Observability Properties and Deterministic Algorithms in Visual-Inertial Structure from Motion

Agostino Martinelli
INRIA
agostino.martinelli@ieee.org

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Robotics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. Martinelli. *Observability Properties and Deterministic Algorithms in Visual-Inertial Structure from Motion*. Foundations and Trends® in Robotics, vol. 3, no. 3, pp. 139–209, 2012.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-739-6

Foundations and Trends® in Robotics
Volume 3, Issue 3, 2012

Editorial Scope

Topics

Foundations and Trends® in Robotics publishes survey and tutorial articles in the following topics:

- Mathematical modelling
- Kinematics
- Dynamics
- Estimation methods
- Artificial intelligence in robotics
- Software systems and architectures
- Sensors and estimation
- Planning and control
- Human-robot interaction
- Industrial robotics
- Service robotics

Information for Librarians

Foundations and Trends® in Robotics, 2012, Volume 3, 4 issues. ISSN paper version 1935-8253. ISSN online version 1935-8261. Also available as a combined paper and online subscription.

Foundations and Trends® in Robotics
Vol. 3, No. 3 (2012) 139–209

DOI: 10.1561/2300000030

Observability Properties and Deterministic Algorithms in Visual-Inertial Structure from Motion

Agostino Martinelli
INRIA
agostino.martinelli@ieee.org

Contents

Abstract

This paper discusses the visual inertial structure from motion problem (VI-SfM problem) with special focus on the following three fundamental issues: observability properties, resolvability in closed form and data association. Regarding the first issue, after a discussion about the current state of the art, the paper investigates more complex scenarios. Specifically, with respect to the common formulation, which assumes three orthogonal accelerometers and three orthogonal gyroscopes, the analysis is extended to cope with the cases of a reduced number of inertial sensors and any number of point features observed by monocular vision. In particular, the minimal case of a single accelerometer, no gyroscope and a single point feature is addressed. Additionally, the analysis accounts for biased measurements and unknown extrinsic camera calibration. The results derived for these new and very challenging scenarios have interesting consequences both from a technological and neuroscientific perspective. Regarding the second issue, a simple closed form solution to the VI-SfM is presented. This solution expresses the structure of the scene and the motion only in terms of the visual and inertial measurements collected during a short time interval. This allows introducing deterministic algorithms able to simultaneously determine the structure of the scene together with the motion without the need for any initialization or prior knowledge. Additionally, the closed-form solution allows us to identify the conditions under which the VI-SfM has a finite number of solutions. Specifically, it is shown that the problem can have a unique solution, two distinct solutions or infinite solutions depending on the trajectory, on the number of point-features and on their arrangement in the 3D space and on the number of camera images. Finally, the paper discusses the third issue, i.e., the data association problem. Starting from basic results in computer vision, it is shown that, by exploiting the information provided by the inertial measurements, a single point correspondence (in the case of a planar motion) and two point correspondences (for a general 3D motion) are sufficient to characterize the motion between two camera poses. This allows us to use an 1-point RANSAC algorithm (in the planar case) or a 2-point RANSAC algorithm (in the general 3D case) to detect outliers. The paper concludes with some discussion about connections to related research fields both in the framework of computer science and neuroscience.

A. Martinelli. *Observability Properties and Deterministic Algorithms in Visual-Inertial Structure from Motion*. Foundations and Trends® in Robotics, vol. 3, no. 3, pp. 139–209, 2012.

DOI: 10.1561/2300000030.

1

Introduction

The term *Structure from Motion* (SfM) was coined by the computer vision community to define the problem of estimating the three-dimensional structure of the scene and the motion from two-dimensional image sequences. In this paper we consider the same estimation problem but the sensor suit is also composed by inertial sensors (accelerometers and gyroscopes). We will refer to this problem as to the *Visual-Inertial Structure from Motion* problem (from now on the VI-SfM problem). The VI-SfM problem has particular interest and has been investigated in many disciplines, both in the framework of computer science (e.g., [5, 23, 24, 37, 49]) and in the framework of neuroscience (e.g., [4, 11, 14]). These sensors require no external infrastructure and this is a key advantage for robots operating in unknown environments where GPS signals are shadowed. For this reason, vision and inertial sensing have received great attention by the mobile robotics community in the last years and many approaches have been introduced.

According to Corke et al. [7], we distinguish between loosely coupled and tightly coupled approaches. In the former, the sensor data processing takes place in separate modules, which exchange information each other. The information delivered by the inertial can be used to speed up the tracking task of the features by predicting their locations within the next frame; in turn, data from the visual sensor allows updating the calibration parameters of inertial sensors. In the latter (tightly coupled approaches), all measurements, both visual and inertial, are combined and processed using a common filter-based approach.

