of an InfraRed (IR) laser emitter, an IR camera for measuring depth information, and a RGB camera. The chosen resolution of the RGB camera was 640 × 480. OpenNI library has been used in order to make the Asus work on the PC and the whole algorithm has been implemented under ROS (Robotic Operating System) for ensuring a real-time approach. For the proposed experiments, the depth information was used during initialization only. In the proposed experiments the association of the measurements was unknown.

7.4.2 Estimation

According to the motion parameter dynamics in Equation (7.1), given the image-space measurements (7.3) (with $N = 6$), a nonlinear estimation scheme was designed. The aim of the filter is to estimate the state $x\left(t\right)$ of the system, consisting of the motion variables, $T\left(t\right), v\left(t\right)$ and the angular parametrization of the rotation matrix $R\left(t\right)$, which reflect the pose of the hand palm with respect to the camera frame. In this case an estimation scheme based on the Unscented Kalman Filter [17] was selected. In particular, given the non linearity of the kinematic model with respect to the state and the orientation noise terms, the Augmented Unscented Kalman Filter algorithm presented in [17] was used. The peculiarity of the adopted estimation scheme, compared with the classical UKF approach [46], is the possibility to easily deal with non-affine noise terms in the state/measurement model. For the remaining part, the technique is a classical UKF as in [46]. The full algorithm can be found in [17] and it will be omitted for brevity. It is worth to mention how the equations of the predicted measurements and their covariance (eq. (7.11) and (7.12)) are modified in the case of Unscented Filtering. In this case, given the predicted state-related Sigma-Points [17], $\mathbf{X}^{x}_{n,t/t-1}$, $n = 1, \ldots, L$, computed by employing the nonlinear state model, their transformation through the conditioned measurement function can be obtained, as in the classical UKF:

$$\mathbf{Y}^{j}_{n,t/t-1} = h\left(\mathbf{X}^{x}_{n,t/t-1} \middle| a_i = j\right) \tag{7.21}$$

The superscript j on the transformed Sigma-Points of the output, indicates that $Y^j_{n,t/t-1}$ refers to the predicted projection of the marker j, for which the association is being tested. The mean and covariance of the measurement vector are calculated as:

$$\widehat{y_j^-} = \sum_{n=0}^{L} W_m^n \mathbf{Y}^j_{n,t/t-1} \tag{7.22}$$

$$P^-_{yy,j} = \sum_{n=0}^{L} W_c^n \left(\mathbf{Y}^j_{n,t/t-1} - \widehat{y_j^-}\right)\left(\mathbf{Y}^j_{n,t/t-1} - \widehat{y_j^-}\right)^T + R \tag{7.23}$$

where W_c^n and W_m^n are the weights associated to the Sigma-Points [17], $\widehat{y_j^-}$ is the predicted projection of the marker j and $P^-_{yy,j}$ its covariance. Thus, the probability of the association $a_i = j$ (eq. (7.10)) can be computed as:

$$p\left(a_i = j \middle| y_i, \mathbf{y}_{0:t-1}\right) \propto \mathcal{N}\left(y_i - \widehat{y_j^-}, P^-_{yy,j}\right) p\left(a_i = j\right) \tag{7.24}$$

7.4.3 Filter initialization

The initialization phase is responsible of the estimation of the initial relative pose between the camera and the reference frame on the wrist, and needs to be reasonably accurate. For this reason, the estimation is formulated as a Least-Squares optimization problem. During this phase, the marker are required to be visible, such that the association between markers and measurements can be made without effort, after the detection phase. Therefore, no probabilistic optimization needs to be carried out. Finally, the hand must be in neutral configuration. The measurements employed during the initialization phase are the projection of the markers on the image plane and the measurement of their depth, relative to the camera, which are obtained via the available IR camera. The equation mapping the available measurements into the estimation variables are:

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} \pi\left(g\left(\theta\right) T_{wi}\right) \\ \vdots \\ e_3^T \left(g\left(\theta\right) T_{wi}\right) \\ \vdots \end{bmatrix} = \begin{bmatrix} h_y\left(g\left(\theta\right)\right) \\ h_z\left(g\left(\theta\right)\right) \end{bmatrix} \tag{7.25}$$

$$\bar{y} = \bar{h}\left(g\left(\theta\right)\right) \tag{7.26}$$

where T_{wi} are the position of the markers with respect to the reference frame placed on the wrist. Note that the relative transformation $g\left(\theta\right)$, between the camera and the wrist, is parametrized via $\theta \in \mathbb{R}^6$, which encodes the unknown pose parameters (translation and angular parametrization) to be estimated. y_i and z_i are respectively the measured projections and depths of the markers. Finally $e_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$. The locally optimal estimation of the foregoing transformation is found by minimizing the 2-norm cost function:

$$\min_{\theta} \|\bar{y} - \bar{h}\left(g\left(\theta\right)\right)\|^2 \tag{7.27}$$

The linearization of the nonlinear function $\bar{h}\left(g\left(\theta\right)\right)$ around an initial estimation of the pose parameter θ_0, gives:

$$\begin{aligned} J_\theta &= \|\bar{y} - \bar{h}\left(g\left(\theta\right)\right)\|^2 && (7.28) \\ &\approx \|\bar{y} - \bar{h}\left(g\left(\theta_0\right)\right) - H_{\theta_0}\left(\theta - \theta_0\right)\|^2 && (7.29) \\ &= \|\widetilde{y} - H_{\theta_0}\theta\|^2 && (7.30) \end{aligned}$$

