Inertial Structure From Motion with Autocalibration

Eagle Jones, Andrea Vedaldi, and Stefano Soatto

UCLA Vision Lab, Los Angeles, CA 90095, USA
[eagle|vedaldi|soatto]@cs.ucla.edu

Abstract. We present a technique to fuse inertial and visual information for real-time navigation applications. The combined model exhibits bounded bias, autocalibrates the camera-to-IMU transformation (eliminating the need for precise measurement or construction), and updates an estimate of the gravity vector. We analyze of the observability of the combined system, and discuss the implementation of a filter to estimate ego-motion from inertial and vision measurements as part of an integrated sensor system.

1 Introduction

Reliable estimation of the trajectory of a moving body ("ego-motion") is key to a number of applications, particularly to autonomous navigation of ground and air vehicles. Current methods typically employ global positioning measurements, sometimes integrated with inertial sensors. This depends on external signals, which are unavailable, or of low quality, in some of the most interesting scenarios (urban environments or indoors.) Integrating vision in the ego-motion estimation process holds great potential in this area. Vision and inertial sensors have naturally complementary characteristics, and it is no surprise that they are present in most animals with high mobility. Despite these promising characteristics, the potential of integrated vision and inertial navigation has not yet been realized. In this manuscript we address some of the key issues that have hindered progress, including managing the gravity vector and the calibration between the camera and inertial sensors.

While a multitude of filtering techniques may be employed to estimate ego-motion given all prior visual and inertial measurements, a *necessary condition* for any of them to operate correctly is that the underlying model be *observable.* We address this issue by showing that, in the presence of known gravity and known camera-inertial calibration, ego-motion is observable, under certain conditions.

A significant practical difficulty in integrating vision and inertial measurements is the need for accurate calibration of the mutual position and orientation between the camera and the inertial measurement unit (IMU). In this paper we show that such a calibration is actually unnecessary.

Any terrestrial system which incorporates accelerometer data is subject to the effects of gravity. The unavoidable biases in estimating such a large acceleration as $9.8m/s^2$, if not properly handled, compound under double integration to cause the rapid divergence of motion estimates. Many tricks of the trade have been employed to minimize the problems associated with gravity, but we show that when vision measurements are available the gravity vector becomes observable. Therefore, we simply add it to the state of our model and estimate it on-line in a straightforward and principled manner.

In addition to our analysis, we present a complete implementation of a nonlinear filter to estimate ego-motion from vision and inertial data. We have tested our algorithm, both in simulation with ground truth and on real data. Our experiments are performed on an embedded platform that includes range and positioning devices for validation.

This manuscript follows a number of attempts in the computer vision community to build a robust "visual odometry" module, including [15,3,20,4]. More specifically, some have proposed a variety of models incorporating inertial measurements, either as inputs to the model [17], or as states [16,5].

2 Formalization

Our exposition employs some notation that is standard in the robotics literature; readers interested in a thorough exposition should consult Chapter 2 of [14] or [12] for more details. We represent the motion of the (camera/IMU) body via $g = (R, T) \in SE(3)$, with $R \in SO(3)$ a rotation (orthogonal, positive-determinant) matrix, and $T \in \mathbb{R}^3$ a translation vector. $\widehat{V}^b = g^{-1}\dot{g} \in se(3)$ is the so-called "body velocity," i.e., the velocity of the moving body relative to the inertial frame, written in the coordinates of the moving body's reference frame. In homogeneous coordinates, we have

$$g = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}; \quad \widehat{V}^b = \begin{bmatrix} \widehat{\omega}^b & v^b \\ 0 & 0 \end{bmatrix}; \; V^b = \begin{bmatrix} \omega^b \\ v^b \end{bmatrix} \tag{1}$$

where $\widehat{\omega} \in so(3)$ is the skew-symmetric matrix constructed from the coordinates of $\omega \in \mathbb{R}^3$, and $v \in \mathbb{R}^3$ is the translational velocity. For a motion with constant rotational velocity ω, we have $R(t) = \exp(\widehat{\omega}t)$ if $R(0) = I$. The null rigid motion is $e = (I, 0)$. When writing the change of coordinates from a frame, say "b" for the moving body, to another frame, say "s" for the spatial (inertial) frame, we use a subscript g_{sb}, again following [14]. With this notation in place, we proceed to formalize the problem of estimating body pose and velocity.

