DISS. ETH NO. 20628

Real-Time Scalable Structure from Motion: From Fundamental Geometric Vision to Collaborative Mapping

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Laurent Kneip

Diplom-Ingenieur Univ. Friedrich-Alexander
Universität Erlangen-Nürnberg, Germany

Born November 23rd, 1982 in Ettelbruck
Citizen of Luxembourg

accepted on the recommendation of

Prof. Roland Siegwart,	ETH Zurich
Prof. Marc Pollefeys,	ETH Zurich
Prof. Davide Scaramuzza,	University of Zurich

2012

Abstract

Good egomotion estimation forms the backbone of any modern high-performance localization and navigation system. In contrast to global positioning systems or laser-based range measurement devices, the use of cameras represents an increasingly interesting alternative promising the applicability in a vast number of related scenarios. The range of possible fields of application easily extends from indoor to outdoor, small-scale to large-scale, and under-water to aerial operations. The computer vision community has investigated camera-based egomotion estimation for more than three decades now—research that has lead to impressive results. Two fundamental ways have been pursued. The first one consists of a purely geometric approach, where the incremental transformation between consecutive images is each time computed by absolute or relative camera pose algorithms, based on the identified feature correspondences. The second approach additionally takes time information into account, and estimates priors about the relative camera displacement by means of a motion model. Each modality can be extended by taking additional sensorial information into account, such as inertial readings from an IMU, GPS data, or laser range information.

The present dissertation represents a contribution towards purely geometric visual egomotion-estimation approaches, which are of major importance for initializing model-based solutions or robustifying estimation of motion with challenging dynamics. The scope ranges from fundamental algebraic geometry to the practical realization of systems for real-time motion estimation on real-life image sequences. The main cornerstones of this dissertation are given by a number of novel geometrical solutions for absolute and relative camera-pose computation in the calibrated case. The presented minimal solution for absolute camera-pose computation sets a new standard in terms of efficiency and numerical robustness. This algorithm is then extended to the non-perspective case via minimal and linear complexity n-point solutions. Finally, the derivation of an intuitive novel epipolar constraint leads to a minimal solution for direct translation-independent computation of the

relative rotation between two frames. Notably, the resulting cost function provides the advantage of remaining robust in vanishing parallax situations. The algorithms all employ a compact formalism based on unit bearing vectors only, which provides increased efficiency and the generality of being applicable to any kind of optical system. Some of the mentioned problems involve the solution of multi-variate polynomial equation systems, which is tackled by a dedicated framework for computing Gröbner bases and automatically generating efficient code.

In the continuation, the presented minimal solutions for absolute and relative camera pose computation form the basic modules of a versatile, real-time visual-odometry implementation, which essentially covers the continuous motion estimation of a single camera. The software is embedded into the robotic operating system ROS, which is a powerful middle-ware for project management, inter-process communication, and sensor interfacing. By allowing the adaptation of a vast number of properties—including the feature and descriptor type, the camera calibration model, as well as the size of the bundle-adjustment window—, this framework provides efficient real-time operation in different scenarios. As a next scientific contribution, the basic vision-only mode is extended by directly incorporating inter-frame rotation priors obtained from an IMU into the geometric computation pipeline. The result is a compact and efficient approach for robust single-camera egomotion computation providing merits especially in situations of critical motion dynamics, where smoothness assumptions about the camera trajectory are potentially violated. In comparison to extensive model-based vision-inertial solutions, the presented algorithm provides advantages in terms of computational efficiency, and remains theoretically consistent in a relative context by taking only short-term integrals of gyroscopic signals into account. For further improvement of the accuracy, the real-time implementation can be easily combined with a loosely-coupled vision-inertial filter. Robust filter-state initialization is provided by a deterministic closed-form solution to compute the visual scale factor and the vertical direction inside the IMU frame.

