Robust and Accurate Localization by Visual-Inertial Odometry

Milad Ramezani

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Infrastructure Engineering
THE UNIVERSITY OF MELBOURNE

September 2018

Milad Ramezani: *Robust and Accurate Localization by Visual-Inertial Odometry*
© September 2018

SUPERVISORS:
Dr. Kourosh Khoshelham
Prof. Clive Fraser

Abstract

REAL-TIME, accurate and seamless localization forms the backbone of various location-aware applications such as autonomous navigation. For instance, in a driverless car or in an autonomous robot operating outdoors or indoors, a navigation system is responsible for providing location-related knowledge so that the system can accomplish tasks with respect to this knowledge. Tasks such as finding an address, gripping or tracking an object, making a location-based decision and so on so forth are highly dependent on a localization. Global Navigation Satellite System (GNSS) is a key technology in meeting this demand, however, in indoor or densely built-up areas, GNSS observations are not fully available. Intense efforts have been devoted to identifying a localization approach which could complement GNSS with inertial sensors, laser scanners, cameras, wireless technologies or the integration of these sensors . Among these localization approaches, the use of cameras has gained popularity over recent years due to their availability, affordability and operability. Applications of cameras range from indoor environments to outdoor environments, micro-scale to macro-scale, under-water to outer space. By having a sequence of images taken from a single camera or multiple cameras, it is possible to estimate the location using multiple view geometry. This localization can further be extended by benefiting from additional information acquired from sensors such as Inertial Navigation Systems (INSs), GNSS observations or laser scanner data.

Beginning with a discussion of limitations of current techniques, this research aims to improve the performance of existing approaches that employ the integration of cam-

eras with an INS, which is known as Visual-Inertial Odometry (VIO). The outcome is the presentation of two novel VIOs: a stereo-VIO in which a stereo camera is exploited to improve the robustness of localization by simultaneously taking advantage of visual constraints across the stereo images and sequential images taken in time, and an Omnidirectional VIO (OVIO) in which an omnidirectional camera with a field of view larger than 180 degrees is employed to explore a larger scene surrounding the system. This way, features from a larger portion of the environment around the system will contribute to the estimation resulting in more accurate localization. The proposed approaches have been evaluated with photorealistic-synthetic and real-world data. The results demonstrate the superiority of the proposed methods compared to their counterparts.

Keywords: Localization, Visual-inertial odometry, SLAM, Kalman filter, Stereo camera, Omnidirectional camera, Multi-view geometry

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Milad Ramezani, September 2018

Acknowledgements

Taking a look back at the past three and a half years, I find out that a PhD program is not only a time of hard work and challenging your abilities and skills, and pushing yourself beyond your personal limits, but also it provides a great opportunity to get involved in teaching and social experiences such as team work and attending conferences. The PhD journey, with its many breakthroughs and failures, under a continuous pressure to keep focused and think critically and positively, can lead to ultimate achievement. I would not have had these factors without their support in my life. Therefore, I would like to particularly thank the following persons:

My supervisor Dr. Kourosh Khoshelham for his continuous support technically and emotionally. It was a great chance for me to pursue my PhD under his supervision.

My co-supervisor Prof. Clive Fraser for accepting me as his student. It was always my desire to be under his supervision when I was doing my master and reading his books and publications.

My brother Dr. Alireza Ramezani for sharing his academic experience with me and broadening my view about research and motivating me to overcome problems.

My office mates, Fuqiang Gu, Debaditya Acharya, Ha Tran and Hanxian He for exchanging information and constructive discussions.

Dr. Lurent Kneip for collaborating with me and giving me very helpful comments on my IROS paper.

Robotics and Perception Group led by Prof. Davide Scaramuzza in the Department of Informatics at the University of Zurich, especially Zichao Zhang for answering my

unlimited questions patiently.

The last but not the least, my parents Nasrin and Ali for providing a wonderful environment for me and my brother back in Iran, my country, to be able to stand at this point, and for their firm support that we have never experienced any deficiencies in our life, and my wife Maryam Haghighat for her kindness and being always along with me in ups and downs and tolerating the distance existing between us during my PhD.

Also, thank you to everybody who I forgot to mention for their help and for making me feel good.

Melbourne, May 2018

Acronyms

2D	2-Dimensional
3D	3-Dimensional
6D	6-Dimensional
AN	Autonomous Navigation
AoA	Angle of Arrival
ARMSE	Average Root Mean Square Error
BIM	Building Information Model
CIM	City Information Model
CL	Cooperative Localization
DGPS	Difference GPS
DoG	Difference of Gaussian
DSRC	Dedicated Short Range Communication
ECEF	Earth Centered Earth Fixed
EKF	Extended Kalman Filter
FoV	Field of View
GIS	Geospatial Information System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
IMU	Inertial Measurement Unit
INS	Inertial Navigation System

IPS	Indoor Positioning System
IR	Infrared Radiation
ITS	Intelligent Transport System
KITTI	Karlsruhe Institute of Technology and Toyota Institute
KLT	Kanade-Lucas-Tomasi
LBA	Local Bundle Adjustment
LBS	Location-Based Service
LoG	Laplacian of Gaussian
LPR	Local Position Radar
LSD-SLAM	Large Scale Direct-SLAM
MSAC	M-estimator Sample Consensus
MSCKF	Multi State Constraint Kalman Filter
ORB-SLAM	Oriented FAST and Rotated BRIEF-SLAM
OVIO	Omnidirectional Visual Inertial Odometry
PDR	Pedestrian Dead Reckoning
RFID	Radio Frequency IDentification
RGB-D SLAM	Red Green Blue-Depth SLAM
RMSE	Root Mean Square Error
RSS	Radio Signal Strength
RTK	Real Time Kinematic
RToF	Roundtrip Time of Flight
SIFT	Scale Invariant Feature Transform
SLAM	Simultaneous Localization and Mapping
SURF	Speed Up Robust Features
SWF	Sliding Window Filter
TDoA	Time Difference of Arrival
UTC	Coordinated Universal Time
UWB	Ultra-WideBand

VANETs	Vehicular Ad hoc NETworks
V-I	Vehicle-Infrastructure
V-V	Vehicle-Vehicle
WLAN	Wireless Local Area Network
ZUPT	Zero Velocity Update

Publications

List of the peer-reviewed publications:

JOURNAL ARTICLES

1. **Ramezani, M.**, Khoshelham, K. and Fraser, C., 2018. Pose estimation by Omnidirectional Visual-Inertial Odometry. In *Robotics and Autonomous Systems*, 105, pp. 26-37.

2. **Ramezani, M.** and Khoshelham, K., 2018. Vehicle Positioning in GNSS-Deprived Urban Areas by Stereo Visual-Inertial Odometry. In *IEEE Transactions on Intelligent Vehicles*. IEEE.

REFEREED CONFERENCE PAPERS

1. **Ramezani, M.**, Khoshelham, K. and Kneip, L., 2017, September. Omnidirectional Visual-Inertial Odometry Using Multi-State Kalman Filter. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on* pp. 1317-1323. IEEE.

2. **Ramezani, M.**, Acharya, D., Gu, F. and Khoshelham, K., 2017, September. Indoor Positioning by Visual-Inertial Odometry. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, pp. 371-376.

3. Khoshelham, K., **Ramezani, M.**, 2017. Vehicle positioning in the absence of GNSS signals: Potential of visual-inertial odometry. In *Urban Remote Sensing Event (JURSE), 2017 Joint*, pp. 1-4. IEEE.

To my parents

Contents

1 Introduction **1**
- 1.1 Research Problem and Motivation . . . 1
- 1.2 Aims and Research Challenges . . . 6
- 1.3 Contributions . . . 8
- 1.4 Outline . . . 9

2 Background **11**
- 2.1 Introduction . . . 11
- 2.2 Absolute Localization . . . 12
  - 2.2.1 Absolute Localization in a Global Frame . . . 12
  - 2.2.2 Absolute Localization in a Local Frame . . . 15
- 2.3 Relative Localization . . . 18
  - 2.3.1 Inertial Localization . . . 18
  - 2.3.2 Optical Sensor-Based Localization . . . 20
- 2.4 Multi-sensor Localization . . . 27
  - 2.4.1 Sensor-Fused Localization . . . 27
  - 2.4.2 Cooperative Localization . . . 28
- 2.5 Visual-Inertial Odometry (VIO) . . . 30
  - 2.5.1 VIO: Survey from a Hardware Perspective . . . 30
  - 2.5.2 VIO: Survey from an Estimation Perspective . . . 32
- 2.6 Summary . . . 33

3 Visual-Inertial Odometry Using a Single Camera **35**
- 3.1 Introduction . . . 35
- 3.2 Conventional MSCKF . . . 35
  - 3.2.1 Formulation of MSCKF State Vector . . . 36
  - 3.2.2 INS state propagation . . . 37
  - 3.2.3 Linearised System Model . . . 39
  - 3.2.4 Discrete-time implementation . . . 40
  - 3.2.5 State Augmentation . . . 41
  - 3.2.6 Measurement Model . . . 42
  - 3.2.7 Update . . . 45
- 3.3 A Simulation Using Predefined Features . . . 47
- 3.4 Summary . . . 48

4 Visual-Inertial Odometry Using a Stereo Camera **49**
4.1 Introduction . . . 49
4.2 Methodology of the Stereo MSCKF . . . 50
4.3 Feature Detection, Matching and Tracking . . . 53
4.3.1 Feature Detection . . . 54
4.3.2 Feature Matching . . . 55
4.3.3 Feature Tracking . . . 55
4.4 Summary . . . 56

5 Visual-Inertial Odometry Using an Omnidirectional Camera **59**
5.1 Introduction . . . 59
5.2 Calibration . . . 60
5.3 Methodology of the Spherical MSCKF . . . 63
5.3.1 The Proposed Measurement Model . . . 64
5.4 Summary . . . 68

6 Experiments and Results **71**
6.1 Introduction . . . 71
6.2 Stereo Visual-Inertial Odometry . . . 71
6.2.1 Feature Detection, Matching and Tracking . . . 72
6.2.2 Accuracy of Stereo Visual-Inertial MSCKF . . . 75
6.3 Omnidirectional Visual-Inertial Odometry . . . 76
6.3.1 Camera Calibration . . . 79
6.3.2 Experiments with Synthetic Data . . . 81
6.3.3 Real-World Experiments . . . 92
6.4 Summary . . . 96

7 Conclusion and Future Work **99**
7.1 Conclusion . . . 99
7.2 Summary of Contributions . . . 101
7.3 Perspectives . . . 102

A Appendix **105**
A.1 3D Scene Reconstruction . . . 105

Bibliography **107**

List of Figures

1.1 Overview of sensors mostly used for localization, and their combination in the form of either sensor fusion or cooperation. 3
1.2 **Left**: Mobile mapping system used by KITTI (Geiger et al., 2012). **Right**: A Lidar sensor integrated with GNSS/INS (A.Arnold and Persaud, 2012). 4

2.1 Positioning solution obtained from triple satellite ranging observations. . 13
2.2 Effect of densely built-up environments on GNSS observations. 14
2.3 An INS with its inertial components and their orientation. 19
2.4 Omnidirectional cameras and their image projection compared to a perspective camera. 21
2.5 An illustration of different methods of VO. 22
2.6 Visual result for LSD-SLAM using an omnidirectional camera (Caruso et al., 2015). 25
2.7 A CL scenario in VANET. 29

3.1 The genral workflow of the mono MSCKF. 36
3.2 **Left**: The simulated indoor environment with the generated feature points. **Right**: The top view of the simulated indoor environment and the estimated trajectory (blue) compared with the INS trajectory (red) and the ground truth (green). 47
3.3 **Left**: Cumulative distribution of translational error for the perspective VIO approach and the INS integration. **Right**: Cumulative distribution of rotational error for the perspective VIO approach and INS integration. 48

4.1 The general workflow of the stereo MSCKF. Images captured from a pair of cameras at each timestep are conducted to the stereo matching process. The stereo matching imposes additional constraints resulting in better estimation. 50
4.2 Steps of feature detection and tracking in mono (**Right**) and stereo (**Left**) sequence of images. 54

5.1 Projection process from 3D space to the 2D sensor plane in a dioptric camera (**Left**) and a catadioptric system (**Right**) (Micušık, 2004). 60
5.2 Different components of projection in a single view-point omnidirectional camera and the triangulation between view C_1 and C_i. 64

6.1 Sample frames from the three tested KITTI datasets. (a) dataset 0001. (b) dataset 0005. (c) dataset 0020. 73
6.2 Features tracked in mono vision (**Top**) and stereo vision (**Bottom**) from KITTI dataset 0001. 74
6.3 Statistics on the number of tracked features in each frame over the three datasets 0001, 0005, and 0020 for both mono and stereo modes. 74
6.4 Trajectories obtained by the stereo VIO as compared with the mono and INS integration for the three KITTI traverses. (a) dataset 0001. (b) dataset 0005. (c)dataset 0020. 77
6.5 Cumulative distribution of errors for the stereo method compared with mono and INS integration for the tested KITTI datasets 0001 (a), 0005 (b) and 0020 (c). 78
6.6 Distribution of reprojection errors. 80
6.7 Sample images used in the calibration including the projected points and their corresponding reprojected points. 83
6.8 Estimated scene including checker-board corners (black dots) and the camera pose (coloured axes). 84
6.9 **Left**: Histogram of estimated residuals. **Right**: Normal probability plot of estimated residuals 84
6.10 Example images from the sequence recorded using a simulated perspective camera (**top**) and a simulated fish-eye camera (**bottom**). 85
6.11 Features after the tracking process in a perspective view (**left**) and fisheye view (**right**). The green pluses show the features tracked along the image sequence. 85
6.12 Top view of the estimated trajectory for perspective VIO (blue), and omnidirectional VIO (magenta) compared with the INS-only integration (red) and the ground truth (green). 86
6.13 **Left**: Cumulative distribution of translational error for the perspective VIO, the omnidirectional VIO and the INS-only integration. **Right**: Cumulative distribution of rotational error for the perspective VIO, the omnidirectional VIO and the INS-only integration. 87
6.14 Sample images of the urban canyon dataset simulated by Zhang et al. (2016) in different FoV cameras namely perspective (**top**), fish-eye (**middle**) and catadioptric (**bottom**). 88
6.15 **Left**: Boxplot of the number of tracked features in each frame for perspective, fisheye and catadioptric views. **Right**: Boxplot of the number of frames from which each feature is observed. 89
6.16 3D estimated trajectories from different integration in the global frame. . . 90
6.17 **Left**: Cumulative distribution of the translational errors for the urban canyon synthetic dataset. **Right**: Cumulative distribution of the rotational errors on the same trajectory. 91
6.18 The sensors used in the real experiment, these being installed on a board that was attached to the roof of a car using rubber magnets. 93
6.19 **Left**: A raw image captured by the fish-eye camera looking upward. **Right**: Features tracked between two consecutive images. 94

6.20 **Top**: 2D top view of the estimated trajectory in frame $\{\mathcal{E}\}$ (**left**), and the comparison between the estimated trajectory and the GNSS ground truth (**right**). **Middle**: Estimated velocity with respect to frame $\{\mathcal{E}\}$ (**right**), and attitude with respect to local navigation frame (**left**). **Bottom**: Acceleration and angular rate biases with respect to frame $\{\mathcal{B}\}$. 95

List of Tables

2.1 Summary of INS and GNSS complementary characteristics according to (Groves, 2013). 27

6.1 Comparison of Tracking Performance in Mono and Stereo on KITTI Traverses 0001, 0005 and 0020. 75
6.2 Comparison of Average Root Mean Square Error (ARMSE) of INS Integration, Mono VIO, and Stereo VIO on KITTI Traverses 0001, 0005 and 0020. . 79
6.3 RMSE of reprojection and the corresponding standard deviations. 82
6.4 Comparison of Average and Total RMSE of the proposed spherical VIO on fisheye and catadioptric VIOs against perspective VIO. 91

Chapter 1
Introduction

REAL-TIME, accurate and seamless localization is a primary requirement for many modern location-based applications such as autonomous navigation in both indoor and outdoor environments, mobile mapping systems and pedestrian localization. The Global Navigation Satellite System (GNSS) is a key technology needed to meet this demand in outdoor environments. However, in indoor or densely built-up areas, e.g. urban canyons, GNSS observations are not always available. To fill this gap, numerous localization approaches have been proposed for more than three decades. Some researchers have exploited the potential of individual sensors such as accelerometers, gyroscopes, laser scanners, cameras and so forth, while others have integrated data captured by different types of sensors. Despite these significant efforts, the provision of a localization system that could complement GNSS is still an active research area. This introductory chapter addresses the importance of localization and its place in many location-aware applications motivating this work. This is followed by an identification of several objectives and open challenges, as well as an overview of the presented contributions.

1.1 Research Problem and Motivation

Over the last few decades, truly real-time and continuous ego-motion estimation of moving entities such as vehicles, robots and pedestrians has drawn the attention of researchers due to its contribution to many civilian applications including Location-Based Services (LBS), Intelligent Transport Services (ITS) and Autonomous Navigation (AN). A location-

based service, which is a software service, employs location information to carry out tasks for entertainment purposes as well as tasks targeting business or assistance in emergency situations (Deuker, 2008). For example, in an accident scenario, both the location of the incident and the location determination of local rescue centres are crucial to dispatch an aid in the fastest time. Meanwhile, the location of the rescue team has to be real-time and accurate for intercommunication purposes. In an intelligent transportation system, accurate and real-time tracking of traffic features such as cars provide better information of transportation for users leading to safer, better coordinated and smarter utilization of transport networks. In an autonomous navigation system, which is, nowadays, a hot topic for driverless cars, all three elements of accuracy, seamlessness and real-time operability are crucially required. On the basis of this location information, a driverless car is able to communicate with other features, e.g vehicles or pedestrians, in its proximity and accomplish its missions.

Since the advent of satellite positioning, known as GNSS, localization has become achievable. Today, using the techniques of differential global Positioning, and Real-Time Kinematic (RTK) positioning, it is feasible to locate static objects and track moving objects. Nevertheless, in GNSS deprived environments such as urban canyons and indoor areas, satellite observations are not always available. In these areas, GNSS signals are mostly blocked or jammed by obstacles such as structures and their components, or they suffer interference by wireless equipment. Moreover, using only GNSS observations, attitude, which is crucial for dynamic platforms such as aerial vehicles, is not fully estimated.

For the purpose of motion estimation, many techniques have been proposed utilizing either inertial, optical or wireless sensors individually or in combination in a form of sensor fusion or Cooperative Localization (CL). Figure 1.1 provides an overview of different sensors commonly used for localization, as well as their integration via either sensor fusion or cooperative localization.

Inertial sensors, including accelerometers and gyroscopes, are the most common sensors that are used in conjunction with GNSS.