We represent with $X_0^i \in \mathbb{R}^3$ the generic point in the inertial frame, and $y^i(t) \in \mathbb{R}^2$ its projection onto the (moving) image plane. Along with the pose g_{sb} of the body relative to the spatial frame and (generalized) body velocity V_{sb}^b, these quantities evolve via

$$\begin{cases} \dot{X}_0^i = 0, \quad i = 1, \dots, N \\ \dot{g}_{sb}(t) = g_{sb}(t)\widehat{V}_{sb}^b(t), \quad g_{sb}(0) = e \end{cases} \tag{2}$$

which can be broken down into the rotational and translational components $\dot{R}_{sb}(t) = R_{sb}(t)\widehat{\omega}_{sb}^b(t)$ and $\dot{T}_{sb}(t) = R_{sb}(t)v_{sb}^b(t)$. The translational component of body velocity, v^b_{sb}, can be obtained from the last column of the matrix $\frac{d}{dt}\widehat{V}^b_{sb}(t)$. That is, $\dot{v}^b_{sb} = \dot{R}^T_{sb}\dot{T}_{sb} + R^T_{sb}\ddot{T}_{sb} = -\widehat{\omega}^b_{sb}v^b_{sb} + R^T_{sb}\ddot{T}_{sb} \doteq -\widehat{\omega}^b_{sb}v^b_{sb} + \alpha^b_{sb}$, which serves to define $\alpha^b_{sb} \doteq R^T_{sb}\ddot{T}_{sb}$. An ideal inertial measurement unit would measure $\omega^b_{sb}(t)$ and $\alpha^b_{sb}(t) - R^T_{sb}(t)\gamma$ where γ denotes the gravity vector in the inertial frame. An ideal vision algorithm capable of maintaining correspondence and overcoming occlusions would measure $y^i(t) = \pi\left(R^T_{sb}(t)(X^i_0 - T_{sb}(t))\right)$. To summarize,

$$\begin{cases} \dot{X}^i_0 = 0, \quad i = 1, \dots, N \\ \dot{R}_{sb}(t) = R_{sb}(t)\widehat{\omega}^b_{sb}(t), \quad R_{sb}(0) = I \\ \dot{T}_{sb}(t) = R_{sb}(t)v^b_{sb}(t), \qquad T_{sb}(0) = 0 \\ \dot{v}^b_{sb}(t) = -\widehat{\omega}^b_{sb}(t)v^b_{sb}(t) + \alpha^b_{sb}(t) \\ y_{imu}(t) = \begin{bmatrix} \omega^b_{sb}(t) \\ \alpha^b_{sb}(t) - R^T_{sb}(t)\gamma \end{bmatrix} \\ y^i(t) = \pi\left(R^T_{sb}(t)(X^i_0 - T_{sb}(t))\right). \end{cases} \tag{3}$$

These equations can be simplified by defining a new linear velocity, v_{sb}, which is neither the body velocity v^b_{sb} nor the spatial velocity v^s_{sb}, but instead $v_{sb} \doteq R_{sb}v^b_{sb}$. Consequently, we have that $\dot{T}_{sb}(t) = v_{sb}(t)$ and $\dot{v}_{sb}(t) = \dot{R}_{sb}v^b_{sb} + R_{sb}\dot{v}^b_{sb} = \ddot{T}_{sb} \doteq \alpha_{sb}(t)$ where the last equation serves to define the new linear acceleration α_{sb}; as one can easily verify we have that $\alpha_{sb} = R_{sb}\alpha^b_{sb}$. The vision measurements remain unaltered, whereas the linear component of the inertial measurements become $R^T_{sb}(t)(\alpha_{sb}(t) - \gamma)$. If we model rotational acceleration $w(t) \doteq \dot{\omega}^b_{sb}$ and translational jerk $xi(t) \doteq \dot{\alpha}_{sb}(t)$ as Brownian motions, our random walk model, with biases and noises, and with all subscripts removed, is

$$\begin{cases} \dot{X}^i_0 = 0, \quad i = 1, \dots, N \\ \dot{R}(t) = R(t)\widehat{\omega}(t), \quad R(0) = I \\ \dot{T}(t) = v(t), \qquad T(0) = 0 \\ \dot{\omega}(t) = w(t) \\ \dot{v}(t) = \alpha(t) \\ \dot{\alpha}(t) = \xi(t) \\ y_{imu}(t) = \begin{bmatrix} \omega(t) \\ R^T(t)(\alpha(t) - \gamma) \end{bmatrix} + \begin{bmatrix} \omega_{bias} \\ \alpha_{bias} \end{bmatrix} + n_{imu}(t) \\ y^i(t) = \pi\left(R^T(t)(X^i_0 - T(t))\right) + n^i(t). \end{cases} \tag{4}$$

where $v \doteq v_{sb} \doteq R_{sb}v^b_{sb}$ and $\alpha \doteq \alpha_{sb} \doteq R_{sb}\alpha^b_{sb}$. In reality, the frames of the IMU and the camera do not coincide, and the IMU measurements are replaced with

$$y_{imu}(t) = R^T_{bi}\begin{bmatrix} \omega(t) \\ R^T(t)(\alpha(t) - \gamma + \ddot{R}(t)T_{bi}) \end{bmatrix} + \begin{bmatrix} \omega_{bias} \\ \alpha_{bias} \end{bmatrix} + n_{imu}(t) \tag{5}$$

where g_{bi} denotes the (constant) body-to-camera transformation, since we attach the body frame to the camera. Our choice of the camera frame as the body origin slightly complicates this model, but simplifies the following analysis. The results, of course, are identical whether derived in the camera or the IMU reference frame.