In the previous equation, $H_{\theta_0} = \left.\frac{\partial \bar{h}}{\theta}\right|_{\theta_0}$. Equation (7.30) is a well known linear quadratic cost function, which minimum is obviously given by:

$$\widehat{\theta} = H_{\theta_0}^{\dagger} \widetilde{y} \tag{7.31}$$

That is, expanding the solution, the optimal estimation of the relative pose between the camera and the wrist reference frames, at the initial time, is given:

$$\widehat{\theta} = R_H \theta_0 + H_{\theta_0}^{\dagger} \left(y - \bar{h} \left(g \left(\theta_0 \right) \right) \right) \tag{7.32}$$

where $R_H = H_{\theta_0}^{\dagger} H_{\theta_0}$ is the range-space projector of the matrix H_{θ_0}. Figure 7.6 shows the results of the initialization phase. The initial pose parameter is used to propagate the UKF sigma-points through the output function (yellow dots). The weighted point average of the sigma-points define the estimated position of the marker projections (blue circles), after the initialization phase. The spreading of the sigma-points are related on the confidence (initial covariance matrix) of the initial pose parameters.

7.4.4 Pose Estimation results

This section summarizes some results of the proposed robust pose estimation algorithm by using measurements which association was unknown and solved via the proposed probabilistic association method. In the first experiment, the subject was asked to perform some smooth angular movements of the wrist, spanning the three degrees of freedom. This experiment is referred to as *Range of Motion* experiment. Figures 7.7 and 7.8 show some results.

In the second test, the subject was asked to approach and grasp an object located on the table. In this case, the goal of the estimation task was to extract position and velocity information from the video

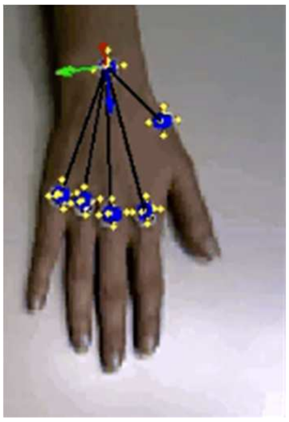


Figure 7.6: *UKF sigma-points after filter initialization (yellow dots). The blue circles represent the estimated marker position after pose initialization.*

sequence. Figure 7.9 to 7.12 are related to the latter experiment and show the reached results. Both the test were performed in real-time under the ROS environment.

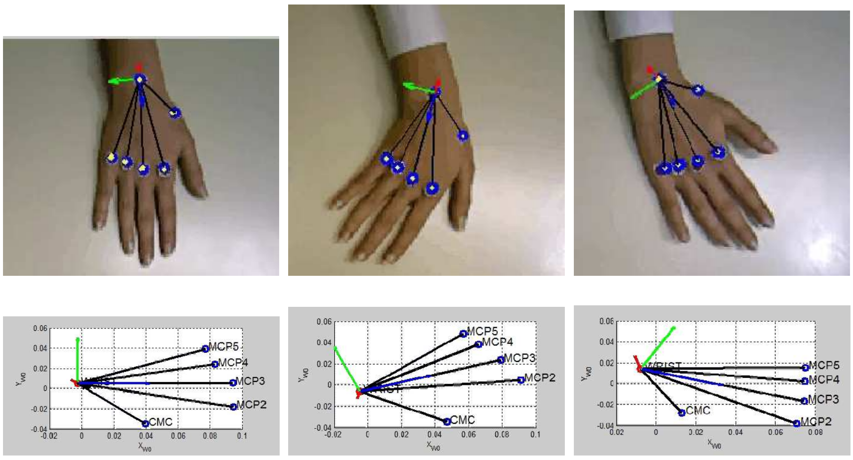


Figure 7.7: *Some example images from the video sequence recorded during the range of motion experiment and the related pose estimations projected onto the X-Y plane (aligned with the desk). The blue circles (on the hand figure - top images) are the estimated (after correction) marker positions projected onto the image space; they are connected with the marker on the wrist via black lines. The yellow dots are the estimated (after correction) UKF sigma points, evaluated through the output function. The bottom figures show the estimated hand pose in the 3D space. 3D marker positions are labeled from the thumb related marker (CMC in Figure) to little finger related marker (MCP5 in Figure).*

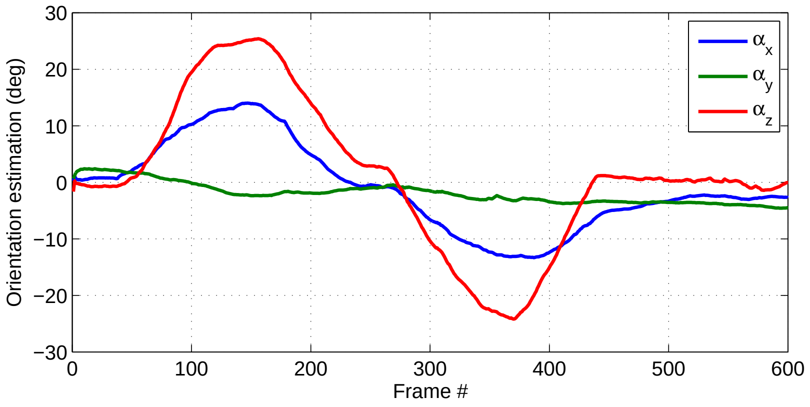


Figure 7.8: *Estimation of the angular movements for the Range of Motion experiment.*