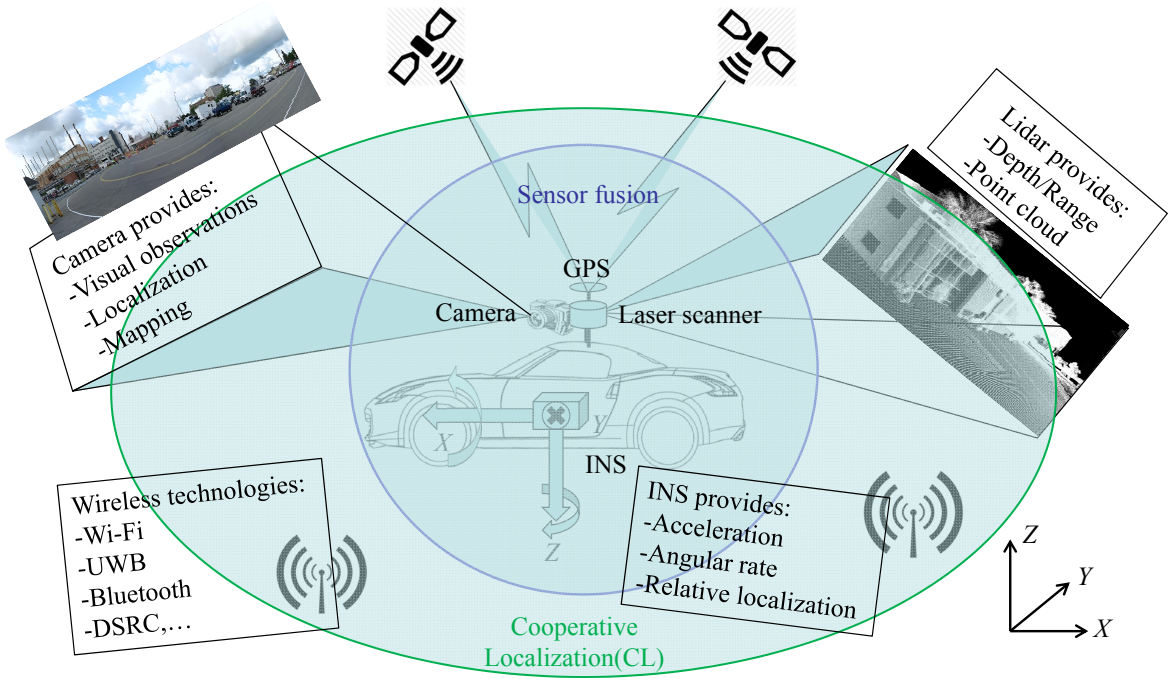


Figure 1.1: Overview of sensors mostly used for localization, and their combination in the form of either sensor fusion or cooperation.

Inertial localization is carried out by double integrating acceleration data from the accelerometer and combining it with attitude measurements from the gyroscope. The benefit of inertial localization is that it is purely self-contained and therefore does not require any external references. Furthermore, inertial sensors can provide seamless localization. Nevertheless, the pose estimated by these sensors suffers from drift over time due to the incremental nature of localization (Groves, 2013). The combination of inertial sensors that provide a navigation solution is known as an Inertial Navigation System (INS). By integrating INS with GNSS, accurate 6D pose estimation can be achieved. However, the solution will dramatically diverge after a short period of GNSS signal loss.

Recently, Lidar sensors are being increasingly used in mobile mapping systems to generate the map and simultaneously estimate the sensor motion (Nüchter et al., 2007; Taguchi et al., 2013). In Lidar-based localization systems, motion estimation is carried out by registering successive laser scans (Khoshelham, 2015). A laser scanner integrated with GNSS/INS, within the mobile mapping system developed by the Karlsruhe Institute and

Toyota Technological Institute(KITTI) (Geiger et al., 2012) is illustrated in Figure 1.2. In the driverless cars of Google and Ford, laser scanners will be used for detecting road obstacles and navigation-related objects, and also for motion estimation in the absence of GNSS observations (Hawkins, 2016). The drawback of laser scanners is their relatively high cost for a large number of users.

A state-of-the-art approach for localization is cooperative localization. cooperative localization systems use data fusion algorithms to integrate position-related information from a group of sensors or entities surrounding the object of interest, which can exchange data via radio signals (*cf.* Figure 1.1). Cooperative localization systems can be beneficial for motion estimation in GNSS deprived environments due to their exploitation of the maximum potential of position-related data. However, this technique is still not mature and is strongly dependant on radio communication, which restricts the operability of the method when sub-meter positioning accuracy is sought (Alam and Dempster, 2013).

Recent advances in the development of visual sensors have resulted in large-scale production of digital cameras with high quality, small size and low cost. Today, cameras can be mounted on any manned or unmanned vehicles due to their light weight and small size. The four elements - low cost, light weight, small size, and moderate consumption of electrical energy - make these sensors ideal for aiding navigation systems in areas

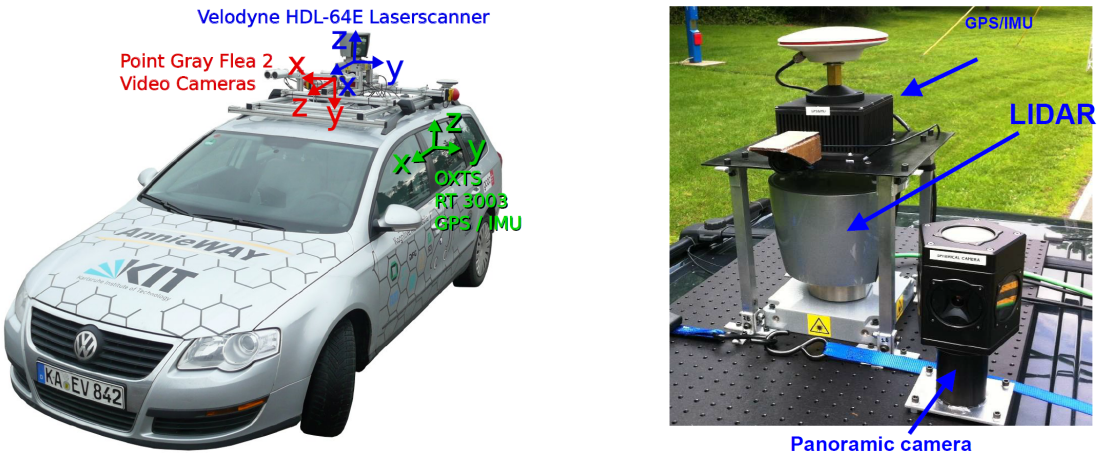


Figure 1.2: **Left**: Mobile mapping system used by KITTI (Geiger et al., 2012). **Right**: A Lidar sensor integrated with GNSS/INS (A.Arnold and Persaud, 2012).

where GNSS is neither fully functional nor its error tolerable.

Pioneers in photogrammetry and computer vision have long endeavoured to acquire geometric information from stereo views or multiple views of a scene typically captured by either a single or stereo cameras temporally, or using a multitude of cameras instantly. In fact, the idea is understanding the mathematics of the projection of a 3D structure onto the 2D image plane of imaging devices. This fact is not achievable to sufficient accuracy and reliability unless the interior geometry of the cameras and a proper exterior orientation with respect to a coordinate system are known. The parameters which define the interior geometry of a camera are called intrinsic or interior orientation parameters, while the parameters which describe the position and the orientation of a camera in space are known as exterior orientation parameters.

A challenge in this domain is motion estimation using image observations, which is known as Visual Odometry (VO). In general terms, VO aims at providing a 6D camera pose at each image acquisition in a sequence of images by reconstructing the 3D environment and back-projecting the scene points onto the image planes from which the scene points are visible, eventually minimizing the 2D residual errors between the back-projected points and their corresponding measurements through an optimization method. If the camera pose parameters are simultaneously estimated with the structure of the environment, the approach is known as Simultaneous Localization and Mapping (SLAM) (Davison, 2003; Mur-Artal et al., 2015; Khoshelham et al., 2013; dos Santos et al., 2016; Strasdat et al., 2010). In a typical visual SLAM algorithm (Davison, 2003), salient features are extracted and matched between pairs of images to reconstruct a map of the scene and estimate the sensor pose with respect to the reconstruction.

Despite the fact that in SLAM approaches the sensor position is estimated with respect to the map, the procedure for reconstructing the map itself is incremental, resulting in error accumulation and trajectory drift. Recent SLAM techniques such ORB-SLAM (Mur-Artal et al., 2015) and RGB-D SLAM (Endres et al., 2012; Khoshelham et al., 2013; dos Santos et al., 2016) detect loop closure and apply pose graph optimization (Strasdat et al.,

2010; Kümmerle et al., 2011) to improve the global consistency of the map and correct the trajectory drift. However, VO and SLAM are unlikely to achieve continuous pose estimation in areas that lack sufficient texture and geometric features.

Through integration of image observations with inertial measurements from an INS, the accuracy and resilience of localization in low-texture situations can be substantially improved. This approach, which is known as Visual-Inertial Odometry (VIO), has been an active research area for more than a decade (Mourikis and Roumeliotis, 2007; Kim and Sukkarieh, 2007; Li and Mourikis, 2013; Forster et al., 2015). VIO employs visual observations to compensate for INS error through an estimation model. Today, most VIO systems utilize image observations from a perspective camera with a limited Field of View (FoV), of generally less than 90°. Achieving a robust localization performance by the state-of-the-art VIO approaches remains a challenge in environments that lack salient features due to illumination changes or insufficient texture in the entire image sequences, as well as the presence of moving objects covering a significant area of the image from a perspective camera.

Overall, a review of different localization approaches illustrates that providing accurate, real-time and seamless localization remains a challenging problem and a fascinating, active research field which mainly motivates this work. Also, because inertial sensors and cameras are becoming increasingly more affordable, as illustrated by sensors in light-weight and small-scale devices such as smartphones, this work mainly focuses on exploring novel techniques to improve the positioning accuracy obtained by the combination of these two sensors, i.e. localization by visual-inertial odometry.

1.2 Aims and Research Challenges

By focusing on inertial and visual sensors and the integration of their measurements, this research aims to introduce novel VIO systems which can not only maintain localization accuracy at an acceptable range, e.g. better than 1 meter in a 100 meter GNSS outage, but

also provide continuous localization in real-time. To achieve these goals, in this study, the robustness of state-of-the-art VIOs has to be improved. However, this improvement imposes changes in the architecture of these systems. Therefore, a number of challenges related to robustness, structure and efficiency of the computation need to be addressed:

- **Robustness**: Existing vision-aided navigation systems are still lacking robustness in visually challenging environments. For instance, VOs cannot provide seamless ego-motion estimation in complex environments that lack sufficient features. Although the employment of an INS in VIOs resolves the issue of continuous pose estimation, these systems are likely to drift, largely in environments that consist of dynamic entities, e.g. vehicles and pedestrians. Indeed, all vision-aided navigation systems are highly sensitive to the accuracy of feature observations. Solutions such as sanitizing visual features and improving the vision geometry, or exploiting the maximum visual information surrounding the sensors, lead to better robustness of these systems.

- **Structure**: To improve the robustness of VIOs, a change in the architecture of these systems is warranted. For instance, as described later in this dissertation, a state-of-the-art VIO approach that employs a multitude of camera poses in ego-motion estimation is known as the Multi-State Constraint Kalman Filter (MSCKF). Despite the fact that this algorithm aims to exploit the maximum potential of constraints that a feature visible from multiple camera poses contributes, it is only suited for perspective cameras and not for cameras such as omnidirectional cameras. To integrate the omnidirectional images with an INS through MSCKF, a change in the architecture of the MSCKF is necessary.

- **Efficiency**: Many frameworks defined for visual and visual-inertial ego-motion estimation still struggle with the trade-off between performance and real-time operation. Vision-based SLAM algorithms suffer from computational complexity and in environments rich in features, some limitations exist in the form of employment

of the maximum geometric constraints that each visible scene feature creates in a sequence of images, thus resulting in a loss of information. This loss of information leads to less accurate localization. Nevertheless, this limitation in the number of camera poses in motion estimation is essential for real-time operability of the system. Improving the performance of a VIO system having the potential of employing the maximum number of geometric constraints created by visual information, while keeping the computational complexity linear in terms of the number of visual features, can possibly retain the balance between high performance and real-time operation.

1.3 Contributions

As mentioned earlier, the main focus of this thesis is upon providing localization by VIO in a way that the above described goals are met. Therefore, the historical achievements towards ego-motion estimation by visual and visual-inertial odometry systems have been studied and are discussed here in terms of robustness and efficiency. Upon realizing the best estimation approach, novel VIO systems, which contribute to the improvement of ego-motion estimation, will be introduced. The major contributions from this research comprise two new VIO approaches. One is based on the integration of a stereo camera with an INS, while the second benefits from 360° images captured by an omnidirectional camera. Both proposed systems constitute an enhancement of the very effective VIO method. They target a broad field of location-based applications in both indoor and outdoor environments.

- **Stereo Visual-Inertial Odometry**

  The first contribution of the conducted research is a stereo visual-inertial odometry system based on the MSCKF. In the proposed MSCKF-stereo VIO, temporal matching of features along the sequence of images constrains the estimated motion, while the stereo pair of images provides an additional constraint, leading to

better motion estimation. This way, not only the problem of scale estimation, which affects the performance of VOs with only a single bearing sensor, is resolved, but also the reconstruction of small and large-scale environments becomes possible, thus contributing to better accuracy of motion estimation.

- **Omnidirectional Visual-Inertial Odometry (OVIO)**

  The second contribution comprises a novel VIO approach in which an omnidirectional camera is integrated with an INS through the MSCKF. In theory, a camera which provides observations from larger FoVs can contribute to better localization due to the fact that more features from different angles are likely to be detected, resulting in stronger geometry. Moreover, through use of an omnidirectional camera, the likelihood of situations that the entire image is covered with dynamic entities will decrease and more usable regions in images will exist for feature extraction and matching. Since the images of an omnidirectional camera have high distortion, the calibration of these cameras, and an introduction of a space in which the distortion does not affect the estimation must be taken into account. To integrate the outputs of an INS with omnidirectional images through the MSCKF, a new measurement model is introduced.

The proposed systems developed in this research are evaluated using synthetic and real data. The performance of each system is compared to the conventional perspective VIO that uses the MSCKF, the INS-only integration and the ground truth.

1.4 Outline

The present dissertation is organised as follows: Chapter 2 provides a literature review covering state-of-the-art localization systems with different classifications. Since the focus of this work is on VIO systems, one section is allocated to the investigation of these systems from both hardware and estimation perspectives. Because the MSCKF filtering method is utilized as the estimator in this research, this methodology is described

in Chapter 3. Chapters 4 and 5 are then allocated to presentation of the formulation of the proposed methodologies, namely stereo MSCKF and spherical MSCKF, respectively. The evaluation of these novel approaches, based on their application to photorealistic-synthetic and real-world data is discussed in Chapter 6. Finally, conclusions are drawn and directions for future research are discussed in Chapter 7.

Chapter 2

Background

2.1 Introduction

THIS chapter discusses different localization systems with respect to a coordinate system. The localization is defined as either absolute or relative localization. Absolute localization refers to location estimation in a global or local coordinate system, whereas relative localization addresses location estimation relative to a previous location. The latter is sometimes called motion estimation. Typically, systems that provide localization on the basis of a global coordinate system require external references such as satellites as in the Global Positioning System (GPS) and Global Navigation Satellite System (GNSS), or infrastructures such as beacons in wireless technologies. Systems that are independent of external references are self-contained but provide localization in a local coordinate system. Today, these systems consist of electromechanical sensors or optical sensors. If various data captured by different on-board sensors is integrated, the method is called sensor fusion and if observations from on-board and off-board sensors, e.g. sensors surrounding the object of interest, are utilized via wireless technologies, the approach is known as Cooperative Localization (CL). All these techniques are thoroughly investigated in this chapter. Finally, a separate section is devoted to visual-inertial odometry to precisely investigate these systems in terms of hardware configurations and estimation algorithms.

2.2 Absolute Localization

Absolute localization systems endeavour to answer the question"*Where am I?*". Correctly answering this question is crucial for many Location-Based Services (LBSs) and mobile mapping systems. To understand "*where I am*", the position must be defined with respect to a reference coordinate system. This coordinate system can be either a global reference system such as WGS84 in GNSS, or a local coordinate system in which the whole environment of interest is defined. The former coordinate system is used for large-scale applications such as Intelligent Transport Systems (ITS) and Autonomous Navigation (AN) in driver-less cars. The latter reference system is broadly used in small-scale applications such as indoor localization or robotics when a robot, for instance, needs find its location with respect to the environment in which it moves. In either of these situations, prior knowledge about the known position of some external reference points is inevitable.

2.2.1 Absolute Localization in a Global Frame

As mentioned, satellite navigation systems that aim to provide a 3D localization solution world-wide are known as GNSS. GNSS-based location is carried out by receiving radio signals transmitted from orbiting satellites. The solution is subject to prior information about the satellite time and its orbiting parameters attained from a navigation data message, known as the ephemeris (Hofmann-Wellenhof et al., 2012). As illustrated in Figure 2.1, by simultaneously receiving observations from at least three satellites, the three dimensional position of an object of interest can be calculated. The intersection of ranging observations results in two positioning solutions, **P**1 and **P**2, one of which is on the surface of the Earth and the other is far from the Earth in space (Groves, 2013). However, because the receiver clock is not exactly synchronised with the satellite clock, an additional parameter known as receiver time error needs to be estimated, therefore the number of parameters that need to be estimated is four and consequently receiving observations from at least four satellites is necessary to estimate 3D position.

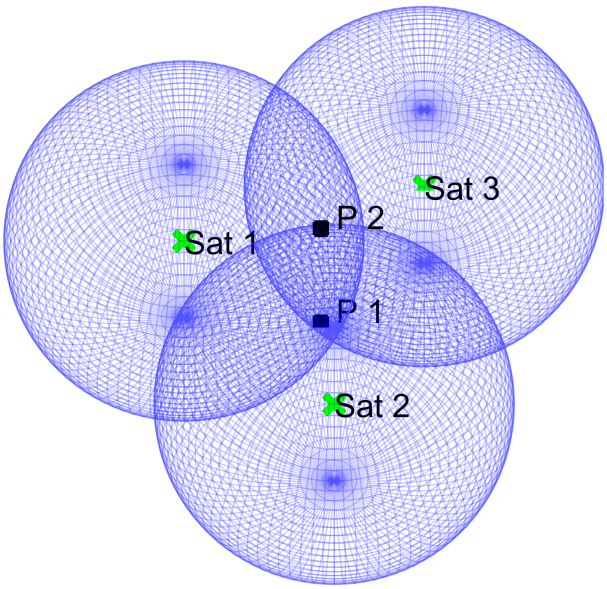


Figure 2.1: Positioning solution obtained from triple satellite ranging observations.

GNSS equipment receives observations from different satellite constellations. Currently, observations from GPS, GLONASS and Galileo are operational. In addition, the integration of satellite observations from other constellations, including the BeiDou Navigation Satellite System, is in progress. By benefiting from two types of observations, i.e. pseudo ranges and carrier phases, GNSS enables positioning from a few meter to a few millimetre accuracy (Hofmann-Wellenhof et al., 2012). However, low cost receivers that are commonly available in commercial devices receive only pseudo ranging observations. The millimetre accuracy is available from post processing techniques such as Differential GPS (DGPS) in which two receivers are used. In DGPS, one receiver, called the base or reference, is stationary with a known position, while the other, known as the rover can be either stationary or moving. The rover position is calculated and corrected on the basis of common observations received from the base. If the connection between the stationary reference and the rover is via telemetry, i.e. radio link, the approach is known as Real Time Kinematic (RTK). The advantage of RTK is that it is almost real-time , therefore it can be used in real-time operations. However, the accuracy achieved by this method is at

decimetre level. Recent advances in RTK have resulted in the achievement of centimetre accuracy for positioning. The most recent approach is PPP-RTK which is an integer ambiguity resolution-enabled Precise Point Positioning (PPP) (Wübbena et al., 2005; Mervart et al., 2008). PPP allows end users to compute their position with centimetre-level accuracy if precise satellite orbit parameters and time are provided (Zumberge et al., 1997; Kouba and Héroux, 2001; Bisnath and Gao, 2009).

Despite the fact that GNSS can provide accurate and almost real-time positioning, it is subject to the provision of clear line of sight between the receiver and a sufficient number of satellites. As illustrated in Figure 2.2, in environments such as indoor or densely built-up areas, known as urban canyons, where GNSS signals are either fully obstructed or affected by buildings, accurate positioning becomes impossible via GNSS alone. Moreover, GNSS is merely able to provide positioning and not attitude estimation. For pose estimation, which includes the estimation of the sensor position and its attitude, the use of other localization sensors is warranted. These sensors are described in the rest of this chapter.