Both (3) and (4) are dynamical models with noise inputs which can be used to determine the likelihood of their outputs; estimating body pose $g_{sb}(t)$ and velocities v_{sb}^b, ω_{sb}^b is equivalent to inferring the state of such a model from measured outputs. This is a *filtering* problem [7] when we impose *causal processing*, that is, the state at time t is estimated using only those measurements up to t, as is necessary in closed-loop applications. These requirements admit a variety of estimation techniques, including sum-of-Gaussian filters [1], numerical integration, projection filters [2], unscented filters [9], and extended Kalman filters [7]. The condition which must be satisfied for any one of these approaches to work is that the model be *observable*. We address this issue in the next section.

3 Observability Analysis

The observability of a model refers to the possibility of uniquely determining the state trajectory (in our case the body pose) from output trajectories (in our case, point feature tracks and inertial measurements). Observability is independent of the amount of (input or output) noise, and it is a necessary condition for *any* filtering algorithm to converge [7]. When the model is not observable, the estimation error dynamics are necessarily unstable. It has been shown that pose is *not* observable from vision-only measurements [3], because of an arbitrary gauge transformation (a scaling and a choice of Euclidean reference frame [13]). The model can be made observable by fixing certain states, or adding pseudo-measurement equations [3]. It is immediate to show that pose is also not observable from inertial-only measurements, since the model consists essentially of a double integrator, and there has been extensive work to make the estimation error explode "slowly enough" that a platform can reach its target. For the purpose of analysis, we start with a simplified version of (4) with no camera-IMU calibration (see Sect. 3.2), no biases $\omega_{bias} = 0; \alpha_{bias} = 0$, no noises $\xi(t) = 0; w(t) = 0;\ n_{imu}(t) = 0; n^i(t) = 0$, since they have no effect on observability, and known gravity (see Sect. 3.3 otherwise).

The observability of a linear model can be determined easily with a rank test [10]. Analysis of non-linear models is considerably more complex [6], but essentially hinges on whether the initial conditions of (4) are *uniquely* determined by the output trajectories $\{y_{imu}(t), y^i(t)\}_{t=1,\dots,T;i=1,\dots,N}$. If it is possible to determine the initial conditions, the model (4) can be integrated forward to yield the state trajectories. On the other hand, if two different sets of initial conditions can generate the same output trajectories, then there is no way to distinguish their corresponding state trajectories based on measurements of the output.

3.1 Indistinguishable trajectories

As a gentle introduction to our analysis we first show that, when only inertial measurements are available, the model (4) is not observable. To this end, consider an output trajectory $y_{imu}(t)$ generated from a particular acceleration $\alpha(t)$. We integrate the model to obtain $v(t) = \int_0^t \alpha(\tau) d\tau + \bar{v}$, and we can immediately see that any initial velocity $\bar{v}$ will give rise to the same exact output trajectory. Hence, from the output, we will never be able to determine the translational velocity, and therefore the position of the body frame, uniquely.

Claim (Inertial only). Given inertial measurements $\{y_{imu}(t)\}_{t=1,\dots,T}$ only, the model (4) is not observable. If $\{R(t), T(t), \omega(t), v(t), \alpha(t) \neq 0\}$ is a state trajectory, then for any $\bar{v}, \bar{T}, \bar{R}$ identical measurements are produced by

$$\begin{cases} \tilde{R}(t) = \bar{R}R(t) \\ \tilde{T}(t) = \bar{R}T(t) + \bar{v}t + \bar{T} \\ \tilde{v}(t) = \bar{R}v(t) + \bar{v} \\ \tilde{\alpha}(t) = \bar{R}\alpha(t) \\ \tilde{\gamma} = \bar{R}\gamma. \end{cases} \tag{6}$$

If the gravity vector γ is known, then from $\tilde{\gamma} = \gamma$ we get that $\bar{R} = \exp(\widehat{\gamma})$, so the rotational ambiguity reduces to one degree of freedom. The claim can be easily verified by substitution to show that $\tilde{R}^T(t)(\tilde{\alpha}(t) - \tilde{\gamma}) = R^T(t)(\alpha(t) - \gamma)$, and assumes that $\|\tilde{\gamma}\| = \|\gamma\|$ is enforced. Note that if we impose $\tilde{R}(0) = R(0) = I$, then $\bar{R} = I$, and $\bar{T} = 0$, but we still have the ambiguity $\tilde{T}(t) = \exp(\widehat{\gamma})T(t) + \bar{v}t$, $\tilde{v}(t) = \exp(\widehat{\gamma})v(t) + \bar{v}$ and $\tilde{\alpha}(t) = \exp(\widehat{\gamma})\alpha(t)$. We will discuss the case $\alpha(t) = 0 \; \forall \, t$ shortly. The volume of the unobservable set grows with time even if we enforce $(R(0), T(0)) = (I, 0)$, as $\|T(t) - \tilde{T}(t)\| = \|(I - \bar{R})T(t) - \bar{v}t - \bar{T}\| = \|\bar{v}t\| \to \infty$. Vision measurements alone are likewise unable make the model observable.