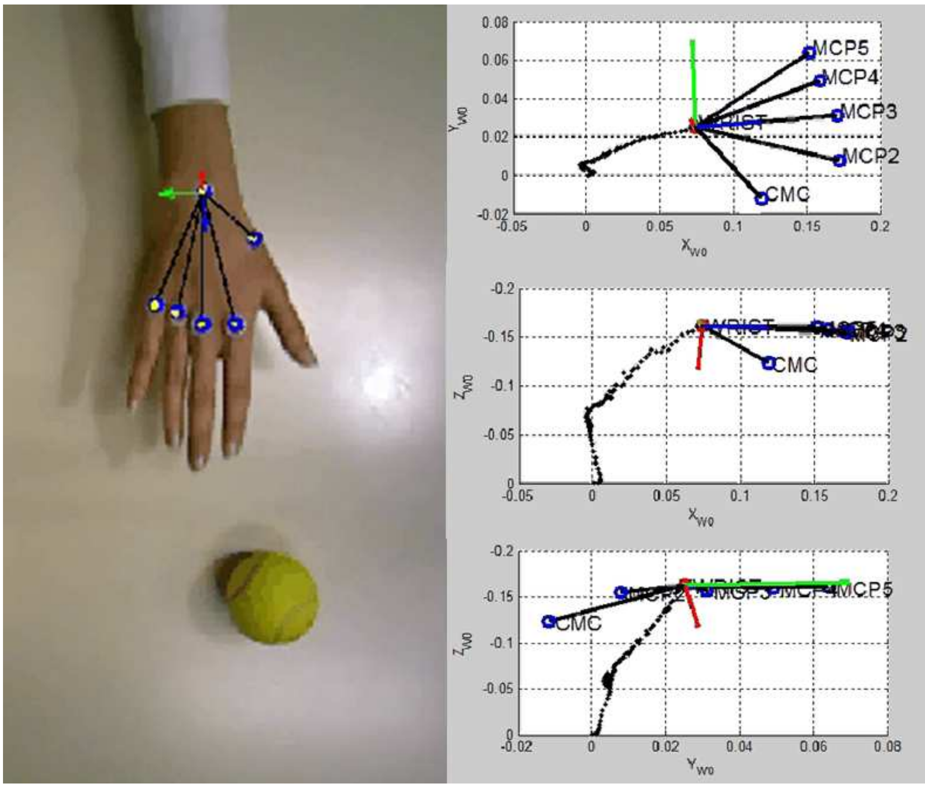


Figure 7.9: *An example image taken during the reaching-and-grasp experiment and the related position estimations projected onto the X-Y, Y-Z and X-Z planes. The blue circles (on the hand figure - left image) are the estimated (after correction) marker positions projected onto the image space; they are connected with the marker on the wrist via black lines. The yellow dots are the estimated (after correction) UKF sigma points, evaluated through the output function. The right figures show the estimated hand pose in the 3D space. 3D marker positions are labeled from the thumb related marker (CMC in Figure) to little finger related marker (MCP5 in Figure).*

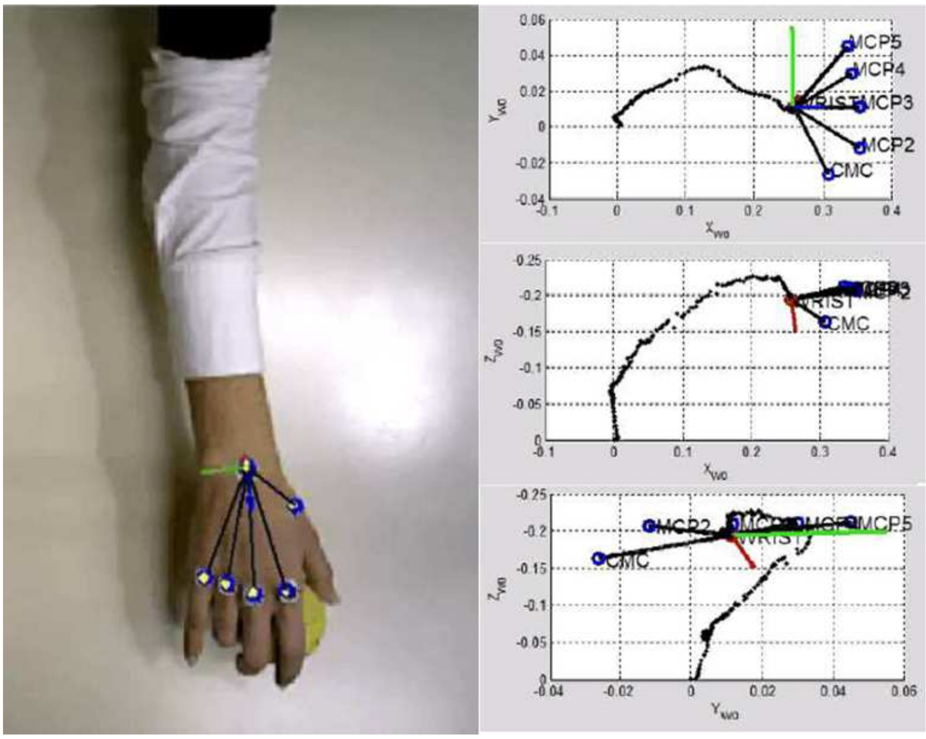


Figure 7.10: *Last image of the reaching-and-grasp experiment and the related position estimations projected onto the X-Y, Y-Z and X-Z planes. The blue circles (on the hand figure - left image) are the estimated (after correction) marker positions projected onto the image space; they are connected with the marker on the wrist via black lines. The yellow dots are the estimated (after correction) UKF sigma points, evaluated through the output function. The right figures show the estimated hand pose in the 3D space. 3D marker positions are labeled from the thumb related marker (CMC in Figure) to little finger related marker (MCP5 in Figure).*