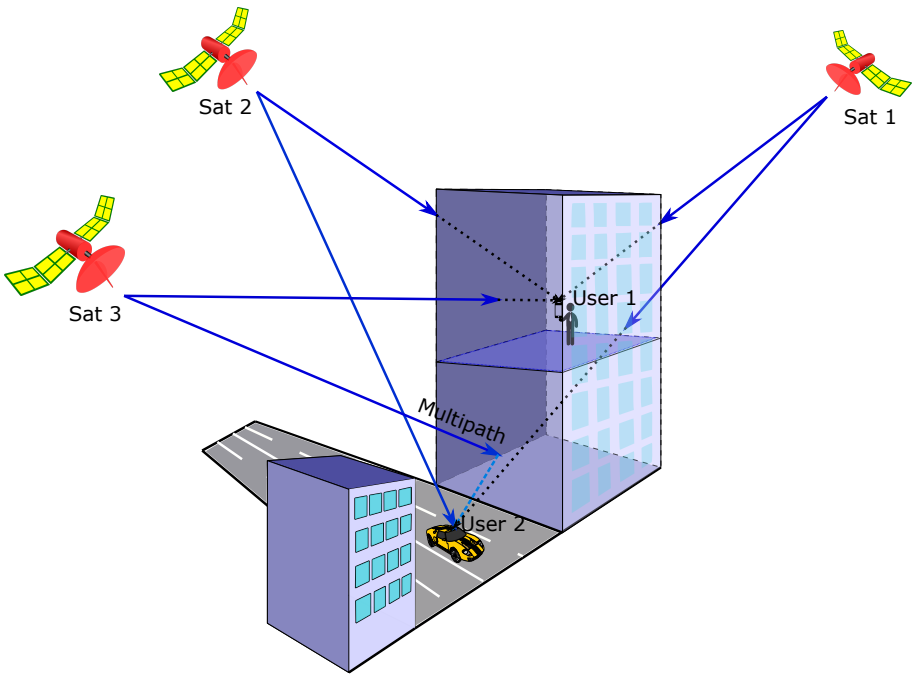


Figure 2.2: Effect of densely built-up environments on GNSS observations.

Exploiting the history of locations provided by GNSS and constraining the locations to navigable paths derived from a map of the environment is a solution to predict a plausible path and infer the present location more precisely and accurately. This technique is known as map-matching in which the estimated locations are matched to a network of navigable paths (e.g., roads for vehicles), typically derived from a map or a Geospatial Information System (GIS). A vast body of literature exists on map-matching to link the GNSS observations with a road network to establish the route taken by a user (Alt et al., 2003; Brakatsoulas et al., 2005; Greenfeld, 2002; Hummel, 2006; Krumm et al., 2007; Li et al., 2007; Quddus et al., 2003). However, achieving a high accuracy by this technique requires a high resolution map and, on the other hand, a higher resolution map leads to the growth of computational cost of localization, thus affecting real-time operability.

2.2.2 Absolute Localization in a Local Frame

Systems that provide absolute localization in a local coordinate system can be grouped into infrastructure-dependent and infrastructure-free localization systems.

Similar to satellite positioning, infrastructure-dependent localization requires external references. However, these references are in the form of an infrastructure of transmitters or beacons with known positions deployed in an environment. By simultaneously receiving information, e.g. range or bearing/angle measurements, from a sufficient number of beacons, it is possible to compute the location of the object of interest in the local coordinate frame of the beacons. The communication between the reference infrastructure and the object of interest is through wireless technology. Today, there exists a range of wireless technologies employed mostly for Indoor Positioning Systems (IPSs). These technologies include ZigBee, Dedicated Short Range Communication (DSRC), Radio Frequency IDentification (RFID), Wireless Local Area Network (WLAN), Ultra Wide Band (UWB) and Bluetooth. The contribution of selected technologies to localization is discussed in the following.

A midrange wireless technology is WLAN, which operates over a range of 50 to 100

metres (Liu et al., 2007). A common sensing measurement used in WLAN is Received Signal Strength (RSS). RSS refers to the intensity of a transmitted signal sensed by a receiver. The intensity of the received signal indicates the distance between two nodes. A number of operational systems are based on this principle. Benefiting from mathematical models such as k nearest neighbourhood in (Saha et al., 2003), maximum likelihood in (Youssef et al., 2003) and (Youssef and Agrawala, 2004), Bayesian networks in (Kontkanen et al., 2004), neural networks in (Saha et al., 2003) or the Monte Carlo approach in (Siddiqi et al., 2003), it has been reported in (Liu et al., 2007) that these systems enable positioning accuracy of roughly 3 to 30 metres to be achieved. In fingerprint-based methods, the localization is carried out within two phases; off-line fingerprint mapping and an online localization phase. In the former phase, the RSS, received from various access points, is collected and the corresponding position is recorded. This information forms a fingerprint map of the area. In the latter phase, the RSS of each node is measured on the path of interest and the measurements are matched with the fingerprint map to calculate signal strength and finally estimate the location. The principal shortcoming of this method, however, is that the offline phase is time consuming and labour-intensive. In addition, accurate matching of the RSS values with the fingerprint map is still a challenge.

UWB (Ingram et al., 2004) is a shorter-range radio technology than WiFi, which is based on transmitting short pulses. In fact, UWB emits signals over a multitude of frequencies in the range of 3.1 to 10.6 Ghz (Bandwidth $>$ 500 MHz). The transmission of multiple bands of frequency is the principal difference between UWB and other wireless technologies. Short pulses allow a precise determination of travel time of the wave from a transmitter to a receiver, the Time of Arrival (ToA), which is the main measurement of UWB. By receiving ToA observations from a sufficient number of UWBs and time synchronising them, a localization accuracy of 20 cm is achievable (Liu et al., 2007). Ubisense is a commercial localization system benefiting from UWB (Ubisense). This system enables not only measurement of ToA, but also measurement of pulse waveform angles, known as Angles of Arrival (AoA), and the Time Difference of Arrival (TDoA).

By taking advantage of the combination of these measurements, this system is reported to achieve decimetre localization accuracy (Gu et al., 2009). If ToA is measured from the transmitter to the corresponding receiver and back, the measurement is called Roundtrip Time of Flight (RToF). This is the measurement that some UWB-based localization systems including Siemens Local Position Radar (LPR) (Vossiek et al., 2003) benefit from. In the literature, it is reported that systems exploiting ToA suffer less from multipath error than those using RSS. In addition, ToA resolves the issue of Line of Sight (LoS) to a considerable extent. Nevertheless, the key issue in the use of UWB technology is its high cost, which is precluding widespread use of this technology.

A common short-range wireless technology is Bluetooth which can exchange data only in the range of 10 to 15 metres (Liu et al., 2007). Since each Bluetooth device has a unique ID, it can be used for locating and tracking objects (Prasithsangaree et al., 2002). A commercial location system based on Bluetooth is Topaz (Topaz) which enables location of tags indoors. The reported accuracy for this system is about 2 metres if only Bluetooth is employed as the wireless technology. Therefore, Topaz exploits the combination of Bluetooth-based localization and IR-based localization, which uses Infrared Radiation (IR) for data exchange, to provide an acceptable room level accuracy (Gu et al., 2009).

Other wireless technologies such as Zigbee (ZigBeeTMAlliance), RFID (Shepard, 2005), DSRC (Yin et al., 2004; Ma et al., 2009) have been used in location systems for communication between transmitting and measuring units/bases. The details of these systems are not discussed because they are outside the scope of this dissertation.

Infrastructure-free localization refers to localization systems that are independent of infrastructure such as WiFi beacons. However, in order to provide absolute localization, an external reference which provides information in a global coordinate system is required. These external references could be either a 2D map of the indoor environment or a 3D model such as a Building Information Model (BIM), or a City Information Model (CIM) in urban areas. Most of the infrastructure-independent localization systems benefit from an optical sensor to match the visual observations with the content of external

references. For instance, Qu (2017) provided a low cost absolute localization technique on the basis of tracking visual landmarks such as traffic signs and road markings in a sequence of images. This approach exploits the observations of visual landmarks and their 3D global coordinates to estimate localization via Local Bundle Adjustment (LBA). However, this approach currently assumes that there is no error in object detection and all the visible landmarks are not tracked simultaneously. Moreover, this approach is not suited for real-time operation due to the use of LBA.

2.3 Relative Localization

Relative localization involves location estimation with respect to the previous pose, given the initial location of the user. Systems that work based on this technique are typically self-contained, i.e. there is no requirement for external references. Among sensors that are used for relative localization, inertial sensors such as accelerometers and gyroscopes, and optical sensors such as cameras and laser scanners, have drawn attention in recent years. In this section, approaches using these sensors are thoroughly reviewed.

2.3.1 Inertial Localization

Sensors operating on the principles of Newton's first and second laws, i.e. bodies tend to be at rest or travel in a straight line at uniform velocity unless a force compels this state to change, and the variation of motion is proportional to the applied force (Newton et al., 1850), are known as inertial sensors. Inertial sensors comprise accelerometers and gyroscopes, which measure specific force and angular rates, respectively (Groves, 2013). Typically, a combination of these sensors forms an Inertial Measurement Unit (IMU), which is used for 3D localization. An IMU encompasses six sensors; three accelerometers in three mutually perpendicular dimensions sensing accelerations in the x, y and z axes, and three gyros measuring rotations in accordance with a specific axes frame. An IMU that is equipped with an on-board processor to estimate the localization is called an

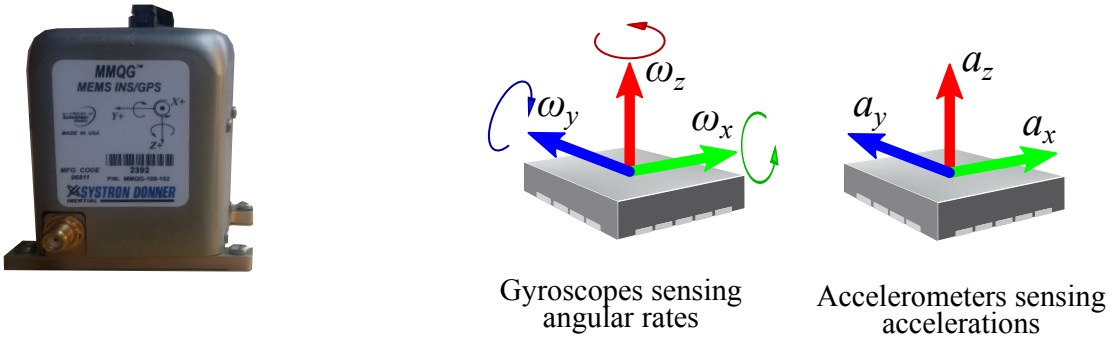


Figure 2.3: An INS with its inertial components and their orientation.

Inertial Navigation System (INS) (Vectornav). Figure 2.3 shows an INS and its inertial components.

A common approach to estimate localization using inertial sensors is dead reckoning. In the dead reckoning approach, the location is estimated based on the outputs of accelerometers and gyroscopes. As noted above, a key advantage of inertial localization is that it is purely self-contained and therefore does not require any external infrastructure such as satellites or beacons. In addition, inertial sensors can provide continuous localization. However, due to the incremental nature of estimation in the dead reckoning approach, the location estimates drift over time (Groves, 2013).

In the pedestrian localization context, localization based on dead reckoning is known as Pedestrian Dead Reckoning (PDR). Research, such as (Abdulrahim et al., 2010) has exploited a Zero Velocity Update (ZUPT) methodology, in which pedestrian steps are detected and the velocity is simply adjusted to zero during each stance phase when the step of the user is on the ground, to add constraints on the INS mechanisation algorithm in order to decrease drift. Nonetheless, this method is applicable only to foot mounted systems, which are able to detect stance phase. State-of-the-art PDR approaches integrate inertial measurements with step length estimation and use motion state recognition and landmarks to constrain the drift of location estimates (Gu et al., 2016a,b). Step length estimation is used when ZUPT is not applicable, i.e. when the IMU is not strapped to a body part e.g. in smartphone based systems.

2.3.2 Optical Sensor-Based Localization

Optical sensors, used broadly in ego-motion estimation, can be grouped into two classes: passive and active systems.

Passive Systems

Passive optical sensors such as cameras receive emissions from objects existing in their Field of View (FoV) and create a single image. Depending on the FoV, cameras are divided into perspective and omnidirectional cameras. Cameras that have a FoV of 360° in azimuth and 90° to 140° in elevation (Strelow and Singh, 2004) are known as omnidirectional cameras. These include dioptric and catadioptric cameras. In a dioptric camera, light beams pass through a combination of lenses, whereas in a catadioptric system, a combination of lenses and curved mirrors is used to project the scene onto the image plane. Figure 2.4 illustrates a catadioptric camera and a dioptric camera, along with their projection process, as compared to a perspective camera.

In the last few years, omnidirectional cameras have gained popularity in ego-motion estimation (Scaramuzza, 2008a; Zhang et al., 2016; Caruso et al., 2015) and a substantial body of research has been devoted to modelling their imaging geometry (Barreto and Araujo, 2001; Geyer and Daniilidis, 2000) and calibration (Kannala and Brandt, 2006; Mei and Rives, 2007; Scaramuzza et al., 2006a). For instance, Geyer and Daniilidis (2000) investigated various types of central catadioptric cameras and provided generic calibration models suitable for all these types of cameras. Central catadioptric cameras refer to cameras whose projection viewpoint is unique. Scaramuzza et al. (2006a) introduced a versatile technique for central omnidirectional camera calibration. This approach represents calibration parameters by a polynomial function whose coefficients are estimated through a two-phase least-squares linear optimization.

Localization by a sequence of images falls into the category of either Visual Odometry (VO) or Visual-Simultaneous Localization and Mapping (V-SLAM). VO, first proposed

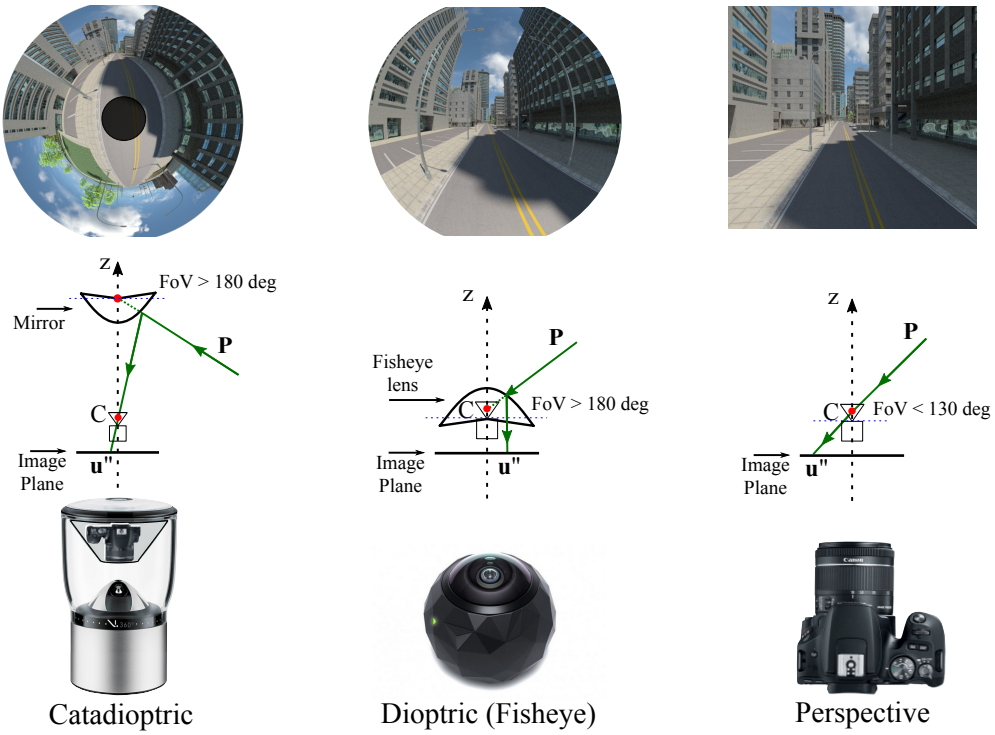


Figure 2.4: Omnidirectional cameras and their image projection compared to a perspective camera.

by Nistér et al. (2004), is a common approach to ego-motion estimation utilizing single or multiple cameras. In a typical visual odometry approach, salient features are detected and matched between pairs of sequential images. Relative poses between the images in a sequence are estimated by employing matched feature points, and these are then refined by using 3D points reconstructed from the image observations. VO is a special case of a problem in computer vision known as Structure from Motion (SfM) (Hartley and Zisserman, 2003; Ma et al., 2012; Snavely et al., 2006, 2008). Regardless of the order of images, SfM estimates camera poses and 3D structure of the visible scene from a multitude of camera images typically in an off-line optimization. In SfM, the ego-motion estimation can be conducted either in 2D space or 3D space. Based on the specification of feature correspondences in two or three dimensions, VO approaches are classified into 2D to 2D, 3D to 3D and 3D to 2D groups (Scaramuzza and Friedrich, 2011). As illustrated in Figure 2.5, in a 2D-to-2D VO approach, features from one frame are used with their correspondences

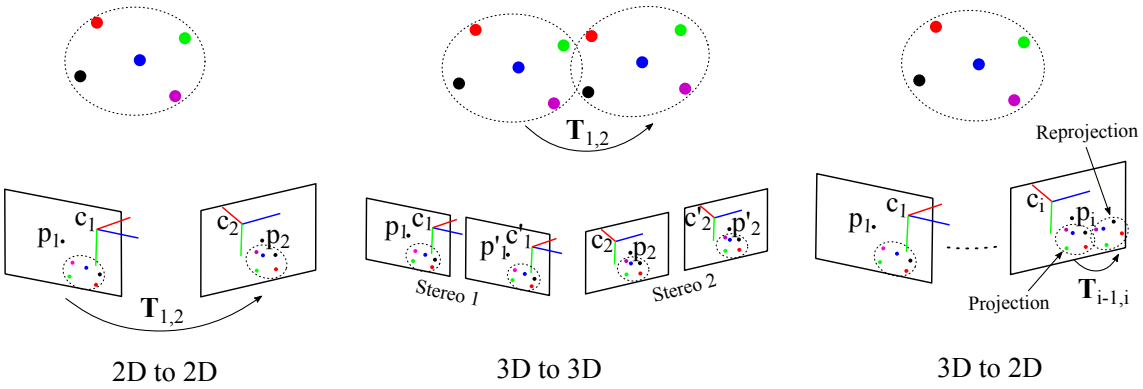


Figure 2.5: An illustration of different methods of VO.

on the next frame to estimate the essential matrix. From the decomposition of the essential matrix, it is possible to achieve the transfer matrix which includes the translation and rotation components. Research by (Nistér, 2004) has been pioneering in this technique. Contrary to the 2D-to-2D approach, the 3D-to-3D motion systems calculate the transfer matrix in the 3D space where the scene is reconstructed. This technique is mostly suitable for systems that use more than one camera such as stereo cameras, because the scene must be reconstructed at each camera pose to be able to calculate the transfer matrix from consecutive reconstructions. However, Moravec (1980), who is a pioneer of this technique, designed a mono camera which slides on a rail, and in a stop-and-go fashion it captures multiple images at equidistant intervals to reconstruct the scene. Finally, by comparing the reconstructed scenes from two consecutive poses, the translation and rotation parameters are estimated. A similar approach, but using a stereo pair of cameras has been used by Lacroix et al. (1999), Olson et al. (2000) and Hannah (1974) to provide localization. The last approach, 3D to 2D, is the most common approach used for ego-motion estimation. In this approach, the transformation is computed from the reprojection of a 3D reconstructed scene and the corresponding observations on a same image. Nistér et al. (2004) noted that motion estimation by this technique is more accurate than by 3D-to-3D approaches due to the minimization of residual errors on the image plane instead of the feature position errors in the 3D space. Works such as (Moreno-Noguer et al., 2007; Kneip et al., 2011; Nistér and Stewénius, 2007) have employed this technique

for ego-motion estimation.