Claim (Vision only). Given only vision measurements $\{y^i(t)\}_{i=1,\dots,N; t=1,\dots,T}$ of N points in general position [3], the model (4) is not observable. Given any state trajectory $\{X_0, R(t), T(t), \omega(t), v(t), \alpha(t)\}$, for any rigid motion $(\bar{R}, \bar{T})$ and positive scalar $\lambda > 0$, identical measurements are produced by

$$\begin{cases} \tilde{X}_0^i = \lambda(\bar{R}X_0^i + \bar{T}) \\ \tilde{R}(t) = \bar{R}R(t) \\ \tilde{T}(t) = \lambda(\bar{R}T(t) + \bar{T}) \\ \tilde{v}(t) = \lambda\bar{R}v(t) \\ \tilde{\alpha}(t) = \lambda\bar{R}\alpha(t) \end{cases} \tag{7}$$

This can be verified by substitution. Note that $\dot{\tilde{X}}_0^i = 0$, so $\lambda, \bar{R}, \bar{T}$ are arbitrary *constants*. Even if we enforce $(R(0), T(0)) = (I, 0)$, the unobservable set can grow unbounded, for instance $\|\tilde{T}(t) - T(t)\| = \|(I - \lambda\bar{R})T(t) - \lambda\bar{T}\| = |1 - \lambda|\|T(t)\|$.

We now fix the global reference frame, or equivalently the initial conditions $(R(0), T(0))$, by constraining three directions on the image plane, as described in [3]. In the combined vision-inertial system, this is sufficient to simultaneously restrain the motion of the IMU (given that the camera and IMU move together as a rigid body). This leaves us with an ambiguity in the scale factor only; that is, $\tilde{R} = R$ and $\tilde{T} = \lambda T$ (therefore $\tilde{\omega} = \omega$ and $\tilde{\alpha} = \lambda\alpha$). We do not yet have constraints on gravity, nor the transformation between camera and IMU. We seek to determine what, if any substitutions λ, $\tilde{g}_{bi}$, and $\tilde{\gamma}$ can be made for the true values 1, g_{bi}, and γ while leaving the measurements (5) unchanged.

Let us define the ambiguities $\bar{R} \doteq \tilde{R}_{bi} R_{bi}^T$ and $\bar{T} \doteq \lambda(\tilde{T}_{bi} - \bar{R}T_{bi})$. This allows us to write $\tilde{R}_{bi} = \bar{R}R_{bi}$ and $\tilde{T}_{bi} = \bar{R}T_{bi} + \lambda\bar{T}$ without loss of generality. The constraint $\tilde{\omega} = \omega$ and the IMU's measurement of angular velocity tell us that $R_{bi}^T\omega(t) = \tilde{R}_{bi}^T\tilde{\omega}(t) = R_{bi}^T\bar{R}^T\omega(t)$, so $\omega(t) = \bar{R}^T\omega(t)$. Hence $\bar{R}$ is forced to be a rotation around the ω axis; it is easy to verify that this implies

$$\bar{R}\hat{\omega} = \hat{\omega}\bar{R}. \tag{8}$$

The accelerometer measurements require that

$$R_{bi}^T R^T(t)\left(\alpha(t) - \gamma + \ddot{R}(t)T_{cb}\right) = \tilde{R}_{bi}^T R^T(t)\left(\tilde{\alpha}(t) - \tilde{\gamma} + \ddot{R}(t)\tilde{T}_{bi}\right). \tag{9}$$

This is satisfied only by assigning

$$\tilde{\gamma} = R\bar{R}R^T(t)\gamma + \left(\lambda I - R(t)\bar{R}R^T(t)\right)\alpha(t) + \ddot{R}(t)\lambda\bar{T}. \tag{10}$$

Note that $R^T(t)\ddot{R}(t) = \dot{\hat{\omega}}(t) + \hat{\omega}^2(t)$, so (8) allows us to write $\bar{R}R^T(t)\ddot{R}(t) = R^T(t)\ddot{R}(t)\bar{R}$. This identity may be used to verify (10) by substitution into (9). We can now fully describe the ambiguities of the system.

Claim (Observability of Combined Inertial-Vision System). Provided the global reference frame is fixed as in [3], two state trajectories for the system (4-5) are indistinguishable if and only if, for constants $\lambda \in \mathbb{R}$ and $(\bar{R}, \bar{T}) \in SE(3)$,

$$\begin{cases} \tilde{X}_0^i = \lambda X_0^i \\ \tilde{R}(t) = R(t) \\ \tilde{T}(t) = \lambda T(t) \\ \tilde{R}_{bi} = \bar{R}R_{bi} \\ \tilde{T}_{bi} = \bar{R}T_{bi} + \lambda\bar{T} \\ \tilde{\omega}(t) = \omega(t) = \bar{R}\omega(t) \\ \tilde{\gamma} = R\bar{R}R^T(t)\gamma + (\lambda I - R(t)\bar{R}R^T(t))\alpha(t) + \ddot{R}(t)\lambda\bar{T}, \end{cases} \tag{11}$$

We now examine a few scenarios of interest. First, in a simple case when gravity and calibration are known, the ambiguity reduces to $0 = (\lambda - 1)\alpha(t)$, which tells us that scale is determined so long as acceleration is non-zero.