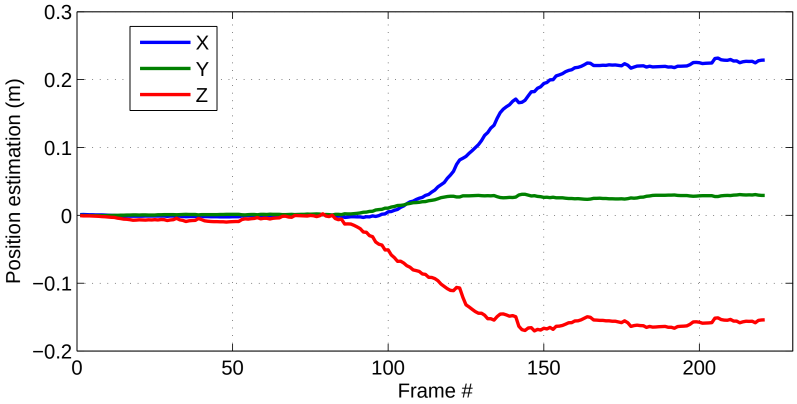


Figure 7.11: *Full position estimation of the wrist in the reaching-and-grasp experiment.*

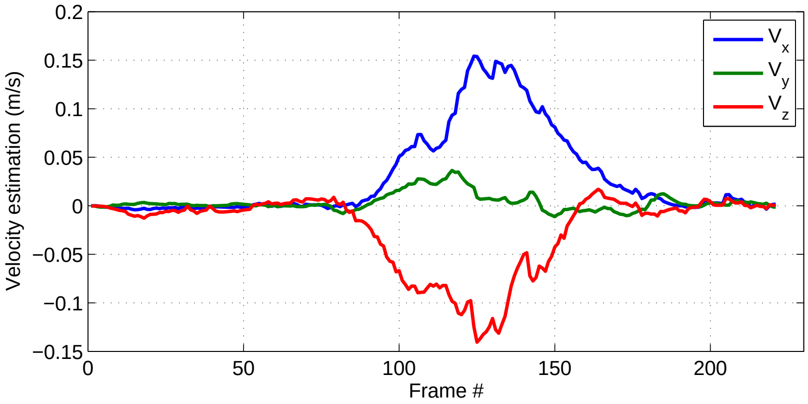


Figure 7.12: *Full velocity estimation of the wrist in the reaching-and-grasp experiment.*

Part III

Closure

Appendix A

Proof of observability

Non-linear observability study can be made in several ways: here we use an analytical approach [16], that is by describing the complete set of initial conditions which render the available measurements identical.

We look for all the initial conditions of the state variables (and associated movements of the states) that produce exactly the same outputs. Let's denote with a *tilde* hat the generic movement of the state variables obtained starting from an initial condition which is different from the *true* one. Without loss of generality we can consider the reference time, which was generally called τ in the previous sections, to correspond to the initial time, $\tau = 0$. Moreover, we assume to work in the case of non-degenerate epipolar constraints, that is relative translation different from 0. Thus we aim at determining all the possible values that the variables $\widetilde{T}(t)$, $\widetilde{v}(t)$, $\widetilde{a}(t)$, $\widetilde{R}(t)$, $\widetilde{\gamma}$, $\widetilde{b}_{\omega}$, $\widetilde{b}_{a}$ and the global scale α can take which render the available outputs identical:

$$\left[\begin{array}{c} \widetilde{R}^T(t)\left(\widetilde{a}(t)+\widetilde{\gamma}\right)+\widetilde{b}_a \\ \widetilde{\omega}(t)+\widetilde{b}_\omega \end{array}\right] = \left[\begin{array}{c} R^T(t)\left(a(t)+\gamma\right)+b_a \\ \omega(t)+b_\omega \end{array}\right] \tag{A.1}$$

$$\phi\left(\widetilde{g}(t)^{-1}\widetilde{g}(0), y_{l0}^i, y_{lt}^i\right) = \phi\left(g(t)^{-1}g(0), y_{l0}^i, y_{lt}^i\right) \tag{A.2}$$

$$\phi\left(g_{lr}\widetilde{g}(t)^{-1}\widetilde{g}(0), y_{l0}^i, y_{rt}^j\right) = \phi\left(g_{lr}g(t)^{-1}g(0), y_{l0}^i, y_{rt}^j\right) \tag{A.3}$$

Proof of Claim 4. It is known that the epipolar constraints are invariant under scaling transformation (ref. (3.3), i.e.:

$$\begin{aligned}\phi\left(g(t)^{-1}g(0), y_{l0}^i, y_{lt}^i\right) &= \phi\left(g(t)^{-1}g(0), \lambda y_{l0}^i, \lambda y_{lt}^i\right) \\ &= \phi\left(\lambda^2 g(t)^{-1}\lambda^2 g(0), y_{l0}^i, y_{lt}^i\right)\end{aligned} \tag{A.4}$$

Putting (A.2), (A.3) and (A.4) together and choosing $\alpha = \lambda^2$, we get:

$$\phi\left(\widetilde{g}(t)^{-1}\widetilde{g}(0), y_{l0}^i, y_{lt}^i\right) = \phi\left(\alpha\left(g(t)^{-1}g(0)\right), y_{l0}^i, y_{lt}^i\right) \tag{A.5}$$

$$\phi\left(g_{lr}\widetilde{g}(t)^{-1}\widetilde{g}(0), y_{l0}^i, y_{rt}^j\right) = \phi\left(\alpha\left(g_{lr}g(t)^{-1}g(0)\right), y_{l0}^i, y_{rt}^j\right) \tag{A.6}$$

this holds provided that:

$$\begin{cases} \widetilde{g}(t)^{-1}\widetilde{g}(0) = \alpha\left(g(t)^{-1}g(0)\right) \\ g_{lr}\widetilde{g}(t)^{-1}\widetilde{g}(0) = \alpha\left(g_{lr}g(t)^{-1}g(0)\right) \end{cases} \tag{A.7}$$