A special issue of the *International Journal of Robotics Research* has recently been devoted to the visual and inertial sensor fusion [10]. In [7], a tutorial introduction to the vision and inertial sensing was presented. This work provides a biological point of view and it illustrates how vision and inertial sensors have useful complementarities allowing them to cover the respective limitations and deficiencies. In [47] the inertial measurements are used to reduce the ambiguities in the structure from motion problem.

The majority of the approaches so far introduced, perform the fusion of vision and inertial sensors by filter-based algorithms. In [3], these sensors are used to perform egomotion estimation. The sensor fusion is obtained by an Extended Kalman Filter (*EKF*) and by an Unscented Kalman Filter (*UKF*). The approach proposed by Gemeiner et al. [16] extends the previous one by also estimating the structure of the environment where the motion occurs. In particular, new landmarks are inserted on line into the estimated map. This approach has been validated by conducting experiments in a known environment where a ground truth was available. Also, in [52], an *EKF* has been adopted. In this case, the proposed algorithm estimates a state containing the robot speed, position and attitude, together with the inertial sensor biases and the location of the features of interest. In the framework of aerial navigation, an *EKF* has been adopted by Kim and Sukkarieh [25] to perform VI-SfM. It was observed that any inconsistent attitude update severely affects any solution. The authors proposed to separate attitude update from position and velocity update. Alternatively, they proposed to use additional velocity observations, such as air velocity observation. Very recently, in the frame work of micro aerial robotics, flight stabilization and fully autonomous navigation have been achieved by using monocular vision and inertial sensors as the only exteroceptive sensors.

Also in this case, the sensor fusion was carried out by a filter based algorithm [53, 54].

Since most of the previous algorithms to fuse visual and inertial measurements are based on linear estimators, in the last years, the effect that the observability properties can have on the consistency of a linearized estimator, has been investigated by Hesch et al. [22], Kottas et al. [26], Li and Mourikis [28]. More in general, recent works have investigated the observability properties of the VI-SfM problem for various scenarios (see Section 2.1 for a detailed state of the art).

The first goal of this paper (chapter 2) is to provide the main results achieved in the last years about the observability properties of the VI-SfM problem and to provide an important step forward by analyzing new and very challenging scenarios. Specifically, in the second part of chapter 2, the observability analysis is extended to cope with the cases of a reduced number of inertial sensors and any number of point features observed by the monocular vision. In particular, the minimal case of a single accelerometer, no gyroscope and a single point feature is addressed. Additionally, the analysis accounts for biased measurements and unknown extrinsic camera calibration. The results derived for these new and very challenging scenarios have interesting consequences both from a technological and neuroscientific perspective.

There are very few methods able to perform the fusion of image and inertial measurements without a filter-based approach. One algorithm of this type has been suggested by Strelow and Singh [49]. This algorithm is a batch method that performs SfM from image and inertial measurements. Specifically, it minimizes a cost function by using the Levenberg Marquardt algorithm [46]. This minimization process starts by initializing the velocities, the gravity and the biases to zero.

When using a recursive estimator (e.g. an *EKF*), or an off-line optimization method to minimize a suitable cost function, an important issue that arises is the initialization problem. Indeed, because of the system non-linearities, an erroneous initialization can irreparably compromise the entire estimation process. This problem has firstly been considered by Lupton and Sukkarieh [31]. In particular, they proposed a method able to estimate the scale factor by using a square root information filter. Additionally, the same authors proposed an *EKF* that does not suffer from the initialization of the speed and of the orientation [32, 33]. An efficient solution to the initialization problem is obtained by introducing a deterministic algorithm able to determine the initial state by using the visual and inertial measurements acquired during a short time interval. This issue has been addressed only very recently by Dong-Si and Mourikis [12] and Martinelli [35, 37, 38]. Specifically, in [35] the first closed-form solution to VI-SfM has been obtained. Then, new extensions of this solution have been derived to cope with the cases of biased accelerometer's measurements [37, 38] and an unknown camera-IMU extrinsic calibration [12]. In chapter 3, we provide a simple closed-form solution to the VI-SfM and, starting from this, we identify the conditions under which the VI-SfM has a finite number of solutions. Specifically, it is shown that the problem can have a unique solution, two distinct solutions or infinite solutions depending on the trajectory, on the number of point-features and on their arrangement in the 3D space and on the number of camera images.