Claim (Inertial & Vision). The model (4) is locally observable provided that $\alpha(t) \neq 0$ and that the initial conditions $(R(0), T(0)) = (I, 0)$ are enforced.

We emphasize that unless the global reference is fixed by saturating the filter along three visible directions, following the analysis in [3], the choice of initial pose is not sufficient to make the model observable since it is not actively enforced by the filter.

The term "locally observable" refers to the fact that infinitesimal measurements are sufficient to disambiguate initial conditions; local observability is a stronger condition than global observability.

3.2 Observability with unknown calibration

Measuring the transformation between the IMU and camera precisely requires elaborate calibration procedures, and maintaining it during motion requires tight tolerances in mounting. To the best of our knowledge there is no study that characterizes the effects of camera-IMU calibration errors on motion estimates. Consider the simplified case of known gravity, and correct rotational calibration, but a small translational miscalibration (for example, due to expansion or contraction of metals with temperature). Our constraint becomes $(1-\lambda)\alpha(t) = \ddot{R}(t)\lambda\bar{T}$, where $\bar{T}$ is the miscalibration. For general motion, this is clearly not satisfiable, and can cause divergence of the filter. In this section we show that such errors can be made to have a negligible effect; indeed, we advocate forgoing such a calibration procedure altogether. Instead, a filter should be designed to automatically calibrate the camera and IMU.

First consider the ambiguity in rotational calibration, $\bar{R}$. Since $\bar{R}\omega(t) = \omega(t)$, $\bar{R}$ must be the identity when $\omega(t)$ is fully general.[1] This reduces the second constraint to $(1-\lambda)\alpha(t) = \ddot{R}\lambda T$. If $\alpha(t)$ is non-zero and not a function of $\ddot{R}$, then $\lambda = 1$ and the model is observable.

Claim (Observability of calibration). The model (4), augmented with (5) and T_{bi}, R_{bi} added to the state with constant dynamics, is locally observable, so long as motion is sufficiently exciting and the global reference frame is fixed.

3.3 Dealing with gravity

We now turn our attention to handling the unknown gravity vector. Because γ has a rather large magnitude, even small estimation errors in R_{sb} will cause a large innovation residual $n_{imu}(t)$. Dealing with gravity is an art of the inertial navigation community, with many tricks of the trade developed over the course of decades of large scale applications. We will not review them here; the interested reader can consult [11]. Rather, we focus on the features of vision-inertial integration. Most techniques already in use in inertial navigation, from error statistics to integration with positioning systems, can be easily incorporated.

Our approach is to simply add the gravity vector to the state of the model (4) with trivial dynamics $\dot{\gamma} = 0$ and small model error covariance. Note that this is *not* equivalent to the slow-averaging customarily performed in navigation filters – the disambiguation of the gravity vector comes from the coupling with vision measurements. Assuming known calibration, we have that $\tilde{\gamma} = \gamma + (\lambda - 1)\alpha(t)$. Since γ and $\tilde{\gamma}$ are constants, λ must be unity as long as $\alpha(t)$ is non-constant.

Claim (Observability of gravity). The gravity vector, if added to the state of (4) with trivial dynamics $\dot{\gamma} = 0$, is locally observable provided that $\alpha(t)$ is not constant and the global reference frame is fixed.

[1] Special cases include not only $\omega(t) = 0$, but also $\omega(t)$ spanning less than two independent directions.

3.4 Summary and notes

The claims just made may be combined if gravity and calibration are unknown.

Claim (Observability of calibration and gravity). The model (4-5) and T_{bi}, R_{bi}, γ added to the state with constant dynamics, is locally observable, so long as motion is sufficiently exciting and the global reference frame is fixed.

When we describe a motion as "sufficiently exciting", this means that the motion is varied enough to span the entirety of the state space. This condition is typically assumed to hold asymptotically in off-line identification problems.

This condition may not be satisfied in common cases of practical import; for example, most ground navigation occurs approximately on a plane with rotation primarily about a single axis. "Cruising", with zero angular velocity and zero linear acceleration is also common. A number of other special cases will allow the constraints (11) to be satisfied. The following experiments demonstrate "calibration sequences" complex enough that we can estimate all parameters. A full derivation of the constraints and a more complete analysis of degenerate cases are available in [8].

4 Experiments

The model we use to design a filter is a modified discrete-time version of (4):