If motion parameters and scene structure are jointly estimated, the method is known as Simultaneous Localization and Mapping (SLAM). SLAM has become a loaded term in recent years, with much of the debate revolving around the idea of global map consistency (Mur-Artal et al., 2015; Endres et al., 2012). Indeed, this is a key difference between SLAM and VO, which concerns only local consistency of the trajectory. If cameras are only the sensors used for localization and mapping, the method is defined as Visual-SLAM (V-SLAM) (Scaramuzza and Friedrich, 2011). V-SLAM approaches operate on the basis of either a filtering method (Aulinas et al., 2008), fusing visual information from a window of camera poses with a probability distribution, or a non-linear optimization, utilizing a global optimization (Rosen et al., 2015). In recent years, V-SLAM approaches, employing either a mono or a stereo camera, have achieved successful performance in small-scale (Newcombe et al., 2015) and large-scale environments using submap joining (Ni and Dellaert, 2010; Zhao et al., 2013). In order to reduce the computational effort, some works such as (Pirchheim et al., 2013; Williams et al., 2007) benefit from the keyframe optimization algorithm. Although V-SLAM outperforms VO in terms of drift reduction, error accumulation is inevitable, affecting the ego-motion estimation after long operation. Algorithms such as ORB-SLAM (Mur-Artal et al., 2015) and RGB-D SLAM (Endres et al., 2012) detect loop closures and apply pose graph optimization (Strasdat et al., 2010; Kümmerle et al., 2011) to improve the global consistency of the map and correct trajectory drift to a certain degree. However, loop closing and trajectory correction impose a delay in localization affecting the real-time operability of the system.

What all vision-based localization systems have in common is that they are highly sensitive to the quality of visual features. In environments with the presence of dynamic features, e.g. pedestrians and cars, or insufficient salient features, VO and V-SLAM approaches can fail, resulting in discontinuity and inaccuracy in motion estimation. The aforementioned problems are more common when a perspective camera, which normally has a limited FoV, is utilized in the design of vision-based navigation systems. By con-

trast, omnidirectional cameras, which provide super wide FoV, are less likely to end up in situations where there are no salient features in the view, or a large portion of the images is covered by dynamic entities. Moreover, using omnidirectional cameras, it is more likely that features will be tracked over a longer sequence of images, resulting in more accurate localization (Zhang et al., 2016). In recent years, omnidirectional cameras have often been employed in the design of VO and V-SLAM systems to estimate ego-motion (Caruso et al., 2015; Scaramuzza, 2008a). Scaramuzza (2008a) is amongst the pioneers who used omnidirectional cameras for ego-motion estimation. He detailed a 2D VO using an omnidirectional camera to track robust scale invariant features, mostly from the ground plane, and estimate high resolution estimates of the azimuth utilizing an appearance-based approach. By combining the feature tracker with a visual compass, more accurate pose estimation of a vehicle was achieved compared to use of the standard feature-based approach. Tardif et al. (2008) presented an omnidirectional VO based on a Ladybug 2 camera, which includes 5 perspective cameras, to estimate motion. Similar to the work of Scaramuzza (2008a), the system developed by Tardif et al. (2008) was restricted to planar motion. In a recent effort, Caruso et al. (2015) introduced a Large-Scale Direct SLAM (LSD-SLAM) (Engel et al., 2014) for omnidirectional cameras and tested it in several large-scale environments. Figure 2.6 shows the result of one of their evaluations in a large-scale indoor environment. As reported, the omnidirectional LSD-SLAM outperformed the traditional approach using a perspective camera. This implies that additional information provided by omnidirectional cameras will contribute to better localization accuracy due to their super wide FoV. A downside of LSD-SLAM, however, is that because photometric error is used instead of reprojection error, which is common in indirect SLAMs, it is highly affected by a change in illumination. In addition, in order to decrease the computational cost, many assumptions have to be considered. One of the assumptions is that the camera rotation between two consecutive frames will be small for propagation of inverse depth variance (colour-coded version of inverse depth maps are shown in Figure 2.6), affecting the operability of the system on dynamic plat-

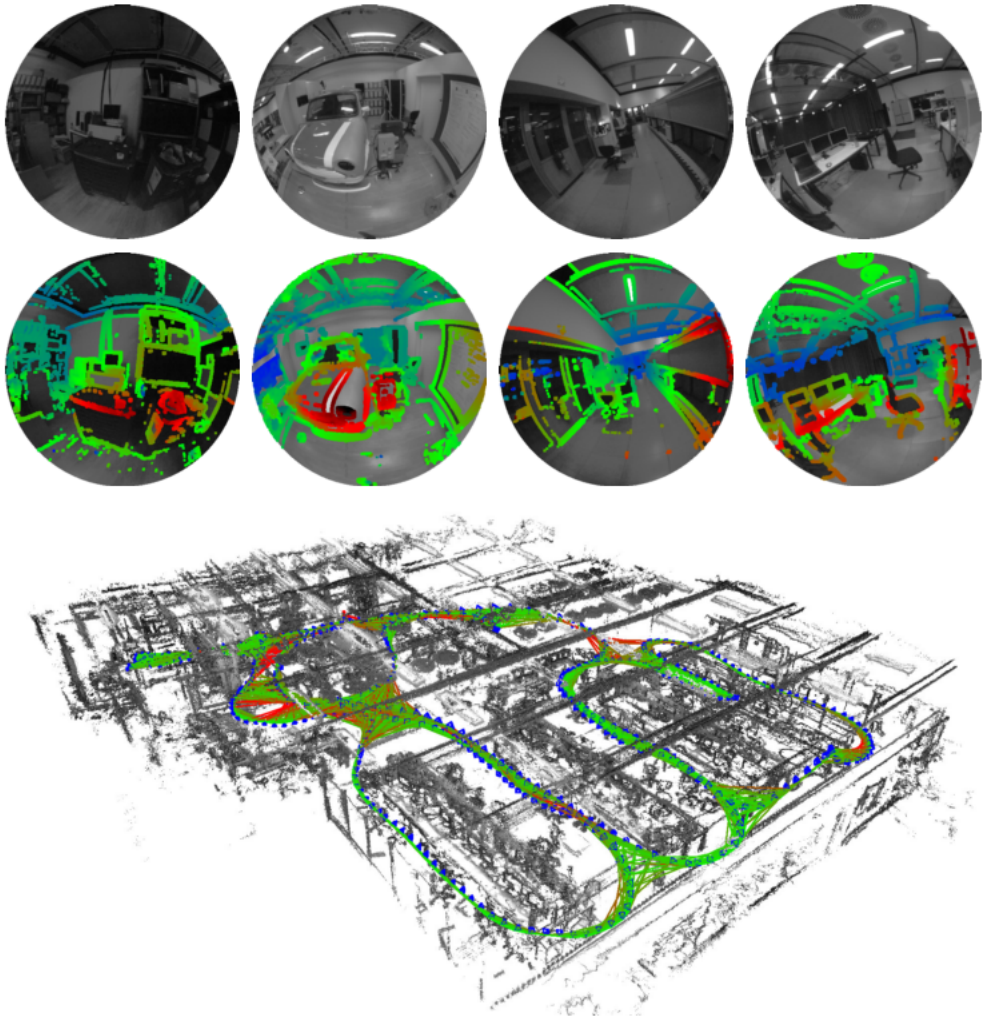


Figure 2.6: Visual result for LSD-SLAM using an omnidirectional camera (Caruso et al., 2015).

forms. Moreover, this technique requires several seconds of initialization which makes the system unsuitable for quick response localization-aware applications.

Active Systems

Active sensors such as laser scanners measure the travel time of light transmitted by these sensors and reflected from objects existing in the range of the sensor's operation. Recently, Lidar sensors have been increasingly employed in SLAM systems to generate the map and simultaneously estimate the ego-motion (Nüchter et al., 2007; Taguchi et al.,

2013). In Lidar-based SLAM, motion estimation is performed by registering successive laser-scan epochs (Khoshelham, 2015, 2016), which is generally more accurate than localization using image observations. In ITS, including driverless cars, Lidar sensors are used for detecting road obstacles and navigation-related objects, but also for position estimation in the absence of GNSS signals (Levinson et al., 2007; Kaijaluoto et al., 2015). They are also broadly exploited in mobile mapping systems to map the environments in detail for GIS purposes (Kim et al., 2017; Javanmardi et al., 2017; Manandhar and Shibasaki, 2001). A principal disadvantage of Lidar sensors, which remains a challenge, is the expense of the technology, which can limit their adoption.

There are also sensors, such as the Kinect, that exploit both technologies. Basically, a Kinect consists of an RGB camera and a depth sensor which contains a structured-light projector operating in the IR spectrum and a monochrome camera. By combining observations from the monochrome camera and the IR projector, estimation of depth is achieved and the scene can be reconstructed in 3D. This system, which is known as an RGB-D sensor, has been broadly employed in robot localization (Ganganath and Leung, 2012; Biswas and Veloso, 2012; Henry et al., 2012; Sturm et al., 2012). Khoshelham and Elberink (2012) analysed the accuracy and resolution of Kinect V1 depth data for indoor mapping applications. According to their report, the random error of depth and its resolution are quadratically in direct and inverse proportion to distance from the sensor, respectively. In other words, an increase in distance from the sensor results in a quadratic ascent in the random error of depth measurements, along with a quadratic descent in resolution. Similarly, Yang et al. (2015) investigated the accuracy error distribution of Kinect for Windows V2 as well as its resolution. Comparing the results of evaluation of two versions, a substantial improvement was seen in the Kinect V2 both in terms of accuracy and resolution within a distance between the sensor and target object of a few metres. For the Kinect V1, data acquisition within 1 to 3 meters is recommended, whereas a range of 0.5 to 3.5 meter is advised for Kinect V2. Due to this short operating range, both systems are not suitable for large indoor areas and outdoor environments.

2.4 Multi-sensor Localization

Multi-sensor localization refers to methodologies in which more than one sensor, either on-board or off-board, are employed for localization purposes. In this dissertation, localization by only the use of on-board sensors, i.e. sensors which are attached to the agent of interest, is called localization by sensor fusion. If the sensors are off-board, i.e. deployed surrounding the agent and not necessarily attached to it, the localization is termed as co-operative localization (CL). This section discusses both various sensor fusion approaches used for localization, and CL techniques.

2.4.1 Sensor-Fused Localization

The most common navigation system integrated with other systems is INS due to its complementary characteristics. Because INS is self-contained and provides inertial measurements continuously at high rate, it is a perfect match for GNSS. As summarized in Table 2.1, while INS provides short-term accurate localization including translation, velocity and attitude, GNSS provides long-term localization accuracy, but the localization components are limited to global location and velocity. By integrating GNSS with INS, a global navigation system which operates seamlessly at high rates is possible. Nevertheless, this integration would likely fail in GNSS-deprived environments due to the rapid error accumulation in INS (Groves, 2013).

The integration of an INS with optical sensors is another type of sensor fusion which has drawn attention in location-aware applications (Mourikis and Roumeliotis, 2007; Kim

Table 2.1: Summary of INS and GNSS complementary characteristics according to (Groves, 2013).

Characteristic / System	Accuracy Endurance	Operating Rate	Products
GNSS	Long-term	Low (10Hz)	Position, Velocity
INS	Short-term	High (>50Hz)	Position, Velocity, Attitude

and Sukkarieh, 2007; Li and Mourikis, 2013; Forster et al., 2015). Similar to INS-GNSS fusion, the integration of INS with a camera can overcome their individual shortcomings and create a more robust localization system. While INS provides a good initialization for camera motion, image observations acquired from a camera decrease the INS drift to a considerable degree. The idea of a localization system based on the fusion of vision and inertial measurements is known as Visual-Inertial Odometry (VIO). Since VIO plays a fundamental role in this thesis, separate subsections are allocated to describing this technique from a hardware perspective and an estimation perspective for existing VIO approaches.

Other sensors, even though expensive, such as laser scanners, have been fused with INS or cameras for localization and mapping purposes (Kaijaluoto et al., 2015; Talaya et al., 2004; Grejner-Brzezinska et al., 2011). For instance, Kaijaluoto et al. (2015) introduced a precise indoor localization system using a 3D mobile laser scanner. According to their report, this integration can achieve a localization accuracy of a few centimetres. As an example of camera and laser scanner integration, Lin et al. (2012) fused a 2D laser scanner with a stereo camera for mapping and localization in 3D environments. An advantage of using a 2D laser scanner is its lower price compared to a 3D laser scanner. In addition, since 2D laser scanners can provide highly accurate range and bearing measurements, they have been utilized more broadly in SLAM. The integration of cameras with 2D laser scanners can provide a robust localization system for accomplishing localization tasks in 3D environments.

2.4.2 Cooperative Localization

As shown in Figure 1.1, in the adopted classification, cooperative localization means the use of sensors on-board and off-board for sharing information to estimate motion of an agent. Measurements taken by off-board sensors are exchanged via wireless technologies. A common CL approach is Vehicular Ad hoc NETworks (VANETs) (Saha and Johnson, 2004), which are used for vehicle localization on the basis of vehicle-vehicle (V-V)

and vehicle-infrastructure (V-I) communications. In a simple scenario, as illustrated in Figure 2.7, the location of the car which is in a GNSS deprived environment can be acquired with the aid of additional ranging measurements through V-V and V-I communications in a network in which vehicles and infrastructures form the vertices. Despite the fact that CL researchers such as Benslimane (2005), Drawil and Basir (2010), Parker and Valaee (2007), Patwari et al. (2005), Patwari et al. (2003) and Shao et al. (2010) reported an achievement of sub-meter localization accuracy, there exist constraints and problems of radio ranging/range rating and vehicular communication. Alam and Dempster (2013) concluded that CL is still premature, therefore it is not capable to bridge the gap and address the localization accuracy required for location-aware services such as Intelligent Transport System (ITS).

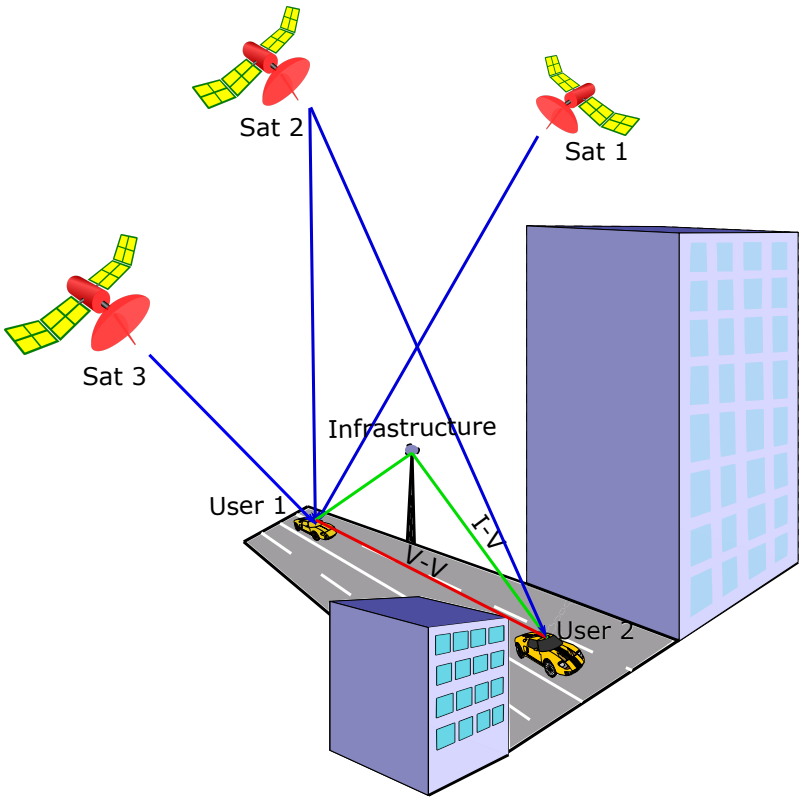


Figure 2.7: A CL scenario in VANET.

2.5 Visual-Inertial Odometry (VIO)

As mentioned earlier, VO and visual-SLAM are likely to fail in environments with insufficient texture and geometric features, resulting in discontinuities in localization. To provide seamless motion estimation, visual observations can be fused with the outputs of accelerometers and gyroscopes. This approach is known as Visual-Inertial Odometry (VIO). A significant body of literature exists on VIO (Mourikis and Roumeliotis, 2007; Kim and Sukkarieh, 2007; Li and Mourikis, 2013; Forster et al., 2015). This section discusses various VIO systems from both hardware and integration algorithm perspectives.

2.5.1 VIO: Survey from a Hardware Perspective

Many VIO systems comprise an INS and a single perspective camera forming their hardware. Research by Mourikis (2006), Leutenegger et al. (2015) and Li and Mourikis (2013) has employed the standard projection model to integrate image observations captured from the single camera with inertial measurements. In fact these systems benefit from epipolar constraints produced by consecutive images which are in the form of temporal stereo-pairs to decrease the drift of the INS over time. To improve the motion estimation, VIO systems using stereo cameras have been developed (Tardif et al., 2010; Oskiper et al., 2007). By exploiting a stereo camera as a rigid body, depth information can be obtained from both static and temporal stereos. Static stereo refers to the fixed-baseline images of the stereo camera, whereas temporal stereo relates to images from the same camera captured at different camera poses in time.

In addition, by combining static with temporal stereos, the system exploits the reconstruction of very small and very large scaled environments. This way, ego-motion estimation can be significantly improved especially in indoor environments where it is more likely that objects and features are very close to the agent. There are also several VIO systems that have been developed on the basis of an integration of inertial measurements with image observations captured by a rolling-shutter camera (Li and Mourikis,

2014; Li et al., 2013).

In contrast with a global-shutter camera, in which all pixels of an image are captured at once, in a rolling-shutter camera the image is captured row by row, usually with a constant delay. Most works using a rolling-shutter camera and inertial measurements have focused on the image rectification issue. Using inertial sensors, it is possible to remove the distortion caused by the rolling shutter (Hanning et al., 2011; Karpenko et al., 2011; Jia and Evans, 2012). In order to estimate ego-motion employing a rolling-shutter camera, Klein and Murray (2009), employed the estimates of camera poses to remove the effect of the rolling-shutter in the projection of the visual features and eventually considered the image measurements as if they were captured from a global-shutter camera. This method, however, leads to loss of accuracy in very dynamic camera motion with significant accelerations (Li and Mourikis, 2014).To solve this issue, Li and Mourikis (2014) considered the linear and rotational velocities of a perspective rolling-shutter camera at the time of image capture as parameters to be estimated in order to model the rolling-shutter effect in the measurement model.

Perspective cameras have limited Field of View (FoV). Due to this fact, VIO systems based on a perspective camera, cannot exploit the maximum potential of visual information surrounding the localization system. Moreover, it is more likely that images captured by a perspective camera will lack sufficient salient features in the presence of moving entities occupying large parts of the image format. An alternative that can reduce this issue is the use of an omnidirectional camera. Despite the advantages of omnidirectional cameras for ego-motion estimation, there is currently only a small body of research on the use of these cameras in VIO. In (Li et al., 2014) an effort was made in order to fuse omnidirectional vision with inertial sensors using an adaptive estimator. This method estimates the position and linear velocity of a mobile robot, but not other pose parameters such as orientation. Furthermore, the performance of the algorithm has not been tested in a large scale environment.