The final part of this dissertation then moves towards multi-camera extensions of the presented visual-odometry pipeline. The first system performs real-time visual odometry on two rigidly-coupled cameras with non-overlapping fields of view. Metric scale information is recovered by merging the relative displacement information of two individual monocular visual odometry nodes. The dependency between both trajectories is expressed by the hand-eye calibration constraint. The second proposed system consists of a collaborative SLAM back-end merging the output of multiple freely moving cameras. It again represents an efficient modular design that first

executes visual odometry in each camera individually, and then merges the obtained keyframes along with relative transformation information into different submaps. Via a recognition of previously visited places, the system is able to close loops and merge different submaps. This allows to potentially reach a collaborative situation in which all cameras operate on one and the same map.

Keywords: Structure from motion, Geometric vision, Camera pose, Minimal solution, Gröbner basis, Visual odometry, Vision-inertial egomotion computation, Multi-camera system, Non-overlapping stereo odometry, Collaborative SLAM

Kurzfassung

Ein gute Abschätzung der Eigenbewegung bildet das Rückgrat eines jeden modernen, leistungsfähigen Lokalisations- und Navigationssystems. Verglichen mit globalen Positionierungssystemen oder Laser-basierter Entfernungsmessung bildet der Einsatz von Kameras hier eine zunehmends interessante Alternative, die in einer Vielzahl von Szenarien eingesetzt werden kann. Die Spannweite der möglichen Anwendungsgebiete reicht vom Innen- zum Ausseneinsatz, von klein- zu grossmassstäblichen Problemen, sowie von Unterwasser- bis hin zu Lufteinsätzen. Die visuelle Abschätzung der Eigenbewegung wurde im Bereich Computer Vision seit nun mehr als drei Jahrzehnten mit eindrücklichen Ergebnissen intensiv erforscht. Zwei grundlegende Lösungswege haben sich hierbei herauskristallisiert. Die erste Methode besteht aus einem rein geometrischen Lösungsansatz, bei dem basierend auf den gefundenen Bildpunkt-Korrespondenzen die inkrementelle Transformation zwischen aufeinanderfolgenden Bildern durch den Einsatz von Algorithmen zur Berechnung der absoluten und relativen Kamera-Position ermittelt wird. Der zweite Ansatz zieht zusätzlich noch die Zeitinformation in Betracht, und kann so Prädiktionen über die relative Kamerabewegung basierend auf einem Bewegungsmodell generieren. Beide Ausführungsarten können durch den Einsatz zusätzlicher Sensoren—wie zum Beispiel einer Inertialmesseinheit, eines GPS Empfängers, oder eines Laser-Entfernungsmessers—erweitert werden.

Die vorliegende Dissertation stellt einen Beitrag in Richtung rein geometrischer Ansätze für visuelle Bewegungsabschätzung dar, welche für die Initialisierung von modellbasierten Lösungen sowie für robuste Schätzung von Bewegungen mit anspruchsvoller Dynamik von zentraler Bedeutung sind. Der Umfang reicht von grundlegender algebraischer Geometrie bis hin zur praktischen Umsetzung von Echtzeitsystemen zur kontinuierlichen Bewegungsschätzung über reale Bildsequenzen. Die Grundpfeiler dieser Dissertation bestehen aus einer Reihe von neuartigen geometrischen Lösungen für die Berechnung der absoluten und relativen Kamera-Position, wobei nur

der Fall von kalibrierten Kameras behandelt wird. Durch eine neuartige Minimallösung für die Berechnung der äusseren Kameraausrichtung werden neue Massstäbe im Sinne von Effizienz und numerischer Robustheit gesetzt. Dieser Algorithmus wird in der Folge mit minimalen sowie n-Punkt Lösungen auf den nicht-perspektivischen Fall erweitert. Die Herleitung einer intuitiven und neuartigen Epipolarbedingung führt schlussendlich zu einer Minimal-Lösung zur Berechnung der relativen Rotation zwischen zwei Kamerabildern, welche sich als unabhängig von der relativen Translation erweist. Die hierbei auftretende Kostenfunktion bietet darüberhinaus den Vorteil der Robustheit in Situationen in welchen die Norm der relativen Translation verschwindend gering ist. Ein hoher Grad an Effizienz sowie die Verallgemeinerung auf unterschiedlichste optische Systeme werden in allen Algorithmen durch eine ausschliessliche Formulierung in Funktion von normalisierten Richtungsvektoren erreicht. Einige der erwähnten Probleme erfordern die Lösung von multivariaten, polynomialen Gleichungssystemen, welche durch den Einsatz einer dedizierten Software zur Berechnung von Gröbner Basen und Erstellung von effizientem Code in Angriff genommen werden.