$$\begin{cases} y_0^i(t+1) = y_0^i(t) + n_0^i(t) \quad i = 4, \dots, N(t) \\ \rho^i(t) = \rho^i(t) + n_\rho^i(t) \quad i = 1, \dots, N(t) \\ T(t+1) = T(t) + v(t), \quad T(0) = 0 \\ \Omega(t+1) = Log_{SO(3)}(\exp(\widehat{\Omega}(t)) \exp(\widehat{\omega}(t)), \quad R(0) = I \\ v(t+1) = v(t) + \alpha(t) \\ \omega(t+1) = \omega(t) + w(t) \\ \alpha(t+1) = \alpha(t) + \xi(t) \\ \xi(t+1) = \xi(t) + n_\xi(t) \\ w(t+1) = w(t) + n_w(t) \\ \gamma(t+1) = \gamma(t) + n_\gamma(t) \\ T_{cb}(t+1) = T_{cb}(t) + n_{T_{cb}}(t) \\ \Omega_{cb}(t+1) = \Omega_{cb}(t) + n_{\Omega_{cb}}(t) \\ y^i(t) = \pi\left(e^{\widehat{\Omega}_{cb}(t)}(e^{-\widehat{\Omega}(t)}(e^{-\widehat{\Omega}_{cb}(t)}(y_0^i(t)e^{\rho^i(t)} - T_{cb}(t)) - T(t))) + T_{cb}(t)\right) + n^i(t) \\ y_{imu}(t) = \begin{bmatrix} \omega(t) + \omega_{bias} \\ e^{-\widehat{\Omega}(t)}(\alpha(t) - \gamma(t)) + \alpha_{bias} \end{bmatrix} + n_{imu}(t) \\ norm_\gamma = \|\gamma(t)\| \end{cases} \tag{12}$$

where all noises are assumed to be white, zero-mean Gaussian processes, with covariances set in a tuning procedure. Where the analysis was simplified by attaching the body frame to the camera, our implementation is simplified by attaching the body frame to the IMU. Thus g_{bi} has been supplanted by g_{cb} and the vision measurements are transformed rather than the IMU measurements. The last (pseudo-)measurement sets the norm of gravity to a constant, with degenerate (zero) covariance. Notice that the number of visible features $N(t)$ can change with time, and the index i in the first equation (describing point feature positions in the camera at time t) starts from 4. Fixing the first three points[2] is equivalent to fixing three directions in space, which fixes the global reference as described in [3]. Depths are represented using the exponential map to ensure that their estimates are positive. We remind the reader that $\omega = \omega_{sb}^{b}$, whereas $v = v_{sb}$ and $\alpha = \alpha_{sb}$ are defined as in (4).

To overcome failures of the low-level feature tracking module, we employ a robust version of an extended Kalman filter (EKF), similar to [19], which allows a variable number of points. We have implemented our filter in simulation as well as on an embedded platform which we developed.

4.1 Simulation platform

Our simulation platform follows the parameters of [3], allowing us to generate point tracks and inertial measurements with ground truth reference. We generated sets of 10 repeated trials for each of 100,000 combinations of parameters, generating over $1M$ test results. Obviously we cannot summarize all the results in the space available, so we limit ourselves to reporting representative trials to validate the analysis we have performed.

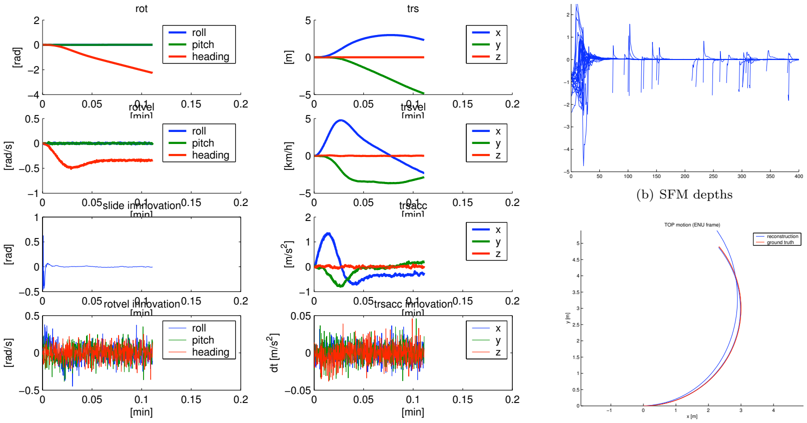
Fig. 2 shows a typical outcome when the covariance of the vision measurements is inflated, so that only the inertial measurements are used. Gravity and calibration are assumed known, but a small error in the gravity direction is reflected in a non-zero mean component of the innovation. This is intentionally left uncorrected in order to emphasize the resulting drift in the estimated trajectory, which is significant even though the innovation remains substantially white. In Fig. 3 we show the results of a similar experiment where vision measurements are used along with inertial. The non-zero component of the innovation is still visible, but now the bias is substantially reduced. The bias affects both the estimated trajectory and the estimated positions of the points in space, seen as red crosses. Scale is estimated correctly.

4.2 Autocalibration

In Fig. 4 we report a representative example that illustrates the proposed approach to deal with unknown camera-imu calibration and estimate gravity.[3] All

[2] We choose the first three points for simplicity; in practice one may want to choose three points that are sufficiently far apart to ensure stability of the fixed frame.

[3] Note that throughout this paper, autocalibration refers to the camera-to-IMU calibration, not to the intrinsic parameters of the camera – those can be inferred through standard calibration procedures as customary in this application domain [18].