2.5.2 VIO: Survey from an Estimation Perspective

From an estimation perspective, existing VIO algorithms can be classified into non-linear optimization and recursive methods. In non-linear optimization methods, the state vector, including parameters to be estimated, is obtained by iterative minimization of a least-squares error function. In (Strelow and Singh, 2004), the sensor pose, scene structure and other unknowns are estimated using all image observations and inertial measurements, simultaneously. However, the motion estimation in this method is subject to the availability of all the observations from images, thus limiting the real-time operation of the system. To tackle the real-time operability issue, Leutenegger et al. (2015) bounded the size of the optimization window by discarding the older states. This method is defined in (Sibley et al., 2010) as a Sliding Window Filter (SWF). The drawback of the approach, however, is that owing to the discarding of multiple camera poses, some observations of a visible feature will not contribute to the motion estimation, resulting in a loss of constraints.

In contrast to non-linear optimization, recursive methods estimate sensor motion and parameters recursively in a filter as inertial measurements and image observations become available. Therefore, this methodology is appropriate for real-time applications even in large-scale environments with very long sequences (Strelow and Singh, 2004). A common filtering estimator is the Extended Kalman Filter (EKF) (Mourikis and Roumeliotis, 2007; Huster et al., 2002; You and Neumann, 2001; Kottas and Roumeliotis, 2013), in which the state vector will be updated by visual observations extracted from the images once it and its uncertainty have been propagated in a linear system model. Provided that the scene structure is simultaneously estimated with the current INS pose, the approach is termed EKF-based Simultaneous Localization and Mapping (SLAM) (Langelaan, 2007; Strelow and Singh, 2004; Ong et al., 2003; Eade and Drummond, 2006). A benefit of EKF-SLAM algorithms is that the correlations between sensor motion and 3D features are taken into account since the camera pose and 3D coordinates of visual features are estimated together. A rigorous handling of these correlations, however, imposes computa-

tional complexity that can compromise this approach in areas rich in features (Mourikis, 2006).

Several filtering algorithms have been proposed in which, contrary to SLAM, the 3D coordinates of features are not jointly estimated with the state vector (Mourikis, 2006; Roumeliotis et al., 2002; Diel et al., 2005; Bayard and Brugarolas, 2005; Soatto et al., 1996; Soatto and Perona, 1997; Prazenica et al., 2005). These methods exploit visual observations to impose epipolar constraints between either a pair of images (Roumeliotis et al., 2002; Diel et al., 2005; Bayard and Brugarolas, 2005) or a multitude of camera poses (Mourikis, 2006; Nistér et al., 2006). Amongst these methods, Multi-State Constraint Kalman Filtering (MSCKF) (Mourikis, 2006) benefits from all geometric constraints that a scene point produces when observed from multiple camera poses. This method will be thoroughly introduced in Chapter 3. By employing all these geometric constraints, more accurate motion estimation can be achieved.

2.6 Summary

In this chapter, different approaches of state-of-the-art localization systems have been discussed. Based on localization with respect to a coordinate system, the localization systems have been classified as absolute or relative localization. GNSS with the integration of sensors such as INS has been discussed as a means to provide an accurate solution for localization in cases where there are sufficient GNSS observations. Other solutions have been proposed to solve the localization problem in GNSS-denied environments such as urban canyons and indoor areas. As noted, although numerous techniques employing wireless technologies exist, each has its own shortcomings and providing a robust real-time localization system remains a challenge. Using sensors such as laser scanners can produce an accurate and robust localization. Nevertheless the problem of these sensors is their high cost which currently precludes wider adoption. Recently, localization via the integration of visual and inertial sensors within VIO frameworks has been studied,

however, little attention has been paid to improving the robustness of the localization solution via stereo and omnidirectional VIO.

Chapter 3

Visual-Inertial Odometry Using a Single Camera

3.1 Introduction

THIS chapter describes the Multi-State Constraint Kalman Filter (MSCKF) (Mourikis and Roumeliotis, 2007) and its full formulation. Because the MSCKF is the basis of the stereo-VIO and OVIO approaches proposed in this thesis, allocation of one chapter to this method is beneficial for better clarification of the following chapters.This estimator is a recursive estimator in which an Extended Kalman Filter (EKF) plays a key role in propagating INS parameters in time and augmenting them with camera poses at each image acquisition. It also updates them with image observations through a measurement model. A principal benefit of this method compared to others is its ability to use the maximum potential of constraints that a visible feature creates, while the complexity remains linear in the number of features. A simulation is designed in the following of this chapter to show the performance of the MSCKF compared to INS-only integration.

3.2 Conventional MSCKF

Basically, the MSCKF comprises three steps: 1- system model in which a state vector including system position, attitude, velocity and INS biases is propagated in accordance with the INS mechanism and it is augmented with camera poses at each time that an

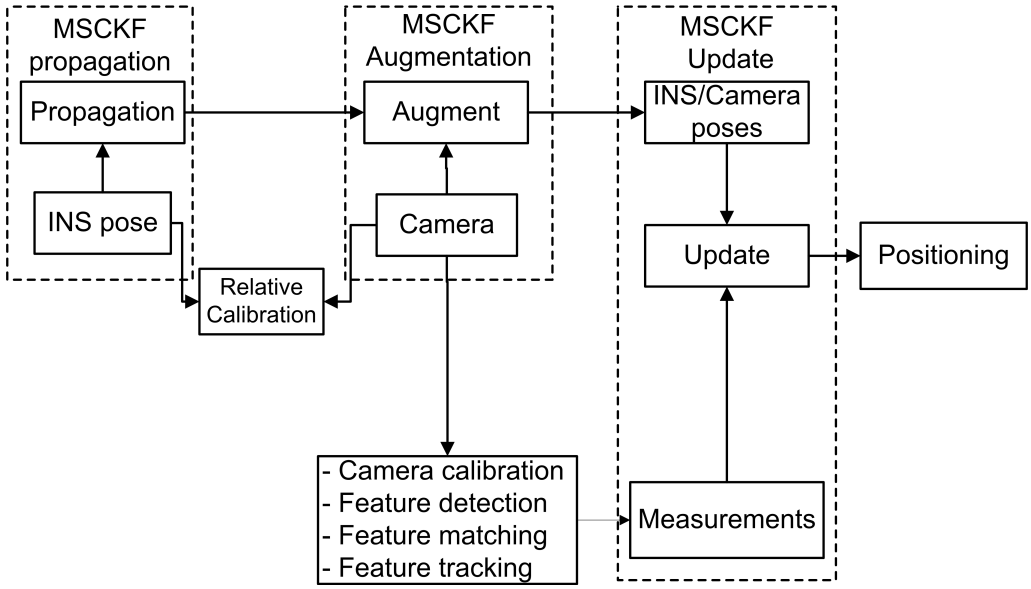


Figure 3.1: The genral workflow of the mono MSCKF.

image is taken; 2- measurement model which forms a mathematical model for image observation according to the type of camera; and 3- update where visual measurements influence the propagated state to minimize the error vector. The general workflow of the mono MSCKF is illustrated in Figure 3.1.

3.2.1 Formulation of MSCKF State Vector

The state vector of INS, $\mathbf{X}_{INS} \in \mathbb{R}^{16}$, is described as:

$$\mathbf{X}_{INS} = \begin{bmatrix} \mathbf{q}_{\mathcal{E}}^{\mathcal{B}\mathrm{T}} & {}^{\mathcal{E}}\mathbf{v}_{\mathcal{B}}^{\mathrm{T}} & {}^{\mathcal{E}}\mathbf{p}_{\mathcal{B}}^{\mathrm{T}} & \mathbf{b}_a^{\mathrm{T}} & \mathbf{b}_g^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{3.1}$$

where $\mathbf{q}_{\mathcal{E}}^{\mathcal{B}}$ is the 4×1 unit quaternion representing the rotation of the INS resolved in the body frame $\{\mathcal{B}\}$, vectors ${}^{\mathcal{E}}\mathbf{v}_{\mathcal{B}}$, and ${}^{\mathcal{E}}\mathbf{p}_{\mathcal{B}}$ are the linear velocity and position of the INS with respect to the Earth Centered Earth Fixed (ECEF) frame $\{\mathcal{E}\}$, respectively. The 3×1 vectors $\mathbf{b}_g$ and $\mathbf{b}_a$ express the gyroscope and accelerometer biases, respectively. Follow-

ing equation (3.1), the INS error-state, $\delta\mathbf{X}_{INS} \in \mathbb{R}^{15}$ is given by:

$$\delta\mathbf{X}_{INS} = [\ \delta\mathbf{\Phi}_{\mathcal{B}}^{\mathrm{T}} \quad {}^{\mathcal{E}}\delta\mathbf{v}_{\mathcal{B}}^{\mathrm{T}} \quad {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{B}}^{\mathrm{T}} \quad \delta\mathbf{b}_{a}^{\mathrm{T}} \quad \delta\mathbf{b}_{g}^{\mathrm{T}}\ ]^{\mathrm{T}} \tag{3.2}$$

where the 3×1 vector $\delta\mathbf{\Phi}_{\mathcal{B}}$ defines the perturbation of the INS attitude in the body frame. This vector can be linearly approximated by the quaternion error evolution (Sola, 2012):

$$\delta\mathbf{q} \approx [\frac{1}{2}\delta\mathbf{\Phi}^{\mathrm{T}}\ 1]^{\mathrm{T}} \tag{3.3}$$

The full EKF state vector includes the state vector of INS, and M camera poses from which visual features are visible. This $(16+7M)\times1$ EKF state vector is defined as:

$$\mathbf{X} = [\mathbf{X}_{INS}^{\mathrm{T}}\ \mathbf{q}_{\mathcal{E}}^{\mathcal{C}_1\mathrm{T}}\ {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_1}^{\mathrm{T}}\ \cdots\ \mathbf{q}_{\mathcal{E}}^{\mathcal{C}_i\mathrm{T}}\ {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}^{\mathrm{T}}\ \cdots\ \mathbf{q}_{\mathcal{E}}^{\mathcal{C}_M\mathrm{T}}\ {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_M}^{\mathrm{T}}]^{\mathrm{T}} \tag{3.4}$$

In this expression, the rest of the parameters are M camera poses encompassing the camera quaternions, $\mathbf{q}_{\mathcal{E}}^{\mathcal{C}_i}$, expressed in the camera frame $\{\mathcal{C}_i\}$, and translations, ${}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}$, with respect to the ECEF frame $\{\mathcal{E}\}$, where $i = 1\cdots M$. Similar to equation (3.2), the $(15+6M)\times1$ error-state vector is defined as:

$$\delta\mathbf{X} = [\delta\mathbf{X}_{INS}\ \delta\mathbf{\Phi}_{\mathcal{C}_1}^{\mathrm{T}}\ {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_1}^{\mathrm{T}}\ \cdots\ \delta\mathbf{\Phi}_{\mathcal{C}_M}^{\mathrm{T}}\ {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_M}^{\mathrm{T}}]^{\mathrm{T}} \tag{3.5}$$

3.2.2 INS state propagation

Errors involved in the INS error-state propagation include attitude error, velocity error, position error, and biases.

1. Attitude error propagation (Sola, 2012):

   We describe the attitude-error propagation resolved in the body frame.

$$\delta\dot{\mathbf{\Phi}}_{\mathcal{B}} \approx -\lfloor(\hat{\boldsymbol{\omega}})\wedge\rfloor\delta\mathbf{\Phi}_{\mathcal{B}} - \delta\mathbf{b}_{g} \tag{3.6}$$

where $\delta\dot{\mathbf{\Phi}}_{\mathcal{B}}$ is the time derivative of the attitude error and the vector $\hat{\boldsymbol{\omega}}$ is the estimated angular velocity resolved in the body frame. $\hat{\boldsymbol{\omega}}$ is obtained from the sensed angular velocity, $\boldsymbol{\omega}_m$, and the estimated gyroscope bias, $\hat{\mathbf{b}}_g$:

$$\hat{\boldsymbol{\omega}} \approx \boldsymbol{\omega}_m - \hat{\mathbf{b}}_g \tag{3.7}$$

the operator $\lfloor(\mathbf{x})\wedge\rfloor$ denotes the skew symmetric matrix of the vector $\mathbf{x}$ expressed as:

$$\lfloor(\mathbf{x})\wedge\rfloor = \begin{bmatrix} 0 & -x_z & x_y \\ x_z & 0 & -x_x \\ -x_y & x_x & 0 \end{bmatrix} \tag{3.8}$$

2. Velocity error propagation (Groves, 2013):

The velocity error, where it is referenced to and resolved in the ECEF frame, is described as:

$$^{\mathcal{E}}\delta\dot{\mathbf{v}}_{\mathcal{B}}^{\mathrm{T}} \approx \lfloor -(\hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\,\hat{\mathbf{f}}\,)\wedge\rfloor \hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\,\delta\mathbf{\Phi}_{\mathcal{B}}^{\mathrm{T}} - 2\mathbf{\Omega}_{\mathcal{E}}\,{}^{\mathcal{E}}\delta\mathbf{v}_{\mathcal{B}}^{\mathrm{T}} + \frac{2g_0(\hat{L}_b)}{^{\mathcal{E}}r_S(\hat{L}_b)}\frac{^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}}{|\,^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}|^2}\,{}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}^{\mathrm{T}}\,{}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{B}} - \hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\mathbf{b}_a^{\mathrm{T}} \tag{3.9}$$

where $\mathbf{C}_{\mathcal{B}}^{\mathcal{E}}$ is the rotation matrix from frame $\{\mathcal{B}\}$ to frame $\{\mathcal{E}\}$, $\hat{\mathbf{f}}$ is the estimated specific force in the body frame related to the measured specific force, $\mathbf{f}_m$, and the estimated accelerometer bias, $\hat{\mathbf{b}}_a$, and it is given by:

$$\hat{\mathbf{f}} = \mathbf{f}_m - \hat{\mathbf{b}}_a \tag{3.10}$$

The matrix $\mathbf{\Omega}_{\mathcal{E}}$ is the skew-symmetric of the Earth-rotation vector and $|\mathbf{x}|$ is the Euclidean norm of the vector $\mathbf{x}$. The scalar parameter, $g_0(\hat{L}_b)$ is the surface gravity calculated from the Somigliana model (Hofmann-Wellenhof and Moritz, 2006),

and ${}^{\mathcal{E}}r_S(\hat{L}_b)$ is the geocentric radius at the surface given by (Groves, 2013):

$$ {}^{\mathcal{E}}r_S(\hat{L}_b) = R_E(\hat{L}_b)\sqrt{cos^2\hat{L}_b + (1-e^2)^2 sin^2\hat{L}_b} \tag{3.11} $$

In the previous expression, L_b is the geodetic latitude of the body and R_E is the transverse radius of curvature.

3. Position error propagation (Groves, 2013):

The time derivative of the position error in frame $\{\mathcal{E}\}$ is simply the velocity error described as:

$$ {}^{\mathcal{E}}\delta\dot{\mathbf{p}}_{\mathcal{B}}^{\mathrm{T}} = {}^{\mathcal{E}}\delta\mathbf{v}_{\mathcal{B}}^{\mathrm{T}} \tag{3.12} $$

4. Propagation of biases:

The accelerometer and gyroscope biases are assumed not to vary in time. Thus:

$$ \dot{\mathbf{b}}_a = 0, \dot{\mathbf{b}}_g = 0 \tag{3.13} $$

3.2.3 Linearised System Model

The linearised continuous system model defined for the INS error-state, $\mathbf{X}_{INS}(t)$ is:

$$ \delta\dot{\mathbf{X}}_{INS}(t) = \mathbf{F}(t)\delta\mathbf{X}_{INS}(t) + \mathbf{G}(t)\mathbf{W}_{INS}(t) \tag{3.14} $$

where $\delta\dot{\mathbf{X}}_{INS}$ is the time derivative of the INS error-state, $\mathbf{F}$ is the system matrix with respect to the ECEF frame, given by:

$$\mathbf{F} = \begin{bmatrix} -\lfloor(\hat{\omega})\wedge\rfloor & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & -\mathbf{I}_{3,3} \\ \mathbf{F}_{21} & -2\boldsymbol{\Omega}_{\mathcal{E}} & \mathbf{F}_{23} & -\hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{I}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} \tag{3.15}$$

where
$\mathbf{F}_{21} = \lfloor -(\hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\hat{\mathbf{f}})\wedge\rfloor$, and $\mathbf{F}_{23} = \frac{2g_0(\hat{L}_b)}{{}^{\mathcal{E}}r_S(\hat{L}_b)} \frac{{}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}}{|{}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}|^2} {}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}^{\mathrm{T}}$ (Groves, 2013). The $\mathbf{G}$ matrix and $\mathbf{W}_{INS}$ reflect system noise and its propagation in time. Assuming the system noise is zero-mean, white Gaussian, modelling the INS measurement noise, i.e. $\mathbf{W}_{INS} = [\mathbf{w}_{rg}^{\mathrm{T}}\ \mathbf{w}_{rg}^{\mathrm{T}}\ \mathbf{w}_{bdg}^{\mathrm{T}}\ \mathbf{w}_{bda}^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{w}_{rg}$, $\mathbf{w}_{ra}$, $\mathbf{w}_{bdg}$, and $\mathbf{w}_{bda}$ are gyro and accelerometer noises and their corresponding bias variations, respectively, the matrix $\mathbf{G}$ is given by:

$$\mathbf{G} = \begin{bmatrix} -\mathbf{I}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & -\hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{I}_{3,3} & \mathbf{0}_{3,3} \\ \mathbf{I}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} \end{bmatrix} \tag{3.16}$$

3.2.4 Discrete-time implementation

Considering that τ_s is the INS sampling time, the continuous time equation is transformed into a discrete-time equation as (Groves, 2013):

$$\hat{\mathbf{X}}_k = \boldsymbol{\Pi}_{k,k-1}\hat{\mathbf{X}}_{k-1} \tag{3.17}$$

where, $\mathbf{\Pi}_{k,k-1}$ is the 15×15 transition matrix from step $k-1$ to k linked to the system matrix, $\mathbf{F}$, by:

$$\mathbf{\Pi}_{k,k-1} \approx \mathbf{I} + \mathbf{F}(t_k, t_{k-1})\tau_s \tag{3.18}$$

Furthermore, the EKF covariance matrix, which has to be propagated, is introduced as:

$$\mathbf{P}_{k|k} = \begin{bmatrix} \mathbf{P}_{II_{k|k}} & \mathbf{P}_{IC_{k|k}} \\ \mathbf{P}_{CI_{k|k}} & \mathbf{P}_{CC_{k|k}} \end{bmatrix} \tag{3.19}$$

where $\mathbf{P}_{II_{k|k}}$ is the 15×15 covariance matrix of the evolving INS state, $\mathbf{P}_{CC_{k|k}}$ is the $6M\times6M$ covariance matrix of the camera pose parameters, $\tilde{\boldsymbol{\xi}}_{\mathcal{C}_i} = [\boldsymbol{\Phi}_{\mathcal{C}_i}^{\mathrm{T}}\ {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}^{\mathrm{T}}]^{\mathrm{T}}$, $i = 1\cdots M$, and $\mathbf{P}_{IC_{k|k}}$ is the correlation between the errors in the INS state and the camera pose estimates. Using this notation, the propagated covariance matrix is given by (Mourikis, 2006):

$$\mathbf{P}_{k+1|k} = \begin{bmatrix} \mathbf{P}_{II_{k+1|k}} & \mathbf{\Pi}(k+1,k)\mathbf{P}_{IC_{k|k}} \\ \mathbf{P}_{IC_{k|k}}^{\mathrm{T}}\mathbf{\Pi}(k+1,k)^{\mathrm{T}} & \mathbf{P}_{CC_{k|k}} \end{bmatrix} \tag{3.20}$$

3.2.5 State Augmentation

As soon as an image is recorded, the corresponding camera pose parameters are added to the state to augment it. These parameters are related to the INS pose estimate By:

$$\hat{\mathbf{q}}_{\mathcal{E}}^{\mathcal{C}} = \mathbf{q}_{\mathcal{B}}^{\mathcal{C}} \otimes \hat{\mathbf{q}}_{\mathcal{E}}^{\mathcal{B}} \tag{3.21}$$

where $\hat{\mathbf{q}}_{\mathcal{E}}^{\mathcal{C}}$ is the unit quaternion representing rotation from frame $\{\mathcal{E}\}$ to the camera frame, $\{\mathcal{C}\}$, and $\mathbf{q}_{\mathcal{B}}^{\mathcal{C}}$ defines the rotation from the body frame to the camera frame computed off-line from the calibration between camera and INS. This calibration is possible by ceaseless-time batch estimation and continuously incorporating time offset between outputs of the INS and the camera within a maximum likelihood estimator (Furgale et al., 2013). The symbol, $\otimes$, denotes the quaternion multiplication.