Die vorgestellen Minimallösungen für die Berechnung der absoluten und relativen Kamera-Position bilden im weiteren Verlauf die Basiselemente einer vielseitigen Echtzeitimplementierung zur Berechnung der sogenannten visuellen Odometrie, welche im Wesentlichen die kontinuierliche Bewegungsabschätzung einer Kamera umfasst. Die Software ist in das Betriebssystem ROS integriert, welches eine umfassende robotische Umgebung für Projektverwaltung, Interprozess-Kommunikation, und Ansteuerung von unterschiedlichsten Sensoren und Aktoren darstellt. Die Möglichkeit der Anpassung einer Vielzahl an Eigenschaften—wie zum Beispiel des Verfahrens zur Bildpunkt-Extraktion und -Beschreibung, des Kameramodells, sowie der Grösse des Opimierungsfensters für den lokalen Bündelausgleich—erlaubt eine hohe Effizienz und gute Anwendbarkeit in unterschiedlichen Szenarien. Ein weiterer wissenschaftlicher Beitrag besteht aus der Hinzunahme einer Inertialmesseinheit zur Gewinnung von Prädiktionen über die relativen Rotationen zwischen den Bildpaaren, welche direkt in den Prozess der geometrischen Berechnungen miteinbezogen werden. Das Ergebnis ist ein kompakter und effizienter Ansatz für die robuste Berechnung der Eigenbewegung einer Kamera, welcher besonders in kritischen Bewegungssequenzen, in denen Annahmen über die Glätte der Bewegung nicht mehr geltend sind, zum Tragen kommt. Im Vergleich zu umfassenden, modell-basierten Optointertiallösungen bietet der vorgestellte Algorithmus Vorteile in der rechnerischen Effizienz, und bleibt dank der Beschränkung auf Kurzzeitintegrale der Gyroskopsignale in einem relativen Zusammenhang theoretisch konsistent. Die Echtzeitim-

plementierung kann bequem mit einem lose-gekoppelten optoinertialen Filter verbunden werden, wodurch eine weitere Verbesserung der Genauigkeit erzielt wird. In diesem Zusammenhang wird durch eine abgeschlossene Lösung zur Berechnung des metrischen Massstabs sowie der vertikalen Richtung im Sensorreferenzsystem eine weitere deterministische Lösung zur robusten Filterinitialisierung bereitgestellt.

Der abschliessende Teil dieser Dissertation bewegt sich entgegen der Erweiterung des vorgestellten visuellen Odometriesystems auf Mehrkamerasysteme. Die erste Lösung berechnet visuelle Odometrie auf zwei starr miteinander verkoppelten Kameras mit nicht-überlappenden Sichtfeldern. Der metrische Massstab wird durch die Abhängigkeit der relativen Verschiebungsinformationen von zwei monokularen visuellen Odometrien wiederhergestellt, welche durch die sogenannte *hand-eye*-Kalibrierungsbedingung ausgedrückt wird. Das zweite vorgeschlagene System besteht aus einer kollaborativen, simultanen Lokalisierungs- und Kartografierungseinheit welche die Ausgabe von mehreren sich frei bewegenden Kameras miteinander verbindet. Diese Lösung verkörpert erneut einen effizienten und modularen Entwurf, in dem zuerst individuelle visuelle Odometrie auf jeder Kamera gerechnet wird, um dann die ausgegebenen Schlüsselbilder mit relativen Transformationsinformationen in unterschiedliche Teilkarten einzubinden. Das System ist zudem in der Lage bereits bekannte Kartenteile wiederzuerkennen um so Schleifen zu schliessen und Kartenteile miteinander zu verschmelzen. Auf diesem Wege wird potentiel eine kollaborative Situation erreicht, in der die Informationen von allen Kameras in der gleichen Karte miteinander vereint werden.