$\forall \alpha \in \mathbb{R}$. The group transformation $\widetilde{g}(t)$ that makes the first of Equation (A.7) hold is

$$\begin{cases} \widetilde{g}(t) = \alpha\bar{g}(t)\,\alpha g(t); \\ \widetilde{g}(0) = \alpha\bar{g}(0)\,\alpha g(0); \end{cases} \tag{A.8}$$

where $\bar{g}(\cdot)$ is a suitable group transformation. We made it depending on time since we still do not know if such transformation is time-varying or not. The first equality in equation (A.7) becomes:

$$\alpha\left(g(t)^{-1}\,\bar{g}(t)^{-1}\,\bar{g}(0)\,g(0)\right) = \alpha\left(g(t)^{-1}\,g(0)\right) \tag{A.9}$$

which holds $\forall\alpha$ and by choosing $\bar{g}(t) = \bar{g}(0)$, which can be proven by substitution. Thus (A.2) holds for any α and for any $\alpha\bar{g}\,\alpha g(t)$, given an arbitrary (constant) $\bar{g}$.

The fact that Equation (A.9) holds for any α can be interpreted as the classical scale ambiguity of monocular vision, as a matter of fact we derived this result using only the epipolar constraint on left camera points. Enforcement of the second group of epipolar constraints (between left and right cameras) eliminates the scale ambiguity. Indeed, with such conditions and Equation (A.8), the second equality in Equation (A.7) becomes[1]:

$$g_{lr}\alpha\left(g(t)^{-1}\,\bar{g}^{-1}\bar{g}g(0)\right) = \alpha\left(g_{lr}g(t)^{-1}\,g(0)\right) \tag{A.10}$$

which holds *iff* $\alpha = 1$ and for every arbitrary $\bar{g} \in SE(3)$. This can be proven by extracting the translation component of the transformations in Equation (A.10), and by showing that the condition (A.10) simply means that:

$$T_{lr} = \alpha T_{lr} \tag{A.11}$$

which holds *iff* $\alpha = 1$, since g_{lr} is assumed known. The two conditions highlighted so far state that 1) the scale is always recoverable,

[1] The scaling term α is the same for the constraints on the left and right cameras since they share the points y^i_{l0}.

since the stereo extrinsic parameters are assumed to be known; 2) the ambiguity $\bar{g}\,(t)$ is actually constant, that is $\bar{g}\,(t) = \bar{g}$. This ambiguity corresponds to the classical *gauge* ambiguity, related to the choice of the initial conditions [16]. □

As expected, the knowledge of the relative transformation between the left and right cameras is sufficient to disambiguate the global scale factor; however vision alone is not sufficient to render the system locally observable. The *gauge* ambiguity and the sensitivity parameters, together with the gravity remain unobservable. It is possible to reduce the unobservable set by using the measurements of the IMU. It is convenient to make explicit the rotational and translational components of the unobservable group transformation $\widetilde{g}\,(t) = \bar{g}g\,(t)$, which will be used in the remainder of the chapter:

$$\widetilde{R}\,(t) = \bar{R}R\,(t)\,,\ \forall t \geq 0 \tag{A.12}$$

$$\widetilde{T}\,(t) = \bar{R}T\,(t) + \bar{T},\ \forall t \geq 0 \tag{A.13}$$

Proof of Claim 5. The condition in (A.1)

$$\widetilde{\omega}\,(t) + \widetilde{b}_\omega = \omega\,(t) + b_\omega \tag{A.14}$$

is not sufficient alone to recover unique $\widetilde{b}_\omega, \widetilde{\omega}\,(t)$. In fact, by choosing $\widetilde{\omega}\,(t) = \omega\,(t) + \bar{\omega}$, $\forall$ constant $\bar{\omega}$, will make $\widetilde{b}_\omega = b_\omega + \bar{\omega}$ which is still feasible, since b_ω and $\widetilde{b}_\omega$ are constant by means of the model.
If we proceed with the time derivative of the first and second term

of the equality in Equation (A.12), we get:

$$\dot{\widetilde{R}}(t) = \bar{R}\dot{R}(t) \tag{A.15}$$

$$\widetilde{R}(t)\,\widetilde{\omega}(t)\wedge = \bar{R}R(t)\,\omega(t)\wedge \tag{A.16}$$

$$\bar{R}R(t)\,\widetilde{\omega}(t)\wedge = \bar{R}R(t)\,\omega(t)\wedge \tag{A.17}$$

that is $\widetilde{\omega}(t) \equiv \omega(t)$, which means $\widetilde{b}_\omega = b_\omega$, i.e. the angular velocity is observable, as well as the gyroscope biases. □

Proof of Lemma 1 and Claim 6. We proceed with the time derivatives of the translational component in Equation (A.13) [16]:

$$\dot{\widetilde{T}}(t) = \widetilde{v}(t) = \bar{R}v(t) \tag{A.18}$$

$$\dot{\widetilde{v}}(t) = \widetilde{a}(t) = \bar{R}a(t) \tag{A.19}$$

We use Equation (A.1) and substitute the found value of $\widetilde{a}(t) = \bar{R}a(t)$ and $\widetilde{R} = \bar{R}R(t)$, i.e.:

$$R^T(t)\,\bar{R}^T\left(\bar{R}a(t) + \widetilde{\gamma}\right) + \widetilde{b}_a = R^T(t)\,(a(t) + \gamma) + b_a \tag{A.20}$$

Thus

$$\widetilde{\gamma} = \bar{R}\left(\gamma + R(t)\left(b_a - \widetilde{b}_a\right)\right) \tag{A.21}$$

The terms $\widetilde{\gamma}, \bar{R}, \gamma$ in equation (A.21) are constant; moreover $\widetilde{b}_a$ and b_a are constant too, by means of the model (2.17). Thus when the rotational motion is *rich enough*, the only feasible solution is $\widetilde{b}_a = b_a$ in order to keep coherence into the equality. Note that the gravity is observable up to the rotational ambiguity, which models the initial misalignment between the local and global vertical axis, which remains unobservable. □

A similar conclusion, for the case of a tightly-coupled vision-aided inertial navigation system, was drawn recently in the framework of tight coupling approach [16].