(a) IMU innovation

(b) SFM depths

(c) Reconstruction

Fig. 1. *We simulate a vehicle moving around a circle. The vehicle starts from still and then accelerates up to about 10km/h. The camera is mounted 2m above the IMU and points slightly downward and toward the center of motion (as opposed to the heading direction). Both the IMU and vision measurements are affected by noise. (a) Motion estimate plus IMU innovation. (b) Vision estimate of the point feature depths (note the features added at later times). (c) Estimate of the vehicle trajectory (blue) using inertial and vision measurements, compared to ground truth (red) and another estimate obtained by the IMU alone (other blue curve).*

the model parameters $(\Omega_{cb}, T_{cb}, \gamma)$ are observable only under the assumption of sufficiently exciting input sequences as discussed previously. Once the parameters have converged during appropriate types of motion, standard covariance-scheduling procedures can be adopted to "fix" the parameter values (calibration and gravity) during degenerate motion. The interplay between acceleration and gravity results in slower convergence. However, the innovation eventually settles to a moderately colored process with small mean and covariance on the same order of magnitude as the measurement noise. More extensive experiments are reported in [8].

4.3 An embedded vision-inertial module

We implemented an optimized version of our full filter which recovers ego-motion from inertial measurements and video on a modern CPU faster than the sensor data arrives (30 Hz for images and 100 Hz for inertial). The overall system is shown in Fig. 5 and includes various cameras (omni-directional, binocular and trinocular), as well as a BEI-Systems IMU, with the filter implemented on a custom computer running on batteries. A laser range finder and a GPS are used for validation. The experiments we have run on indoor and outdoor scenes do not

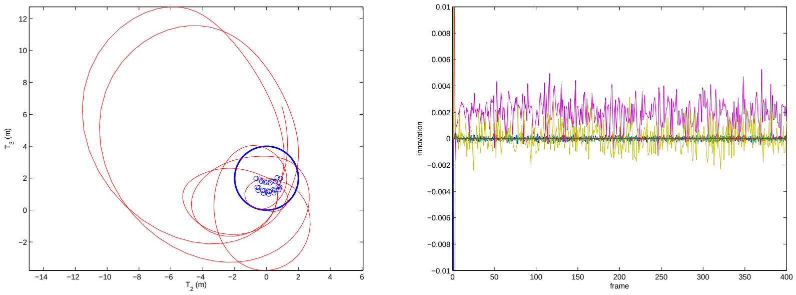


Fig. 2. Inertial-only. *Left: Estimated trajectory of the camera (red solid) compared to ground truth (blue dotted) seen from the bird's eye view; significant drift occurs as a result of double integration and bias; the positions of the points in space (blue circles) are not used in this scenario and are displayed only as a reference. Right: Innovation; it can be seen that there is a small DC component due to simulated bias and alignment error with respect to gravity.*

have accurate enough ground truth, and show the same qualitative behaviour (innovation statistics and state error covariances) as the simulation, except for significantly smaller inertial biases since these are explicitly modeled and handled by the inertial unit. We are in the process of running ground-truth experiments on known 3-D structures, that will be reported in a follow-up technical report.

5 Discussion

We have shown that vision can play an important role in inertial navigation, since it can make pose observable, relative to the initial frame. It can also make the gravity vector observable, and therefore render delicate registration and calibration procedures unnecessary.

We have developed a complete filter based on these conclusions, which we tested extensively in simulation to validate our analysis. Last, but not least, we have implemented an embedded system that runs the proposed vision-inertial filters in real-time. The platform can be mounted on wheels, on a vehicle, or carried by a human, and includes additional instruments (GPS and lidar) for validation.

All the analysis is valid "locally" to the extent in which visual features remain visible. In an experiment where at least 5 features are visible throughout the sequence, positioning relative to the initial frame is possible to the extent described by the analysis. If the initial frame is geo-referenced, so is the entire trajectory that follows. Where visual features disappear, for instance during long forward motion without recognizable landmarks, a bias is accumulated in both visual and inertial measurements, and therefore the benefit of using the two modalities is incremental (i.e. a reduced drift) as opposed to absolute (i.e.

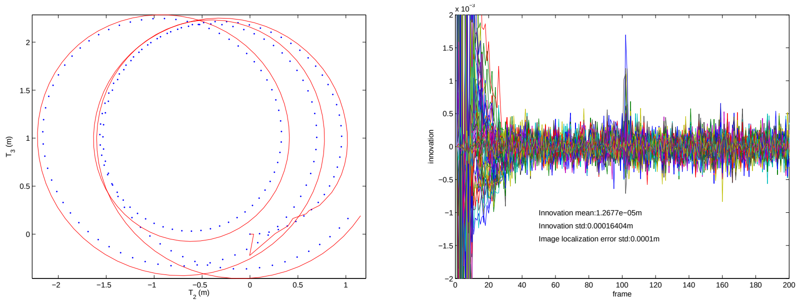


Fig. 3. Vision-Inertial. *Left: The use of vision measurements helps reduce the bias, and enables the estimation of the positions of feature points in space (red crosses; compare with ground truth in blue circles). Notice that the small bias, accumulated during transient, is reflected both in the trajectory and in the reconstruction of the points in space. If known landmarks are visible, this can be used to correct the reconstruction and thus eliminate the bias. Right: The innovation still shows the small bias in acceleration residuals due to the misalignment of gravity.*

the total elimination of drift). Even in this case, however, the important difference between using inertial-only measurements is that the ensuing filter remains observable, rather than diverging – however slowly – as in inertial navigation, and the volume of the unobservable subset remains bounded. We have verified empirically that the estimate of the error covariance remains bounded so long as enough features (typically 30 or more) survive with enough overlap (typically 30 frames or more) throughout the sequence. In sharp turns, where all visual features are lost, inertial measurements provide a valuable transition modality to initialize new features.