Finally, a fundamental issue that arises in any visual motion estimation problem is data association. Any matching algorithm suffers from outliers, which must be detected and removed. To achieve this objective, the *RANdom SAmple Consensus* (RANSAC) introduced by Fischler and Bolles [15] has been extensively used in visual motion estimation. In the past, concerning the case of only visual measurements (i.e., in SfM), the 5-point RANSAC [45] has been adopted [27, 50]. Indeed, five correspondences are in general necessary to identify the five parameters that characterize both the rotation (3 parameters) and the translation up to a scale (2 parameters) between two camera frames. In the special case of a planar motion, the rotation can be characterized by a single parameter and the translation up to a scale by a further parameter. Hence, a 2-point RANSAC can be adopted to detect outliers for any planar motion. In the context of wheeled and indoor robotics, the motion not only is planar but also satisfies the non-holonomic constraint. This further information has been exploited by Scaramuzza [48] to use a 1-point RANSAC for outliers detection.

The information provided by the inertial measurements, dramatically simplifies the data association task. Indeed, the rotation between two camera frames can be efficiently obtained by integrating the inertial measurements. The translation up to a scale only depends on two parameters (for a general $3D$ motion) and one parameter (in the planar case). This makes possible the use of a 2-point RANSAC in the general $3D$ case and a 1-point RANSAC in the planar case [51]. We discuss this issue very briefly in Chapter 4.

References

[1] D. E. Angelaki, A. G. Shaikh, A. M. Green, and J. D. Dickman. Neurons compute internal models of the physical laws of motion. *Nature*, pages 560–564, 2004.

[2] M. Anguelova. *Non linear Observability and Identifiability: General Theory and a Case Study of a Kinetic Model*. PhD thesis, Goteborg, 2004.

[3] L. Armesto, J. Tornero, and M. Vincze. Fast ego-motion estimation with multi-rate fusion of inertial and vision. *The International Journal of Robotics Research*, pages 577–589, 2007.

[4] A. Berthoz, B. Pavard, and L. R. Young. Perception of linear horizontal self-motion induced by peripheral vision (linearvection) basic characteristics and visual–vestibular interactions. *Exp. Brain Res.*, pages 471–489, 1975.

[5] M. Bryson and S. Sukkarieh. Observability analysis and active control for airbone slam. *IEEE Transaction on Aerospace and Electronic Systems*, pages 261–280, 2008.

[6] A. Chiuso, P. Favaro, H. Jin, and S. Soatto. Structure from motion causally integrated over time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 523–535, 2002.

[7] P. Corke, J. Lobo, and J. Dias. An introduction to inertial and visual sensing. *Journal of Robotics Research*, pages 519–535, 2007.

[8] V. Cornilleau-Peres and J. Droulez. The visual perception of three-dimensional shape from self-motion and object-motion. *Vision Research*, pages 2331–2336, 1994.

[9] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1052–1067, 2007.

[10] J. Dias, M. Vinzce, P. Corke, and J. Lobo. Editorial: Special issue: 2nd workshop on integration of vision and inertial sensors. *The International Journal of Robotics Research*, pages 515–517, 2007.

[11] K. Dokka, P. R. MacNeilage, G. C. De Angelis, and D. E. Angelaki. Estimating distance during self-motion: A role for visual-vestibular interactions. *Journal of Vision*, pages 1–16, 2011.

[12] T. C. Dong-Si and A. I. Mourikis. Estimator initialization in vision-aided inertial navigation with unknown camera–imu calibration. In *International Conference on Intelligent Robot and System*, 2012.

[13] A. Farrell. *Aided Navigation: GPS and High Rate Sensors*. McGraw-Hill, 2008.

[14] C. R. Fetsch, G. C. De Angelis, and D. E. Angelaki. Visual-vestibular cue integration for heading perception: Applications of optimal cue integration theory. *Eur J Neurosci.*, pages 1721–1729, 2010.

[15] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, pages 381–395, 1981.

[16] P. Gemeiner, P. Einramhof, and M. Vincze. Simultaneous motion and structure estimation by fusion of inertial and vision data. *The International Journal of Robotics Research*, pages 591–605, 2007.

[17] H. Goldstein. *Classical Mechanics*. Addison-Wesley, 1980.

[18] C. X. Guo and S. I. Roumeliotis. Observability analysis and consistency improvement. In *IEEE International Conference on Robotics and Automation*, 2013.

[19] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transaction on Pattern Recognition and Machine Intelligence*, pages 580–593, 1997.

[20] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*, volume 2. Cambridge Univ Press, 2000.

[21] R. Hermann and A. J. Krener. Nonlinear controllability and observability. *Transaction On Automatic Control*, pages 728–740, 1977.

[22] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis. Towards consistent vision-aided inertial navigation. In *International Workshop on the Algorithmic Foundations of Robotics*, 2012.

[23] E. Jones and S. Soatto. Visual–inertial navigation, mapping and localization: A scalable real-time causal approach. *The International Journal of Robotics Research*, pages 407–430, 2011.

[24] J. Kelly and G. Sukhatme. Visual–inertial simultaneous localization, mapping and sensor-to-sensor self-calibration. *International Journal of Robotics Research*, pages 56–79, 2011.