Remark 1

It is important to notice, however, that extending the state space with the gravity vector is unavoidable. The lack of such term would have destructive effects on the overall estimation. In this case no ambiguity would be associated to the gravity (we cannot write such term with the tilde hat), thus

$$\left(I - \bar{R}^T\right)\gamma = R\left(t\right)\left(b_a - \widetilde{b}_a\right) \tag{A.22}$$

A slight uncertainty in the initial attitude, would render

$$\left(\bar{R}^T - I\right)\gamma = constant \neq 0 \tag{A.23}$$

thus forcing

$$R\left(t\right)\left(b_a - \widetilde{b}_a\right) = constant \neq 0 \tag{A.24}$$

i.e. $\left(b_a - \widetilde{b}_a\right) \neq 0$*, in order to keep the equality to hold. In particular when the rotational motion is rich enough,* $\left(b_a - \widetilde{b}_a\right)$ *needs to vary.*

The previous proof and the remark lead to the following:

Corollary 1

The gravity γ *and the accelerometers biases* b_a *are observable in the combined vision-inertial configuration, provided that they are added to the state with trivial dynamics (null time-derivative) and the angular motion is rich enough.*

Definition of *Rich enough* rotational motion

The constraint in equation (A.21) provides the means to the formal characterization of the nature of the rotational motion in order to have the observability of the bias term b_a and of the gravity term up to the angular ambiguity $\bar{R} \in SO\left(3\right)$. The constant nature of $\widetilde{\gamma}, \bar{R}$ and γ forces $R\left(t\right)\left(b_a - \widetilde{b}_a\right) = R\left(t\right)\Delta b_a$ to be constant too, i.e.:

$$\dot{R}\left(t\right)\Delta b_a = R\left(t\right)\Omega\left(t\right)\Delta b_a = 0 \tag{A.25}$$

being $\Omega\left(t\right) = \omega\left(t\right)\wedge$ the skew-symmetric matrix of the body angular velocity. The derivative of the bias terms are obviously zero, for they are constant.

Equation (A.25) holds $\forall \Delta b_a$ if $\Omega\left(t\right) = 0$, that is $R\left(t\right)$ is constant and in this case the bias term is not observable, as expected.

For nonzero rotational velocities, (A.25) holds $\forall \Delta b_a \in Ker\left(\Omega\left(t\right)\right)$, i.e. $\Delta b_a = 0$ or every Δb_a which is aligned with the vector $\omega\left(t\right)$. It is easy to show that this is the case of constant angular velocities, $\omega\left(t\right) = \bar{\omega}$, non constant angular velocity along one axis only or every (constant/non constant) angular velocity such that the corresponding direction of rotation axis is fixed.

The goal is to find the family of all possible rotational motions such that $Ker\left(\Omega\right)$ reduces to $\Delta b_a = 0$, in the family of feasible (i.e. constant) b_a, $\widetilde{b}_a$. It is straightforward to prove that this space is composed by those rotations that happen along at least two axes and that keep the direction of the resulting angular velocity vector non constant. Formally, given two non-constant angular velocities $\omega_i\left(t\right), \omega_j\left(t\right)$ along two independent (orthogonal) directions $\vec{i}$ and $\vec{j}$ and such that $\omega_i\left(t\right) \neq \omega_j\left(t\right)$, observability is ensured by every angular velocity belonging to the following set:

$$\omega\left(t\right)=\omega_i\left(t\right)\vec{i}+\omega_j\left(t\right)\vec{j} \tag{A.26}$$

In this case the vector Δb_a lying along the vector $\omega\left(t\right)$ would vary, meaning that $\widetilde{b}_a$ would be non constant, missing the constraint $\dot{\widetilde{b}}_a = 0$. Thus the only possible choice is $\widetilde{b}_a - b_a = 0$ and the bias term is observable.

Personal Bibliography

[1] F. Cordella, F. Di Corato, G. Loianno, B. Siciliano and L. Zollo, "Robust Pose Estimation Algorithm for Wrist Motion Tracking" , *IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS13), Tokyo Big Sight, Japan, 2013 – *submitted.*

[2] F. Di Corato, M. Innocenti and L. Pollini, "Experimental Evaluation of a Visual-Inertial Navigation System with Guaranteed Convergence" , *2013 AIAA Guidance, Navigation, and Control Conference* (GNC13), Boston, MA, 2013 – *accepted for publication.*

[3] A. Caiti, V. Calabrò, F. Di Corato, D. Meucci and A. Munafò, "Minimum entropy distributed cooperative algorithms for Autonomous Underwater Vehicles in marine archeology search missions" , *OCEANS 2013 MTS/IEEE Bergen* (OCEANS13), 2013 – *accepted for publication.*

[4] V. Calabrò, F. Di Corato, A. Caiti, "Navigation and Control of an AUV affected by Asymmetric Thruster Response" , *OCEANS 2013 MTS/IEEE Bergen* (OCEANS13), 2013 – *accepted for publication.*

[5] A. Caiti, V. Calabrò, F. Di Corato, D. Meucci and A. Munafò, "Distributed Cooperative Algorithms for Autonomous Underwater Vehicles in Marine Search Missions" , *12th International Conference on Computer Applications and Information Technology in the Maritime Industries* (COMPIT13), 2013.