The position of the camera with respect to the ECEF frame is also defined as (Mourikis,

2006):

$$ {}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{C}}^{\mathrm{T}} = {}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{B}}^{\mathrm{T}} + \hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\, {}^{\mathcal{B}}\mathbf{p}_{\mathcal{C}}^{\mathrm{T}} \tag{3.22} $$

where ${}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{C}}^{\mathrm{T}}$ is the position of the camera with respect to the ECEF frame and ${}^{\mathcal{B}}\mathbf{p}_{\mathcal{C}}^{\mathrm{T}}$ is the displacement vector between the camera frame and the body frame obtained through a calibration process.

Eventually, the covariance matrix of the EKF is augmented, accordingly (Mourikis, 2006):

$$ \mathbf{P}_{k|k} \equiv \begin{bmatrix} \mathbf{I}_{6N+15} \\ \mathbf{J} \end{bmatrix} \mathbf{P}_{k|k} \begin{bmatrix} \mathbf{I}_{6N+15} & \mathbf{J}^{\mathrm{T}} \end{bmatrix} \tag{3.23} $$

where $\mathbf{J}$ is the Jacobian matrix of equations (3.21) and (3.22) with respect to the state vector at timestep k, given by (Mourikis, 2006):

$$ \mathbf{J} = \begin{bmatrix} \mathbf{C}_{\mathcal{B}}^{\mathcal{C}} & \mathbf{0}_{3,3} & \mathbf{0}_{3,3} & \mathbf{0}_{3,6} & \mathbf{0}_{3,6N} \\ \lfloor (\hat{\mathbf{C}}_{\mathcal{B}}^{\mathcal{E}}\, {}^{\mathcal{B}}\mathbf{p}_{\mathcal{C}}) \wedge \rfloor & \mathbf{0}_{3,3} & \mathbf{I}_{3,3} & \mathbf{0}_{3,6} & \mathbf{0}_{3,6N} \end{bmatrix} \tag{3.24} $$

3.2.6 Measurement Model

In a standard EKF, the measurement vector, $\mathbf{y}$ is described as a linear model of the state and the corresponding white noise sources:

$$ \mathbf{y} = \mathbf{H}\delta\mathbf{X} + \mathbf{n} \tag{3.25} $$

where $\mathbf{n}$ is white noise and $\mathbf{H}$ is the measurement Jacobian matrix determined from the known properties of the system.

Considering a single feature observed from a set of N_j camera poses, $i \in \mathcal{S}_j$, the position of feature j in the frame $\{\mathcal{C}_i\}$ of a perspective camera is defined as:

$$ {}^{\mathcal{C}_i}\mathbf{p}_{f_j}^{\mathrm{T}} = \begin{bmatrix} {}^{\mathcal{C}_i}X_j \\ {}^{\mathcal{C}_i}Y_j \\ {}^{\mathcal{C}_i}Z_j \end{bmatrix} = \mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i}({}^{\mathcal{E}}\mathbf{p}_{f_j}^{\mathrm{T}} - {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}^{\mathrm{T}}) \tag{3.26} $$

where, ${}^{\mathcal{C}_i}\mathbf{p}_{f_j}^{\mathrm{T}}$ is the position of feature f_j with respect to the ith camera frame, $\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i}$ is the rotation matrix from the Earth frame to the ith camera frame, ${}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}^{\mathrm{T}}$ is the position of ith camera in the global frame, and ${}^{\mathcal{E}}\mathbf{p}_{f_j}^{\mathrm{T}}$ is the position of feature f_j in frame $\{\mathcal{E}\}$, which is unknown and must be estimated (*cf.* Appendix).

Suppose that the camera is pre-calibrated, the visual measurement model is then described as:

$$ \mathbf{z}_i^{(j)} = \frac{1}{{}^{\mathcal{C}_i}Z_j} \begin{bmatrix} {}^{\mathcal{C}_i}X_j \\ {}^{\mathcal{C}_i}Y_j \end{bmatrix} + \boldsymbol{\eta}_i^{(j)} \tag{3.27} $$

where, $\mathbf{z}_i^{(j)}$ is the measurement of feature f_j in the perspective camera frame $\mathcal{C}_i$. $\boldsymbol{\eta}$ is the 2×1 noise vector on image coordinates, with corresponding covariance matrix $\mathbf{R} = \sigma_{im}^2 \times \mathbf{I}_{2,2}$.

Having estimated the global position of the feature, the related measurement residual is described as:

$$ \mathbf{r}_i^{(j)} = \mathbf{z}_i^{(j)} - \hat{\mathbf{z}}_i^{(j)} \tag{3.28} $$

where $\mathbf{z}_i^{(j)}$ is the measurement of feature f_j in the camera frame $\mathcal{C}_{li}$ calculated from equation (3.27). Linearising the residuals with respect to camera pose parameters and the 3D position of features, the previous equation is expressed as:

$$ \mathbf{r}_i^{(j)} \simeq \mathbf{H}_{\mathbf{X}_i}^{(j)} \delta\mathbf{X} + \mathbf{H}_{f_i}^{(j)} \, {}^{\mathcal{E}}\delta\mathbf{p}_{f_j} + \mathbf{n}_i^{(j)} \tag{3.29} $$

where ${}^{\mathcal{E}}\delta\mathbf{p}_{f_j}$ is the error in the position estimate of f_j, and $\mathbf{H}_{\mathbf{X}_i}^{(j)}$ and $\mathbf{H}_{f_i}^{(j)}$ are the Jacobian matrices of measurement $\mathbf{z}_i^{(j)}$ with respect to the state and the feature position, respec-

tively, given by:

$$\mathbf{H}_{\mathbf{X}i}^{(j)} = \begin{bmatrix} \mathbf{0} & \mathbf{J}_i^{(j)} \lfloor ({}^{\mathcal{C}_i}\hat{\mathbf{p}}_{f_j} \wedge \rfloor & -\mathbf{J}_i^{(j)} \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_i} & \mathbf{0} \end{bmatrix} \tag{3.30}$$

$$\mathbf{H}_{f_i}^{(j)} = \mathbf{J}_i^{(j)} \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_i} \tag{3.31}$$

In equation (3.30), the dimensions of the left and the right zeros are $2\times(15+6(i-1))$ and $2\times 6(N_j - i)$, respectively. The Jacobian matrix $\mathbf{J}_i^{(j)}$ is expressed by:

$$\mathbf{J}_i^{(j)} = \frac{1}{({}^{\mathcal{C}_i}\hat{Z}_j)^2} \begin{bmatrix} {}^{\mathcal{C}_i}\hat{Z}_j & 0 & -{}^{\mathcal{C}_i}\hat{X}_j \\ 0 & {}^{\mathcal{C}_i}\hat{Z}_j & -{}^{\mathcal{C}_i}\hat{Y}_j \end{bmatrix} \tag{3.32}$$

By stacking the residuals $\mathbf{r}_i^{(j)}$, we obtain a $2N_j\times 1$ residual vector, $\mathbf{r}^{(j)}$, where N_j is the number of camera poses in which feature f_j is visible. This block of residuals is defined as:

$$\mathbf{r}^{(j)} \approx \mathbf{H}_{\mathbf{X}}^{(j)}\, \delta\mathbf{X} + \mathbf{H}_f^{(j)}\, {}^{\mathcal{E}}\delta\mathbf{p}_{f_j} + \mathbf{n}^{(j)} \tag{3.33}$$

where $\mathbf{H}_{\mathbf{X}}^{(j)}$ and $\mathbf{H}_f^{(j)}$ are vector blocks including $\mathbf{H}_{\mathbf{X}i}^{(j)}$ and $\mathbf{H}_{fi}^{(j)}$, respectively.

Since the error state, $\delta\mathbf{X}$, correlates with the error of feature position, ${}^{\mathcal{E}}\delta\mathbf{p}_{f_j}$, the residuals cannot be directly applied in an EKF estimation. To resolve this issue, the residual vector, $\mathbf{r}^{(j)}$ is projected onto the left null-space of the $2N_j\times 3$ matrix $\mathbf{H}_f^{(j)}$ (Mourikis, 2006):

$$\mathbf{A}^{\mathrm{T}}\mathbf{r}^{(j)} \simeq \mathbf{A}^{\mathrm{T}}\mathbf{H}_{\mathbf{X}}^{(j)}\delta\mathbf{X} + \mathbf{A}^{\mathrm{T}}\mathbf{n}^{(j)} \tag{3.34}$$

where, $\mathbf{A}$ is a $2N_j\times(2N_j-3)$ unitary matrix with the columns from the basis of the left null-space of $\mathbf{H}_f$. With a slight change in notation, we define the standard measurement model used is an EKF as:

$$\mathbf{r}_0^{(j)} = \mathbf{H}_0^{(j)}\delta\mathbf{X}^{(j)} + \mathbf{n}_0^{(j)} \tag{3.35}$$

where

$$\mathbf{r}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{r}^{(j)},\ \mathbf{H}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{H}_{\mathbf{X}}^{(j)},\ \mathbf{n}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{n}^{(j)} \tag{3.36}$$

Projecting $\mathbf{r}^{(j)}$ onto the left null-space of the matrix $\mathbf{H}_f^{(j)}$, the term related to the error in the position estimate of f_j is cancelled out, and $\mathbf{r}_0$ can now be used in the filter update step.

3.2.7 Update

Having matched salient features from pairs of images sequentially, and having tracked them over a sequence of camera poses, the observations are entirely used as geometric constraints to update the EKF state. By considering that K salient features are visible from N_j sets of camera poses, $j = 1 \cdots K$, the block vector of the entire residuals is described as:

$$\mathbf{r}_0 = \mathbf{H}_0 \delta \mathbf{X} + \mathbf{n}_0 \tag{3.37}$$

where, $\mathbf{r}_0$ is a block vector including the vectors of residuals, $\mathbf{r}_0^{(j)}$, and $\mathbf{n}_0$ is the measurement noise comprising the components $\mathbf{n}_0^{(j)}$. The measurement matrix $\mathbf{H}_0$ is a matrix with block rows $\mathbf{H}_0^{(j)}$.

Due to the fact that image observations can be considered as statistically independent of each other, and their corresponding noise vectors are uncorrelated, the full noise covariance matrix is formed as a diagonal matrix, i.e. $\mathbf{R}_0 = \sigma_{im}^2 \mathbf{I}_\lambda$, where $\lambda = \sum_{j=1}^{K} (2N_j - 3)$ is the length of the entire residuals.

In practice, a large number of observations leads to an exponential growth in the computational cost, therefore affecting the real-time performance of the algorithm. To increase the efficiency of the algorithm when many features are tracked, the QR decomposition of matrix $\mathbf{H}_0$ is employed (Bayard and Brugarolas, 2005), which is defined as:

$$\mathbf{H}_0 = \mathbf{Q}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_\mathcal{H} \\ \mathbf{0} \end{bmatrix} \tag{3.38}$$

where $\mathbf{Q}$ is an orthogonal matrix whose columns are orthogonal unit vectors, i.e. $\mathbf{Q}^\mathrm{T}\mathbf{Q} =$

$\mathbf{I}$, $\mathbf{R}$ is an upper triangular matrix whose $(\lambda-6M-15)$ lower rows are entirely zeros and $\mathbf{R}_{\mathcal{H}}$ is a $(15+6M)\times(15+6M)$ upper triangular matrix. $\mathbf{Q}_1$ has dimensions $\lambda\times(6M+15)$ and $\mathbf{Q}_2$ $\lambda\times(\lambda-6M-15)$ and both have orthogonal columns. Decomposing matrix $\mathbf{H}_0$ into two components, equation (3.37) is re-formed as:

$$\mathbf{r}_0 = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\mathcal{H}} \\ \mathbf{0} \end{bmatrix} \delta\mathbf{X} + \mathbf{n}_0 \tag{3.39}$$

$$\mathbf{Q}^{\mathrm{T}}\mathbf{r}_0 = \begin{bmatrix} \mathbf{R}_{\mathcal{H}} \\ \mathbf{0} \end{bmatrix} \delta\mathbf{X} + \mathbf{Q}^{\mathrm{T}}\mathbf{n}_0 \Rightarrow \begin{bmatrix} \mathbf{Q}_1^{\mathrm{T}}\mathbf{r}_0 \\ \mathbf{Q}_2^{\mathrm{T}}\mathbf{r}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{\mathcal{H}} \\ \mathbf{0} \end{bmatrix} \delta\mathbf{X} + \begin{bmatrix} \mathbf{Q}_1^{\mathrm{T}}\mathbf{n}_0 \\ \mathbf{Q}_2^{\mathrm{T}}\mathbf{n}_0 \end{bmatrix} \tag{3.40}$$

In the preceding equation, the component $\mathbf{Q}_2^{\mathrm{T}}\mathbf{r}_0$ conveys pure noise because the corresponding rows in the measurement matrix are zeros, so with no useful information. Therefore, this component is discarded to allow the EKF state to be updated by the residuals having reduced dimension.

Lastly, the EKF state is updated by the Kalman gain matrix, $\mathbf{K}$, given by:

$$\mathbf{K} = \mathbf{P}\mathbf{R}_{\mathcal{H}}^{\mathrm{T}}(\mathbf{R}_{\mathcal{H}}\mathbf{P}\mathbf{R}_{\mathcal{H}}^{\mathrm{T}} + \mathbf{R}_{\mathcal{N}})^{-1} \tag{3.41}$$

where $\mathbf{P}$ is the state covariance matrix and $\mathbf{R}_{\mathcal{N}}$ is the covariance matrix of the noise component, $\mathbf{Q}_1^{\mathrm{T}}\mathbf{r}_0$, computed from propagation of uncertainty, i.e. $\mathbf{R}_{\mathcal{N}} = \mathbf{Q}_1^{\mathrm{T}}\mathbf{R}_0\mathbf{Q}_1$.

The state vector is corrected with the reduced measurement using:

$$\hat{\mathbf{X}}_{k+1|k+1} = \hat{\mathbf{X}}_{k+1|k} + \mathbf{K}\mathbf{Q}_1^{\mathrm{T}}\mathbf{r}_0 \tag{3.42}$$

Simply, the posterior error covariance matrix is updated with:

$$\mathbf{P}_{k+1|k+1} = (\mathbf{I} - \mathbf{K}\mathbf{R}_{\mathcal{H}})\mathbf{P}_{k+1|k}(\mathbf{I} - \mathbf{K}\mathbf{R}_{\mathcal{H}})^{\mathrm{T}} + \mathbf{K}\mathbf{R}_{\mathcal{N}}\mathbf{K}^{\mathrm{T}} \tag{3.43}$$

where, $\mathbf{I}$ is a $(6M+15)\times(6M+15)$ identity matrix.

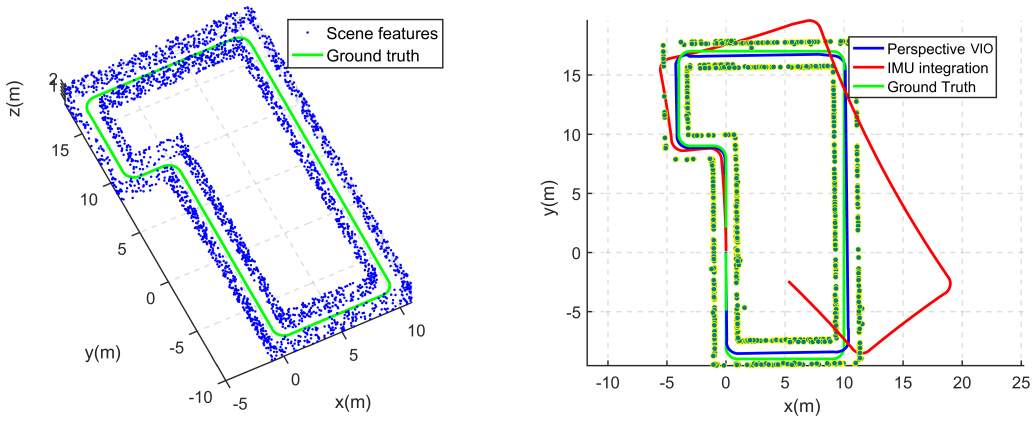


Figure 3.2: **Left**: The simulated indoor environment with the generated feature points. **Right**: The top view of the simulated indoor environment and the estimated trajectory (blue) compared with the INS trajectory (red) and the ground truth (green).

3.3 A Simulation Using Predefined Features

In order to evaluate the performance of mono-VIO a simulation using predefined features has been carried out. In this simulation, a total of 3000 features were randomly generated on the walls of a corridor along a ground truth trajectory with an approximate length of 77 meters. The image measurements of the features were recorded at 1 Hz simulating a perspective camera moving at a constant velocity. The INS measurements were simulated at 100 Hz using points on the ground truth trajectory. The INS measurements were added with a constant error and a random noise for simulating the biases associated with the inertial measurements. Figure 3.2 shows the simulated environment with the generated features points, the ground truth trajectory, the trajectory obtained from the simulated INS measurements only, and the trajectory estimated by the VIO approach. It is evident that while the INS trajectory quickly drifts away, the VIO trajectory closely follows the ground truth.

Figure 3.3 illustrates the cumulative distribution of translational and rotational errors for the perspective MSCKF approach and for the INS integration. The total RMSE, defined as the root mean square of RMSEs along the trajectory, for translation and rotation,

is about 0.5 m and 0.04 rad, respectively, for the VIO approach, whereas these values for the INS-only integration are 6.15 m and 0.24 rad. Considering the trajectory length, the perspective VIO reaches an overall localization accuracy of 0.6% of the trajectory length compared to 8% for the INS-only integration. Note that this accuracy is achieved without considering any feature tracking error.

3.4 Summary

This chapter has introduced an EKF-based estimator for real-time visual-inertial odometry. The formulation of this methodology was thoroughly described. As noted, this methodology endeavours to benefit from the maximum potential of constraints that a visible scene feature contributes to from a multitude of successive images. Moreover, the computational complexity of the method is linear in the number of features. The result of the simulation, which focuses on the performance of the MSCKF by using a set of predefined features, shows the MSCKF capability to accurately estimate motion in large-scale environments. In the following chapters, this method is compared to the result of the proposed algorithms.

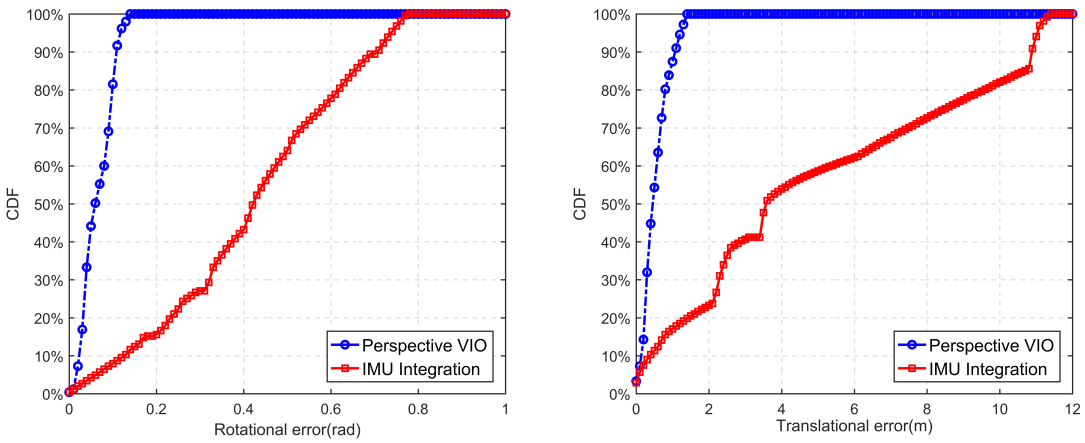


Figure 3.3: **Left**: Cumulative distribution of translational error for the perspective VIO approach and the INS integration. **Right**: Cumulative distribution of rotational error for the perspective VIO approach and INS integration.