[25] J. Kim and S. Sukkarieh. Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, pages 62–71, 2007.

[26] D. G. Kottas, J. A. Hesch, S. L. Bowman, and S. I. Roumeliotis. On the consistency of vision-aided inertial navigation. In *International Symposium on Experimental Robotics*, 2012.

[27] M. Lhuillier. Automatic structure and motion using a catadioptric camera. In *IEEE Workshop on Omnidirectional Vision*, 2005.

[28] M. Li and A. I. Mourikis. Improving the accuracy of ekf-based visual inertial odometry. In *International Conference on Robotics and Automation*, 2012.

[29] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 61–62, 1987.

[30] H. C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Royal Society of London B: Biological Sciences*, pages 385–397, 1980.

[31] T. Lupton and S. Sukkarieh. Removing scale biases and ambiguity from 6dof monocular slam using inertial. In *International Conference on Robotics and Automation*, 2008.

[32] T. Lupton and S. Sukkarieh. Efficient integration of inertial observations into visual slam without initialization. In *International Conference on Inteligent Robot and System*, 2009.

[33] T. Lupton and S. Sukkarieh. Visual–inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *Transaction on Robotics*, pages 61–76, 2012.

[34] P. R. MacNeilage, M. Banks, G. C. S. De Angelis, and D. E Angelaki. Vestibular heading discrimination and sensitivity to linear acceleration in head and world coordinates. *Journal of Neuroscience*, pages 9084–9094, 2010.

[35] A. Martinelli. Closed-form solution for attitude and speed determination by fusing monocular vision and inertial sensor measurements. In *International Conference on Robotics and Automation*, 2011.

[36] A. Martinelli. State estimation based on the concept of continuous symmetry and observability analysis: The case of calibration. *IEEE Transactions on Robotics*, pages 239–255, 2011.

[37] A. Martinelli. Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale and bias determination. *IEEE Transactions on Robotics*, pages 44–60, 2012.

[38] A. Martinelli. Closed-form solution of visual–inertial structure from motion. *International Journal of Computer Vision*, 2013.

[39] A. Martinelli. Visual-inertial structure from motion: Observability and resolvability. In *International Conference on Intelligent Robot and System*, 2013.

[40] D. M. Merfeld, L. Zupan, and R. J. Peterka. Humans use internal models to estimate gravity and linear acceleration. *Nature*, pages 615–618, 1999.

[41] L. Mirsky. *An Introduction to Linear Algebra*. ISBN 0-486-66434-1, 1990.

[42] F. M. Mirzaei and S. I. Roumeliotis. A kalman filter-based algorithm for imu-camera calibration: Observability analysis and performance evaluation. *IEEE Transactions on Robotics*, pages 1143–1156, 2008.

[43] A. I. Mourikis, N. Trawny, and S. I. Roumeliotis. Vision-aided inertial navigation for spacecraft entry, descent, and landing. *IEEE Transactions on Robotics*, pages 264–280, 2009.

[44] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 756–770, 2004.

[45] D. Nistér. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, pages 321–329, 2009.

[46] W. H. Press, S. A. Teukolsky, S. T. Vetterling, and B. T. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2007.

[47] G. Quian, Q. Zheng, and R. Chellappa. Reduction of inherent ambiguities in structure from motion problem using inertial data. In *IEEE International Conference on Image Processing*, 2000.

[48] D. Scaramuzza. 1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. *International Journal of Computer Vision*, pages 74–85, 2011.

[49] D. Strelow and S. Singh. Motion estimation from image and inertial measurements. *International Journal of Robotics Research*, 2004.

[50] J. Tardif, Y. Pavlidis, and K. Daniilidis. Monocular visual odometry in urban environments using an omnidirectional camera. In *International Conference on Intelligent Robot and System*, 2008.

[51] C. Troiani, A. Martinelli, C. Laugier, and D. Scaramuzza. 1-point-based monocular motion estimation for computationally- limited micro aerial vehicles. In *European Conference on Mobile Robotics*, 2013.

[52] M. Veth and J. Raquet. Fusing low-cost image and inertial sensors for passive navigation. *Journal of the Institute of Navigation*, 2007.

[53] S Weiss. *Vision Based Navigation for Micro Helicopters*. PhD thesis, Diss. ETH No. 20305, 2012.

[54] S. Weiss, D. Scaramuzza, and R. Siegwart. Monocular-slam-based navigation for autonomous micro helicopters in gps-denied environments. *Journal of Field Robotics*, 2011.

[55] M. Wexler, F. Panerai, I. Lamouret, and J. Droulez. Self-motion and perception of stationary objects. *Nature*, pages 85–88, 2001.

[56] O. J. Woodman. *An introduction to inertial navigation*. Technical Report, University of Cambridge, Computer Laboratory, UCAM-CL-TR-696, 2007.