[6] F. Di Corato, M. Innocenti and L. Pollini, "Robust Vision-Aided Inertial Navigation Algorithm via Entropy-Like Relative Pose Estimation" , *Gyroscopy and Navigation Journal*, Volume 4, number 1, pp. 1-13, January 2013.

[7] F. Di Corato, M. Innocenti and L. Pollini, "Visual-Inertial Navigation with Guaranteed Convergence" , *Proceedings of the 2013 IEEE Workshop on Robot Vision* (WORV13), Clearwater Beach, FL, 2013.

[8] F. Cordella, F. Di Corato, L. Zollo, B. Siciliano and P. van der Smagt, "Patient performance evaluation using Kinect and Monte Carlo-Based fnger tracking" , *IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, pp. 1967-1972, 2012.

[9] F. Di Corato, M. Innocenti and L. Pollini, "Combined Vision-Inertial Navigation for Improved Robustness" , *Itzhack Y. Bar-Itzhack Memorial Symposium on Estimation, Navigation, and Spacecraft Control*, Haifa, Israel, 2012.

[10] F. Di Corato, M. Innocenti and L. Pollini. "Least entropy-like estimation for vision-based dead-reckoning and inertial navigation" , *Convegno SIDRA*, Pisa, 2011

PERSONAL BIBLIOGRAPHY

[11] F. Di Corato, M. Innocenti and L. Pollini, "An Entropy-Like Approach to Vision-Aided Inertial Navigation" , *18th IFAC World Congress, Proceedings of the*, pp. 13789-13794, volume 18, 2011.

[12] F. Di Corato, L. Pollini, M. Innocenti and G. Indiveri, "An Entropy-like approach to vision based autonomous navigation" , *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 1640-1645, 2011.

[13] L. Pollini, M. Innocenti, F. Di Corato, M. Cellini, M. Franchi, R. Mati and V. Niccolai, "The ICARO Autopilot: A Flexible Controller for small Unmanned Air Vehicles" , *4th US-European Workshop and Flight Competition for Micro Aerial Vehicles* (IMAV09), Pensacola, FL, June 2009.

[14] L. Pollini, M. Innocenti, F. Di Corato, M. Cellini, M. Franchi, R. Mati and V. Niccolai, "TheICARO autopilot: a project at the university of Pisa" , *International Symposium on Light Weight Unmanned Aerial Systems and Subsystems* (UAS-LW), Ostend, Belgium, March 2009.

Technical Reports and Documentation

[1] F. Di Corato and L. Pollini, "Bibliography Review for estimation algorithms in Dynamic Positioning operations" , *Deliverable D1; Project: SONSUB–CASTORONE*, December, 2009

[2] F. Di Corato and L. Pollini, "Results and implementations of the solutions for the state estimation in Dynamic Positioning operations with on-line wave frequency optimal estimation" , *Deliverable D2; Project: SONSUB–CASTORONE*, March, 2010

[3] F. Di Corato and L. Pollini, "A Unified Loosely-coupled Approach to Simultaneous Stereo Cameras-IMU Calibration and In-Motion Fine Alignment" , *Technical Report*, June, 2012

Bibliography

[1] G. Campa, M. Mammarella, M.R. Napolitano, M.L. Fravolini, L. Pollini, and B. Stolarik. A comparison of pose estimation algorithms for machine vision based aerial refueling for uavs. In *Control and Automation, 2006. MED '06. 14th Mediterranean Conference on*, pages 1–6, 2006.

[2] A. Chiuso, P. Favaro, H. Jin, and S. Soatto. Structure from motion causally integrated over time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):523–535, April 2002.

[3] J. De Geeter, H. Van Brussel, J. De Schutter, and M. Decreton. A smoothly constrained kalman filter. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(10):1171–1177, 1997.

[4] F. Di Corato, M. Innocenti, and L. Pollini. Robust vision-aided inertial navigation algorithm via entropy-like relative pose estimation. *Gyroscopy and Navigation Journal*, 4(1):1–13, January 2013.

[5] F. Di Corato, L. Pollini, M. Innocenti, and G. Indiveri. An entropy-like approach to vision based autonomous navigation.

In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1640–1645, 2011.

[6] David D. Diel, Paul DeBitetto, and Seth Teller. Epipolar constraints for vision-aided inertial navigation. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 221 –228, jan. 2005.

[7] F.Moreno-Noguer, V.Lepetit, and P.Fua. Accurate non-iterative o(n) solution to the pnp problem. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007.

[8] K.K. Gunnam, D.C. Hughes, J.L. Junkins, and N. Kehtarnavaz. A vision-based dsp embedded navigation sensor. *Sensors Journal, IEEE*, 2(5):428–442, 2002.

[9] R.M. Haralick, Hyonam Joo, D. Lee, S. Zhuang, V.G. Vaidya, and M.B. Kim. Pose estimation from corresponding point data. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(6):1426–1446, 1989.

[10] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[11] R. Hermann and Arthur J. Krener. Nonlinear controllability and observability. *Automatic Control, IEEE Transactions on*, 22(5):728–740, 1977.

[12] BertholdK.P. Horn. Relative orientation. *International Journal of Computer Vision*, 4(1):59–78, 1990.

[13] Ronchetti E.M. Huber P.J. *Robust statistics, 2nd edition*. Wiley, New York :, 2009.