Chapter 4

Visual-Inertial Odometry Using a Stereo Camera

THIS chapter presents the first proposed methodology in this thesis, stereo-VIO, using the multi-state constraint Kalman filtering method. Because the main difference of this approach with the mono-MSCKF is in the measurement model while the formulation of the system model and update are the same as the mono-MSCKF, the description of the method focuses on the measurement model. In addition, a description of feature detection, matching and tracking is given in this chapter.

4.1 Introduction

Although the MSCKF takes full advantage of the constraints that a scene point provides, it has only been utilized for the integration of observations from an INS and visual observations taken from a single camera. To improve the quality of feature tracking and effectiveness of measurements in the update step, a fixed baseline between pairs of cameras in the stereo configuration is exploited in order to impose additional constraints on the motion estimation. Figure 4.1 represents the principal steps of the stereo-MSCKF. Features are extracted and matched between the left and right camera images. The state is augmented with the pose parameters of the left camera. Given the calibration parameters between the two cameras, the pose of the right camera is calculated from the left. Finally, the state is updated by the measurements obtained from both cameras at each

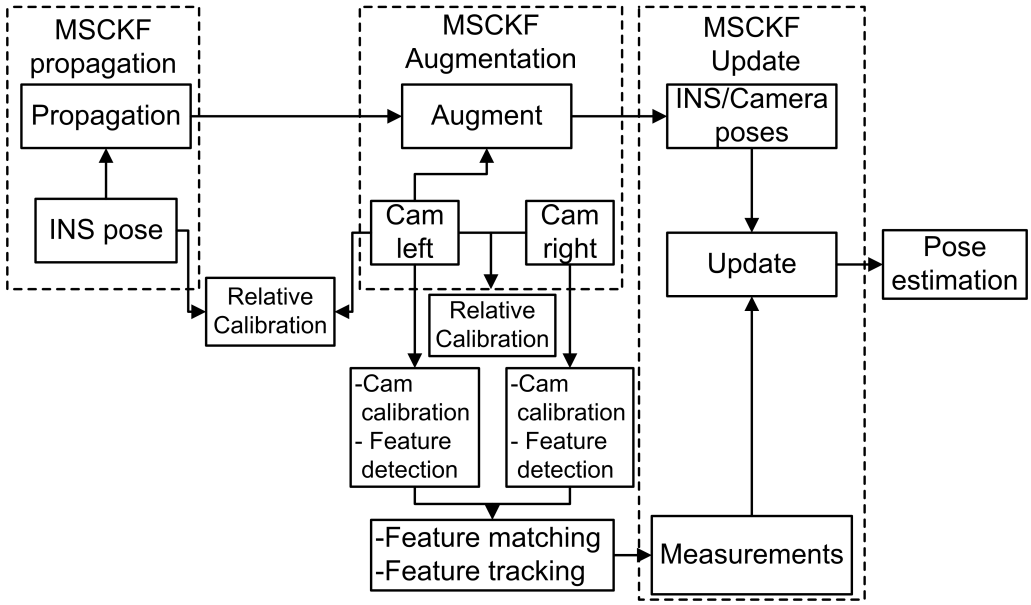


Figure 4.1: The general workflow of the stereo MSCKF. Images captured from a pair of cameras at each timestep are conducted to the stereo matching process. The stereo matching imposes additional constraints resulting in better estimation.

timestep.

4.2 Methodology of the Stereo MSCKF

Similar to the mono-MSCKF, the state vector and the INS error-state vector are defined according to equations (3.1) and (3.2), respectively. Since the poses of the left and right cameras are dependent, only the poses of the left camera are concatenated into the state vector forming the full error-state vector similar to equation (3.5):

$$\delta\mathbf{X} = [\delta\mathbf{X}_{INS} \;\; \delta\mathbf{\Phi}_{\mathcal{C}_{l1}}^{\mathrm{T}} \; {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_{l1}}^{\mathrm{T}} \;\; \cdots \;\; \delta\mathbf{\Phi}_{\mathcal{C}_{lM}}^{\mathrm{T}} \; {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_{lM}}^{\mathrm{T}}]^{\mathrm{T}} \tag{4.1}$$

Each component of the INS error-state vector is propagated according to the presented equations (3.6) to (3.13) and the created system model is thereinafter linearised in accordance with equations (3.14) to (3.20). Now, once a stereo image is captured, the left camera pose parameters are adhered to the state vector using equations (3.21) and

(3.22) and the covariance matrix is also augmented according to equation (3.23).

Using the standard perspective model, the global position of a feature f_j is estimated and used for the calculation of the measurement residual of the left camera :

$$\mathbf{r}_{li}^{(j)} = \mathbf{z}_{li}^{(j)} - \hat{\mathbf{z}}_{li}^{(j)} \tag{4.2}$$

where $\mathbf{z}_{li}^{(j)}$ is the measurement of feature f_j in the left camera frame $\mathcal{C}_{li}$ calculated from equation (3.27). Linearising the residuals with respect to camera pose parameters and the 3D position of features, the previous equation is expressed as:

$$\mathbf{r}_{li}^{(j)} \simeq \mathbf{H}_{l\mathbf{X}_i}^{(j)} \delta\mathbf{X} + \mathbf{H}_{lf_i}^{(j)} \, {}^{\mathcal{E}}\delta\mathbf{p}_{f_j} + \mathbf{n}_{li}^{(j)} \tag{4.3}$$

where ${}^{\mathcal{E}}\delta\mathbf{p}_{f_j}$ is the error in the position estimate of f_j, and $\mathbf{H}_{l\mathbf{X}_i}^{(j)}$ and $\mathbf{H}_{lf_i}^{(j)}$ are the Jacobian matrices of measurement $\mathbf{z}_{li}^{(j)}$ with respect to the state and the feature position, respectively, given by:

$$\mathbf{H}_{l\mathbf{X}i}^{(j)} = \begin{bmatrix} \mathbf{0} & \mathbf{J}_i^{(j)} \lfloor ({}^{\mathcal{C}_{li}}\hat{\mathbf{p}}_{f_j} \wedge \rfloor & -\mathbf{J}_i^{(j)} \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_{li}} & \mathbf{0} \end{bmatrix} \tag{4.4}$$

$$\mathbf{H}_{lf,i}^{(j)} = \mathbf{J}_i^{(j)} \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_{li}} \tag{4.5}$$

In equation (4.4), the dimension of the left and the right zeros are $2\times(15+6(i-1))$ and $2\times6(N_j-i)$ respectively. The Jacobian matrix $\mathbf{J}_i^{(j)}$ is expressed by:

$$\mathbf{J}_i^{(j)} = \frac{1}{({}^{\mathcal{C}_{li}}\hat{Z}_j)^2} \begin{bmatrix} {}^{\mathcal{C}_{li}}\hat{Z}_j & 0 & -{}^{\mathcal{C}_{li}}\hat{X}_j \\ 0 & {}^{\mathcal{C}_{li}}\hat{Z}_j & -{}^{\mathcal{C}_{li}}\hat{Y}_j \end{bmatrix} \tag{4.6}$$

Now we define the pose of the right camera, $\boldsymbol{\xi}_{\mathcal{C}_r} = [\mathbf{q}_{\mathcal{C}_r}^{\mathrm{T}} \; {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_r}^{\mathrm{T}}]^{\mathrm{T}}$, with respect to the left camera pose and the relative calibration parameters between them:

$${}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_r} = {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_l} - \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_l\mathrm{T}} \, \mathbf{C}_{\mathcal{C}_l}^{\mathcal{C}_r\mathrm{T}} \, {}^{\mathcal{C}_r}\mathbf{p}_{\mathcal{C}_l} \tag{4.7}$$

where, $\mathbf{C}_{\mathcal{C}_l}^{\mathcal{C}_r}$ and ${}^{\mathcal{C}_r}\mathbf{p}_{\mathcal{C}_l}$ are the rotation matrix and displacement vector between the left and the right camera, respectively. Matrix $\hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_l}$ represents the rotation between the left camera and the Earth frame.

$$\mathbf{q}_{\mathcal{E}}^{\mathcal{C}_r} = \mathbf{q}_{\mathcal{C}_l}^{\mathcal{C}_r} \otimes \mathbf{q}_{\mathcal{E}}^{\mathcal{C}_l} \tag{4.8}$$

where the unit quaternion $\mathbf{q}_{\mathcal{C}_l}^{\mathcal{C}_r}$ represents the rotation from the left camera coordinate system to the right. We also obtain the position of feature f_j with respect to the right camera frame at the ith camera pair:

$${}^{\mathcal{C}_{ri}}\hat{\mathbf{p}}_{f_j} = \hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_{ri}}\left({}^{\mathcal{E}}\hat{\mathbf{p}}_{f_j} - {}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{C}_{ri}}\right) = \mathbf{C}_{\mathcal{C}_l}^{\mathcal{C}_r}\,{}^{\mathcal{C}_{li}}\hat{\mathbf{p}}_{f_j} + {}^{\mathcal{C}_r}\mathbf{p}_{\mathcal{C}_l} \tag{4.9}$$

The linearisation of measurement residuals in equation (4.2) around the estimates for the left camera and the feature position is now defined as:

$$\mathbf{r}_{ri}^{(j)} \simeq \mathbf{H}_{r\mathbf{X}i}^{(j)}\delta\mathbf{X} + \mathbf{H}_{rfi}^{(j)}\,{}^{\mathcal{E}}\delta\mathbf{p}_{f_j} + \mathbf{n}_{ri}^{(j)} \tag{4.10}$$

where $\mathbf{H}_{r\mathbf{X}i}^{(j)}$ and $\mathbf{H}_{rfi}^{(j)}$ are the Jacobians of the measurement of feature f_j from the right camera pose with respect to the error state and the global position of the feature, respectively, given by:

$$\mathbf{H}_{r\mathbf{X}i}^{(j)} = \begin{bmatrix}\mathbf{0} & \mathbf{J}_i^{(j)}\lfloor({}^{\mathcal{C}_{ri}}\hat{\mathbf{p}}_{f_j})\wedge\rfloor & -\mathbf{J}_i^{(j)}\mathbf{C}_{\mathcal{C}_l}^{\mathcal{C}_r}\hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_{li}} & \mathbf{0}\end{bmatrix} \tag{4.11}$$

$$\mathbf{H}_{rfi}^{(j)} = \mathbf{J}_i^{(j)}\mathbf{C}_{\mathcal{C}_l}^{\mathcal{C}_r}\hat{\mathbf{C}}_{\mathcal{E}}^{\mathcal{C}_{li}} \tag{4.12}$$

By stacking the residuals $\mathbf{r}_{li}^{(j)}$ and $\mathbf{r}_{ri}^{(j)}$ separately, a $4N_j \times 1$ residual vector is obtained, $\mathbf{r}^{(j)} = [\mathbf{r}_l^{(j)\mathrm{T}}\ \mathbf{r}_r^{(j)\mathrm{T}}]^{\mathrm{T}}$, where N_j is the number of stereo camera poses in which feature f_j is visible. By combining equations (4.3) and (4.10), the block of all residuals is formed as:

$$\mathbf{r}^{(j)} \approx \mathbf{H}_{\mathbf{X}}^{(j)}\,\delta\mathbf{X} + \mathbf{H}_f^{(j)}\,{}^{\mathcal{E}}\delta\mathbf{p}_{f_j} + \mathbf{n}^{(j)} \tag{4.13}$$

where

$$\mathbf{H}_{\mathbf{X}}^{(j)} = \begin{bmatrix} \mathbf{H}_{l\mathbf{X}}^{(j)} \\ \mathbf{H}_{r\mathbf{X}}^{(j)} \end{bmatrix}, \ \mathbf{H}_{f}^{(j)} = \begin{bmatrix} \mathbf{H}_{lf}^{(j)} \\ \mathbf{H}_{rf}^{(j)} \end{bmatrix}, \ \mathbf{n}^{(j)} = \begin{bmatrix} \mathbf{n}_{l}^{(j)} \\ \mathbf{n}_{r}^{(j)} \end{bmatrix} \tag{4.14}$$

Since the error state, $\delta\mathbf{X}$, correlates with the error of feature position, ${}^{\mathcal{E}}\delta\mathbf{p}_{f_j}$, the residuals cannot be directly applied in an EKF estimation. To resolve this issue, the residual vector, $\mathbf{r}^{(j)}$ is projected onto the left null-space of the $4N_j \times 3$ matrix $\mathbf{H}_f^{(j)}$ (Mourikis, 2006):

$$\mathbf{A}^{\mathrm{T}}\mathbf{r}^{(j)} \simeq \mathbf{A}^{\mathrm{T}}\mathbf{H}_{\mathbf{X}}^{(j)}\delta\mathbf{X} + \mathbf{A}^{\mathrm{T}}\mathbf{n}^{(j)} \tag{4.15}$$

where $\mathbf{A}$ is a $4N_j \times (4N_j - 3)$ unitary matrix with the columns from the basis of the left null-space of $\mathbf{H}_f$. With a slight change in notation, the standard measurement model used is an EKF is defined as:

$$\mathbf{r}_0^{(j)} = \mathbf{H}_0^{(j)}\delta\mathbf{X}^{(j)} + \mathbf{n}_0^{(j)} \tag{4.16}$$

where,

$$\mathbf{r}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{r}^{(j)}, \ \mathbf{H}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{H}_{\mathbf{X}}^{(j)}, \ \mathbf{n}_0^{(j)} = \mathbf{A}^{\mathrm{T}}\mathbf{n}^{(j)} \tag{4.17}$$

In projecting $\mathbf{r}^{(j)}$ onto the left null-space of the matrix $\mathbf{H}_f^{(j)}$, the term related to the error in the position estimate of f_j is cancelled out, and $\mathbf{r}_0$ can now be used in the filter update step.

The block vector of residuals from the pairs of images is used in the update step, (3.37) to (3.43), to update the error-state vector as well as its relevant covariance matrix.

4.3 Feature Detection, Matching and Tracking

Vision-aided navigation systems, which utilize feature-based matching between images, require the detection of distinctive features on each image and determination of their correspondences on the next frame in a matching process so as to track identical features in a sequence of images. Various steps of feature selection for mono and stereo images, used in this thesis, are illustrated in Figure 4.2. In the mono mode, a feature is detected

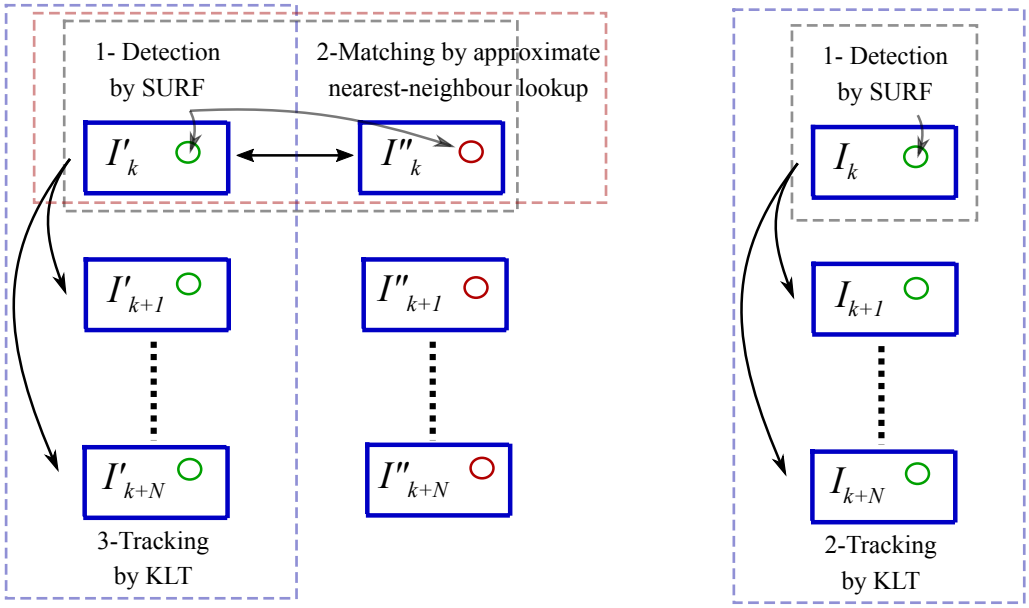


Figure 4.2: Steps of feature detection and tracking in mono (**Right**) and stereo (**Left**) sequence of images.

in the first image of a sequence of images from which the feature is visible and, in the second step, it is tracked by using a feature tracker. In the stereo mode, features are detected in the left and right images of the first stereo image of the tracking window. The corresponding features are then matched by utilizing nearest-neighbour matching. Finally, features are tracked, the same as in the mono mode, by employing a feature tracker.

4.3.1 Feature Detection

Many feature detectors and descriptors have broadly been proposed and investigated in the literature (Salahat and Qasaimeh, 2017; Lindeberg, 2015). In this work, the Speeded Up Robust Features (SURF) detector (Bay et al., 2008) is utilised. Similar to the Scale Invariant Feature Transform (SIFT) (Lowe, 2004), SURF is invariant not only to image scale and rotation, but also to a significant range of distortions caused by affinity, addition of noise and variation in illumination. Bay et al. (2008) based the SURF detector on the Hessian matrix (Lindeberg, 2015) due to its effective performance in terms of computation

time and accuracy. This is contrary to SIFT, which uses a Difference of Gaussian (DoG) approach (Lindeberg, 2015) to provide a close approximation to the scale-normalized Laplacian of Gaussian (LoG). Because the SURF detector relies on integral images for reduction of computation time, it is known as a fast Hessian detector. The descriptor of SURF is on the basis of Haar-wavelet distribution around the interest point. Using Haar-wavelets, only 64 dimensions are utilised to describe the descriptor, thus decreasing the time of computation and matching compared to SIFT where the descriptor includes 128 dimensions.

4.3.2 Feature Matching

Upon extracting features from a pair of images in stereo vision, nearest neighbour matching is employed to find the corresponding features. This approximate matching method is based on finding the minimum Euclidean distance for the descriptor proposed by Lowe (2004). In contrast with many techniques using a branch-and-bound approach that performs searches in depth-first order, the method proposed by Muja and Lowe (2009) applies a priority search on hierarchical k-means trees resulting in a faster matching process of high-dimensional vectors, up to several orders of magnitude, than that of linear search.

4.3.3 Feature Tracking

A common feature tracking method over a sequence of images is the Kanade-Lucas-Tomasi (KLT) feature tracker (Tomasi and Kanade, 1991). The problem of feature tracking in KLT can be characterized as follows: given two successive images whose intensities are represented as a function of 2-D space vector $\mathbf{x}$ and the time variable t, $I(\mathbf{x}+t)$, KLT endeavours to find the displacement $\mathbf{d}$ so that the following equation is minimized over a specified window $\mathcal{W}$:

$$e = \int_{\mathcal{W}} [I(\mathbf{x}-\mathbf{d},t) - I(\mathbf{x},t+k)]^2 w dx \tag{4.18}$$

where, w is a function representing weight and k is the time shift when the next image is taken.