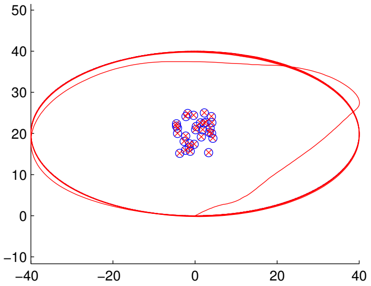


Fig. 4. Autocalibration and gravity. *Calibration parameters and gravity are observable provided that the motion sequence is "generic," which entails non-constant acceleration and rotation about an axis spanning at least two independent directions. All parameters converge to their nominal values. More experiments are discussed in [8], including experiments for pathological but common motion sequences.*

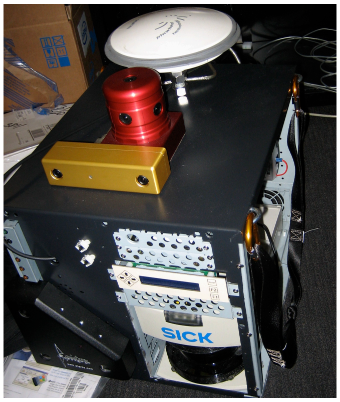


Fig. 5. Embedded platform. *A view of our embedded platform, including monocular-omnidirectional (red), binocular (gold) and trinocular (black) cameras, inertial unit, a custom computer, as well as a lidar and GPS for ground truth generation. Processing is done on a custom computer, and power is drawn from battery packs. The platform can be mounted on one's shoulders with straps, or on top of a vehicle with suction cups, or on a mobile wheeled base with Velcro.*

Acknowledgments

This research was supported by AFOSR FA9550-06-1-0138. We are grateful to Dr. Stergios Roumeliotis for pointing out an error in our analysis in a previous version of this manuscript.

References

1. D. L. Alspach and H. W. Sorenson. Nonlinear bayesian estimation using gaussian sum approximation. *IEEE Trans. Aut. Contr.*, 17(4):439–448, 1972. 4
2. D. Brigo, B. Hanzon, and F. LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *IEEE Trans. on Automatic Control*, 68:181–188, 1998. 4
3. A. Chiuso, P. Favaro, H. Jin, and S. Soatto. Motion and structure causally integrated over time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24 (4):523–535, 2002. 2, 4, 5, 6, 9
4. A. Davison. Real-time simulataneous localisation and mapping with single camera. In *Proc. 9^{th} Int. Conf. on Computer Vision*, 2003. 2
5. E. D. Dickmanns and B. D. Mysliwetz. Recursive 3-D road and relative egostate estimation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(2):199–213, February 1992. 2
6. A. Isidori. *Nonlinear Control Systems*. Springer Verlag, 1989. 4
7. A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970. 4
8. E. Jones, A. Vedaldi, and S. Soatto. Integrated visual and inertial navigation and calibration. Technical report, UCLA CSD Technical Report, April 2008. 8, 10, 12
9. S. J. Julier and J. K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Int. Symp. o Aerospace/Defense Sensing, Simulation and Control*, 1997. 4
10. T. Kailath. *Linear Systems*. Prentice Hall, 1980. 4
11. M. Kayton and W.R. Fried. *Avionics Navigation Systems*. Wiley and Sons, 1996. 7

12. Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to models.* Springer Verlag, 2003. 2
13. P. F. McLauchlan. Gauge invariance in projective 3d reconstruction. In *IEEE Workshop on Multi-View Modeling and Analysis of Visual Scenes, Fort Collins, CO, June 1999*, 1999. 4
14. R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation.* CRC Press, 1994. 2
15. D. Nister. Preemptive ransac for live structure and motion estimation. In *Proc. 9^{th} Int. Conf. on Computer Vision*, 2003. 2
16. G. Qian, R. Chellappa, and Q. Zheng. Robust structure from motion estimation using inertial data. In *Journal of the Optical Society of America A*, 2001. 2
17. S. I. Roumeliotis, A. E. Johnson, and J. F. Montgomery. Augmenting inertial navigation with image-based motion estimation. In *IEEE Intl. Conf. on Robotics and Automation*, 2002. 2
18. R. Tsai. A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE J. Robotics Automat.*, RA-3(4):323–344, 1987. 9
19. A. Vedaldi, H. Jin, P. Favaro, and S. Soatto. KALMANSAC: Robust filtering by consensus. In *Proceedings of the International Conference on Computer Vision (ICCV)*, volume 1, pages 633–640, 2005. 9
20. R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *Proc. CVPR'03 (IEEE Conf. on Computer Vision and Pattern Recognition)*, page 211. 2