[14] T. P. Hutchinson. *Essentials of Statistical Methods in 41 pages*. Rumbsby Scientific Publishing, Adelaide, S.A, 1993.

[15] A. H. Jazwinski. *Stochastic processes and filtering theory*. Dover Books on Electrical Engineering, 1970.

[16] Eagle S. Jones and Stefano Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *Int. J. Rob. Res.*, 30(4):407–430, April 2011.

[17] Rambabu Kandepu, Bjarne Foss, and Lars Imsland. Applying the unscented kalman filter for nonlinear state estimation. *Journal of Process Control*, 18(78):753 – 768, 2008.

[18] Jonathan Kelly and Gaurav S Sukhatme. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *Int. J. Rob. Res.*, 30(1):56–79, January 2011.

[19] Sangho Ko and Robert R. Bitmead. State estimation for linear systems with state equality constraints. *Automatica*, 43(8):1363 – 1368, 2007.

[20] Kurt Konolige, Motilal Agrawal, and Joan Solá. Large-scale visual odometry for rough terrain. In *Robotics Research*, volume 66 of *Springer Tracts in Advanced Robotics*, pages 201–212. Springer Berlin Heidelberg, 2010.

[21] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, January 1981.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[23] C.-P. Lu, G.D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(6):610–622, 2000.

[24] Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models.* SpringerVerlag, 2003.

[25] Marco Mammarella, Giampiero Campa, MarcelloR. Napolitano, and MarioL. Fravolini. Comparison of point matching algorithms for the uav aerial refueling problem. *Machine Vision and Applications*, 21(3):241–251, 2010.

[26] P.F. McLauchlan. Gauge invariance in projective 3d reconstruction. In *Multi-View Modeling and Analysis of Visual Scenes, 1999. (MVIEW '99) Proceedings. IEEE Workshop on*, pages 37 –44, 1999.

[27] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3565–3572, Rome, Italy, April 10-14 2007.

[28] A.I. Mourikis, S.I. Roumeliotis, and J.W. Burdick. Sc-kf mobile robot localization: A stochastic cloning kalman filter for processing relative-state measurements. *Robotics, IEEE Transactions on*, 23(4):717 –730, aug. 2007.

[29] Anastasios I. Mourikis, Nikolas Trawny, Stergios I. Roumeliotis, Andrew E. Johnson, Adnan Ansar, and Larry Matthies. Vision-aided inertial navigation for spacecraft entry, descent, and landing. *Trans. Rob.*, 25(2):264–280, April 2009.

[30] Richard M. Murray, S. Shankar Sastry, and Li Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1994.

[31] David Nistr, Oleg Naroditsky, and James Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23:2006, 2006.

[32] L. Pollini, F. Greco, R. Mati, and M. Innocenti. Stereo vision obstacle detection based on scale invariant feature transform algorithm. In *AIAA Guidance Navigation and Control Conference*, pages 1–13, Hilton Head, South Carolina, 2007.

[33] Lorenzo Pollini, Mario Innocenti, and Roberto Mati. Vision algorithms for formation flight and aerial refueling with optimal marker labeling. In *AIAA Modeling and Simulation Technologies Conference and Exhibit*, pages 15–18. San Francisco, California, 2005.

[34] Richard J. Prazenica, Adam Watkins, and Andrew J. Kurdila. Vision-Based Kalman Filtering for Aircraft State Estimation

and Structure from Motion. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*. San Francisco, California, 2005.

[35] R. M. Rogers. *Applied Mathematics in Integrated Navigation Systems*. American Institute of Aeronautics & Astronautics, Reston, Va, USA, 3rd edition, 2000.

[36] Simo Särkkä, Aki Vehtari, and Jouko Lampinen. Rao-blackwellized monte carlo data association for multiple target tracking. In *Proc. 7th International Conference on Information Fusion*, volume 1, pages 583–590, 2004.

[37] D. Simon. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *Control Theory Applications, IET*, 4(8):1303–1318, 2010.

[38] D. Simon and Tien Li Chia. Kalman filtering with state equality constraints. *Aerospace and Electronic Systems, IEEE Transactions on*, 38(1):128–136, 2002.

[39] S. Soatto, R. Frezza, and P. Perona. Motion estimation via dynamic vision. In *IEEE Transactions on Automatic Control*, volume 41, pages 393–414, 1996.

[40] Stefano Soatto. *A Geometric Framework for Dynamic Vision*. PhD thesis, California Institute of Technology, 1996.

[41] J.-P. Tardif, M. George, M. Laverne, A. Kelly, and A. Stentz. A new approach to vision-aided inertial navigation. In *Intel-*

ligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pages 4161–4168, 2010.

[42] Bruno O.S. Teixeira, Jaganath Chandrasekar, Leonardo A.B. Trres, Luis A. Aguirre, and Dennis S. Bernstein. State estimation for linear and non-linear equality-constrained systems. *International Journal of Control*, 82(5):918–936, 2009.

[43] P.H.S. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138 – 156, 2000.

[44] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, UK, 2000. Springer-Verlag.

[45] John Valasek, Kiran Gunnam, Jennifer Kimmett, Monish D. Tandale, John L. Junkins, and Declan Hughes. Vision-Based Sensor and Navigation System for Autonomous Air Refueling. *Journal of Guidance Control and Dynamics*, 28:979–989, 2005.

[46] E.A. Wan and R. Van der Merwe. The unscented kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 153–158, 2000.

[47] Thomas P Webb, Richard J Prazenica, Andrew J Kurdila, and Rick Lind. Vision-based state estimation for autonomous mi-

cro air vehicles. *Journal of guidance, control, and dynamics*, 30(3):816–826, 2007.