By linearising equation (4.18) it is possible to calculate the displacement **d** on the image precisely only if the two successive images are close to each other. However, the existence of outliers is inevitable. Therefore, to remove outliers, an M-estimator Sample Consensus (MSAC) approach (Torr and Zisserman, 1998) has been applied in this work. This method discards mismatches by randomly selecting samples of minimal sets of points from plausible correspondences, and in a repetitive fashion it tries to find the set which minimizes a cost function defined as:

$$\mathcal{C} = \sum_i \lambda(\frac{v_i}{\sigma}) \tag{4.19}$$

where v_i are point errors, between the measured points and the maximum likelihood estimates, λ is a robust function representing a fixed cost given to outliers and σ is the standard deviation of points.

4.4 Summary

In this chapter, a stereo-VIO based on the multi-state constraint Kalman filtering method has been presented. Since the static stereo adds epipolar constraints perpendicular to the motion, it is expected that the stereo-MSCKF will outperform the mono-MSCKF. In addition, by simultaneously taking advantage of static stereo pairs and temporal stereo pairs taken in time, the system exploits the reconstruction of very small and very large scaled environments resulting in more accurate and robust motion estimation. To detect visual features, SURF has been selected due to its repeatability, meaning that it can find the same point of interest under different circumstances. In addition, it is robust in the presence of noise, detection errors, and geometric and photometric deformations. To match the feature points across pairs of images, the approximate nearest-neighbour approach has been adopted because it is more efficient than linear searching methods. Finally, to track

salient features over a sequence of images, KLT has been chosen due to its simplicity, high speed and accuracy. However, because this tracking method is a window-based method, the tracker only performs well if the window size is selected properly.

Chapter 5

Visual-Inertial Odometry Using an Omnidirectional Camera

THIS chapter details the proposed method of Omnidirectional Visual-Inertial Odometry (OVIO). Before the methodology is introduced, a discussion of omnidirectional camera calibration is necessary due to the high image distortion of these cameras. In the following, the proposed approach to integrate inertial measurements with spherically visual observation through the MSCKF is presented. Since the feature selection is similar to the previous approach, the steps of feature detection, matching and tracking are not mentioned here.

5.1 Introduction

To integrate omnidirectional image observations with inertial measurements, the MSCKF (Mourikis and Roumeliotis, 2007) algorithm is employed. However, the measurement model defined in the conventional MSCKF involves the residual vectors of the image features and their reprojection on the image plane. The image plane is not suited for omnidirectional images, which have large distortion. To resolve this issue, the residual vectors on a plane that is tangent to the unit sphere around each measurement ray can be calculated. Because the system model is similar to the MSCKF detailed in Chapter 3, the following description of the OVIO focuses on the proposed measurement model. However, because dealing with the distortion is crucial, the calibration of these sensors

is initially discussed.

5.2 Calibration

Camera calibration is a prerequisite for vision-aided localization systems. Since the distortion of omnidirectional images is quite high, a proper methodology for their calibration is crucial.

The projection of a scene point $\mathbf{X}$ onto the point $\mathbf{u}''$ is illustrated in Figure 5.1 for a fisheye lens and a type of mirror used in catadioptric systems. The surface of either the lens or the mirror is usually conical such as parabolic, hyperbolic, elliptical or planar. Similar to a standard perspective camera, it is assumed that the omnidirectional cameras employed have a single effective viewpoint. These types of cameras are known as central omnidirectional systems.

The geometry of central omnidirectional image formation has been thoroughly dis-

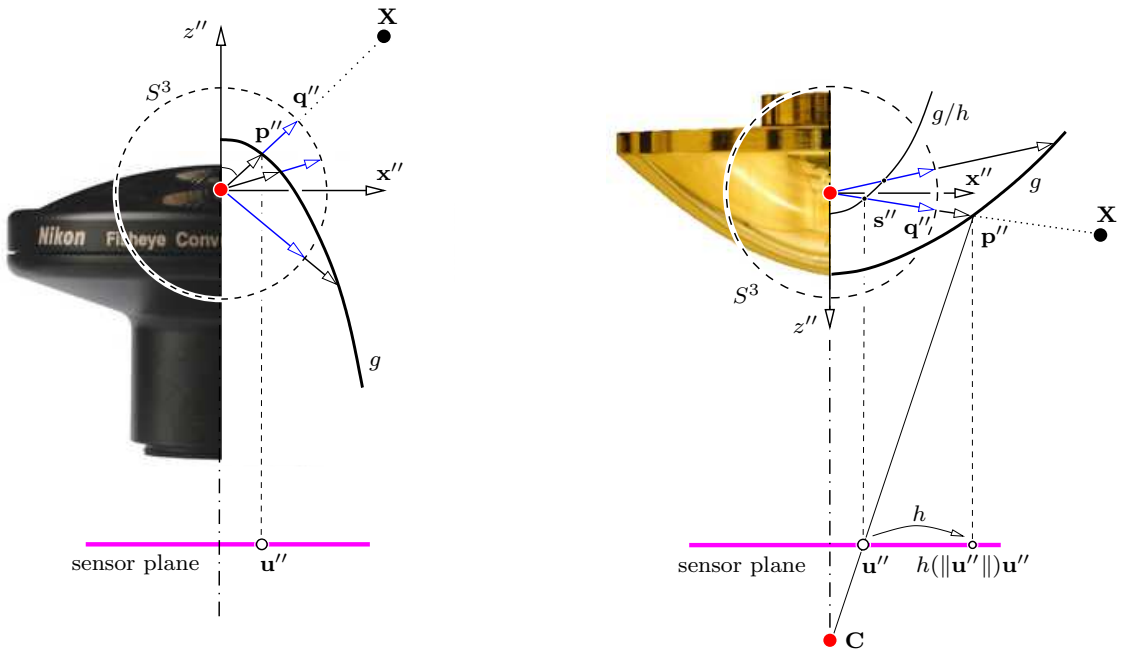


Figure 5.1: Projection process from 3D space to the 2D sensor plane in a dioptric camera (**Left**) and a catadioptric system (**Right**) (Micušık, 2004).

cussed in (Barreto and Araujo, 2001; Geyer and Daniilidis, 2000). As seen in Figure 5.1, in omnidirectional cameras, projection occurs in two steps: first the light travels straight from a scene point **X** toward the surface of either the lens or mirror. All light beams pass through the center of the lens or the mirror along vector $\mathbf{p}'' = [\mathbf{x}''^T, z'']^T$. Vector $\mathbf{p}''$ indicates the light beam between the projection center and the surface of the mirror or lens. Then, depending on the geometrical surface of the lens or mirror, the light beam is projected onto the sensor plane or the image plane if no misalignment exists between the sensor plane and the image plane. In the cases shown in Figure 5.1, the light travels vertically from the surface of the lens, which has a parabolic surface, to the sensor plane (**Left**), whereas for the hyperbolic shape of the mirror, it travels such that it passes through another focus which is typically the center of a perspective camera (**Right**).

Provided that there is a misalignment between the sensor plane and the image plane, the vector $\mathbf{u}''$ can be well related to the point on the image plane $\mathbf{u}'$, which is expressed in pixels, by an Affine transformation:

$$\mathbf{u}'' = \mathbf{A}\mathbf{u}' + \mathbf{t} \tag{5.1}$$

where, $\mathbf{A} \in \mathbb{R}^2$ and $\mathbf{t} \in \mathbb{R}$ are the components of an Affine transformation reducing misalignments caused by rotation, scaling, shearing and translation.

Vector $\mathbf{p}''$ can now be related to vector u'', which is collinear to $\mathbf{x}''$, using two highly non-linear functions g and h:

$$\mathbf{p}'' = \begin{bmatrix} h(||\mathbf{u}''||)\mathbf{u}'' \\ g(||\mathbf{u}''||) \end{bmatrix} = \begin{bmatrix} \mathbf{u}'' \\ \frac{g(||\mathbf{u}''||)}{h(||\mathbf{u}''||)} \end{bmatrix} \tag{5.2}$$

where g and h are non-linear functions of the Euclidean distance $||\mathbf{u}''||$ of the point $\mathbf{u}''$ to the sensor axis origin, i.e. the centre of symmetry, and their parameters vary depending on the type of refraction in dioptric systems or reflection in catadioptric systems (Scaramuzza, 2008b). These functions and their proportion g/h are visualized in Figure 5.1.

In order to simplify these functions, Scaramuzza et al. (2006a) proposed a very flexi-

ble, easy-to-use and independent-of-the-sensor calibration technique in which function h is considered to be equal to 1, therefore equation (5.2) can be written as:

$$\mathbf{p}'' = \begin{bmatrix} \mathbf{u}'' \\ g(||\mathbf{u}''||) \end{bmatrix} \tag{5.3}$$

Further, Scaramuzza et al. (2006a) used a polynomial to express function g:

$$g(||\mathbf{u}''||) = \sum_{i=0}^{N} a_i ||\mathbf{u}''||^i \tag{5.4}$$

where coefficients a_i, $i = 1 \cdots N$ are the polynomial parameters which need to be estimated within the calibration process.

By observing a planar pattern, e.g a chessboard, from various angles of view, the affine parameters and the polynomial coefficients, which constitute the interior parameters, are solved within a four-step optimization problem (Scaramuzza, 2008b). In the first step, only the camera exterior parameters are estimated by assuming that $\mathbf{A} = \mathbf{I}_2$ and $\mathbf{t} = 0$. In the second step, the interior parameters a_i, $i = 1 \cdots N$ are estimated. The third step aims to refine both the interior and the exterior parameters estimated so far. These three steps are performed via linear least-squares minimization. In the last step, the entire parameters including affine components are refined iteratively through a non-linear optimization which is defined on the basis of the maximum likelihood criterion.

Among several omnidirectional camera models (Kannala and Brandt, 2006; Mei and Rives, 2007; Scaramuzza et al., 2006a; Urban et al., 2015), this generic model has been chosen in this work to describe the relation between a scene point and its correspondence on the image plane. As mentioned, the advantage of this model is its flexibility with various types of central omnidirectional camera models through fitting a polynomial to the non-linear function to find the relationship between the image point and the corresponding point on the mirror or lens.

5.3 Methodology of the Spherical MSCKF

Recall that the MSCKF comprises three main parts: the system model, the measurement model and the update. In the system model, parameters in the state vector are propagated in accordance with the INS mechanism and augmented with camera poses once an image is captured. The measurement model in the MSCKF is defined on the image plane, which is not appropriate for omnidirectional images due to their high distortion. Therefore, a new measurement model must be defined. The following description of the OVIO details the formulation of the proposed measurement model. The new measurement model links the omnidirectional observations with the system model to update the state vector and its covariance matrix. An overview of the proposed procedure for integrating inertial measurements with omnidirectional image observations within a MSCKF framework is given in Algorithm 1.

Algorithm 1 The proposed OVIO based on the MSCKF

System model

For each INS output,

- Propagate the state vector and its uncertainty.

Measurement model

For each omnidirectional image:

- Add the camera pose to the state vector and augment the covariance matrix.
- Detect, match and track the projected features corresponding to the scene point on the image plane.
- Calculate the normalized points on the unit sphere as well as the corresponding normalized reprojected points.
- Reduce the points to the tangential plane to calculate the residual.
- Stack the 2D residuals.

Update

Given the measurements of the scene point and the corresponding residuals:

- Update the state vector and its uncertainty.

5.3.1 The Proposed Measurement Model

As noted, in the conventional MSCKF, which has been designed for only perspective cameras, the measurement model is described on the image plane (Mourikis, 2006). From the calibration section it can be seen that this plane is inappropriate for highly distorted images acquired by omnidirectional cameras. To deal with this distortion a measurement model on a plane around each measurement ray that is tangent to the unit sphere is defined. As seen in Figure 5.2, central omnidirectional cameras have a single viewpoint C_i, $i = 1 \cdots N_j$, that all rays pass through. N_j is the number of camera poses from which the scene point $\mathbf{X}^j$ $\mathbb{R}^3$ is observed. The normalized ray direction $\mathbf{x}_{ij}^s$ is the mapping of the scene point to the unit sphere. Collinear with this vector, there exists a 3D vector $\mathbf{p}''$, which is the map of the ray to the mirror in catadioptric cameras or the fish-eye lens in dioptric systems (see Figure 5.1). This vector can be written as equation (5.3) and function g is calculated via the calibration process described in the previous section.

Let us assume that the feature $\mathbf{X}^j$ is observed from a set of N_j camera poses C_i , $i \in \mathbb{S}_j$.

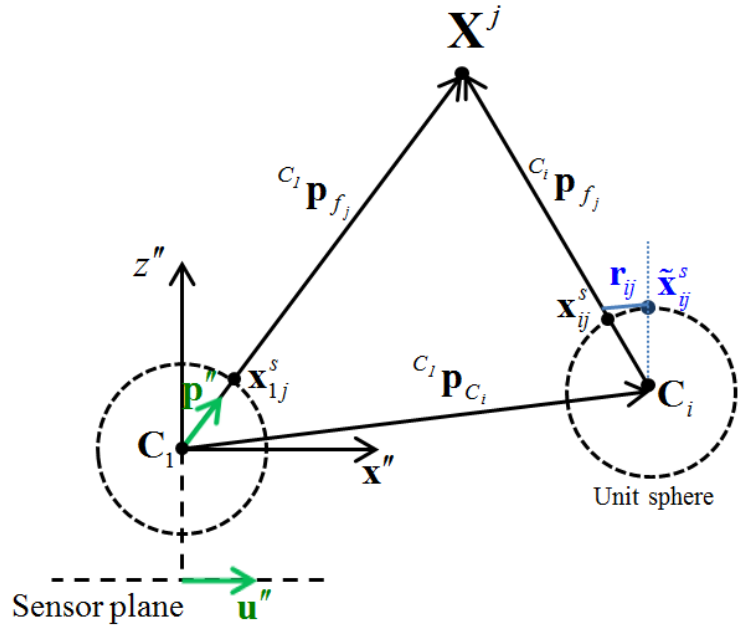


Figure 5.2: Different components of projection in a single view-point omnidirectional camera and the triangulation between view C_1 and C_i.

The normalized point $\mathbf{x}_{ij}^s$ is the unit spherical point of $\mathbf{X}^j$ with respect to view C_i and it is given by

$$\mathbf{x}_{ij}^s = \begin{bmatrix} x_{ij}^s \\ y_{ij}^s \\ z_{ij}^s \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{u}_{ij}'' \\ g(||\mathbf{u}_{ij}''||) \end{bmatrix} \right) \tag{5.5}$$

where $\mathcal{N}(.)$ maps points from the mirror or lens onto the unit sphere to spherically normalize them:

$$\mathbf{x}^s = \mathcal{N}(\mathbf{x}) = \frac{\mathbf{x}}{||\mathbf{x}||} \tag{5.6}$$

Given the covariance of a point on the sensor plane as $\mathbf{Q}_{\mathbf{u}''\mathbf{u}''} = \sigma^2 \mathbf{I}_2$, the covariance matrix of the spherically normalized point varies depending on the location of the point on the sensor plane. According to (Förstner, 2012):

$$\mathbf{Q}_{\mathbf{x}^s\mathbf{x}^s} = \mathbf{J}^s \begin{bmatrix} \mathbf{Q}_{\mathbf{u}''\mathbf{u}''} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 0 \end{bmatrix} \mathbf{J}^{s\mathrm{T}} \tag{5.7}$$

where matrix $\mathbf{J}^s$ is the Jacobian given by

$$\mathbf{J}^s = \frac{\partial \mathbf{x}^s}{\partial \mathbf{x}} = (\mathbf{I}_3 - \mathbf{x}^s \mathbf{x}^{s\mathrm{T}}) / ||\mathbf{x}|| \tag{5.8}$$

Using the normalized ray direction, the collinearity equations are now formulated as

$$\mathbf{x}_{ij}^s = \mathcal{N}([\mathbf{I}_3 | \mathbf{0}_{3,1}] \mathbf{M}_{\mathcal{C}_i}^{-1} \mathbf{X}^{sj}) \tag{5.9}$$

where $\mathbf{X}^s = [\mathbf{X}_0; X_h]$ is the homogeneous vector of the world point and the matrix $\mathbf{M}_{\mathcal{C}_i}$ is the camera motion which consists of the rotation matrix of the Earth frame with respect to the camera system $\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i}$ and the position of the projection center in frame $\{\mathcal{E}\}$, ${}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i}$:

$$\mathbf{M}_{\mathcal{C}_i} = \begin{bmatrix} \mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i\mathrm{T}} & {}^{\mathcal{E}}\mathbf{p}_{\mathcal{C}_i} \\ \mathbf{0}_{3,1}^{\mathrm{T}} & 1 \end{bmatrix} \tag{5.10}$$

Equation (5.9) consists of three collinearity equations per measured camera ray and four arrays for the feature point. Due to the redundancy in equation (5.9), the covariance matrices corresponding to the spherical measurements are singular. Therefore, before the spherical measurements can be used in the measurement model, they must be reduced to the minimum parameters using (Förstner, 2012):

$$\mathbf{x}_r = (null(\tilde{\mathbf{x}}^{s\mathrm{T}}))^{\mathrm{T}}\mathbf{x}^s = \begin{bmatrix} \mathbf{r}^{\mathrm{T}} \\ \mathbf{s}^{\mathrm{T}} \end{bmatrix} \mathbf{x}^s \tag{5.11}$$

where vectors $\mathbf{r}$ and $\mathbf{s}$ express the 2D tangential plane of the unit sphere evaluated at the reprojected point $\tilde{\mathbf{x}}^s$.

By mapping the omnidirectional measurements onto the tangential plane, as well as their corresponding reprojected points, and assuming that the noise has a zero-mean Gaussian distribution and is uncorrelated to the system model, the residuals can be determined and linearized around the reprojected points on the tangential plane. The reprojected points are obtained from their 3D estimated position with respect to frame $\{\mathcal{E}\}$. The linearization of the residuals can be written as

$$\begin{aligned} \mathbf{r}_{ij} = &\mathbf{J}_s^{\mathrm{T}}(\tilde{\mathbf{x}}_{ij}^s)\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i} \lfloor (\hat{\mathbf{X}}_0^j)\wedge \rfloor \delta\boldsymbol{\Phi}_{\mathcal{C}_i} \\ &-\mathbf{J}_s^{\mathrm{T}}(\tilde{\mathbf{x}}_{ij}^s)\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i} \hat{X}_h^j \, {}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_i} \\ &+\mathbf{J}_s^{\mathrm{T}}(\tilde{\mathbf{x}}_{ij}^s)[\mathbf{I}_3|\mathbf{0}_{3,1}]\mathbf{M}_{\mathcal{C}_i}^{-1} null(\hat{\mathbf{X}}^{sj\mathrm{T}})\delta\mathbf{X}_r^{sj} \end{aligned} \tag{5.12}$$

In the preceding equation, vectors $\delta\boldsymbol{\Phi}_{\mathcal{C}_i}$ and ${}^{\mathcal{E}}\delta\mathbf{p}_{\mathcal{C}_i}$ represent perturbations of the camera pose that are augmented in the MSCKF and the vector $\delta\mathbf{X}_r^{sj}$ is the error of the reduced coordinate of the spherical scene feature $\mathbf{X}^j$. The symbol $\lfloor (\hat{\mathbf{X}}_0^j)\wedge \rfloor$ is the skew-symmetric of $\hat{\mathbf{X}}_0^j$:

$$\lfloor (\hat{\mathbf{X}}_0^j)\wedge \rfloor = \begin{bmatrix} 0 & -\hat{Z}_0^j & \hat{Y}_0^j \\ \hat{Z}_0^j & 0 & -\hat{X}_0^j \\ -\hat{Y}_0^j & \hat{X}_0^j & 0 \end{bmatrix} \tag{5.13}$$

The Jacobian matrix $\mathbf{J}_s(\tilde{\mathbf{x}}_{ij}^s)$ is given by

$$\mathbf{J}_s(\tilde{\mathbf{x}}_{ij}^s) = \frac{1}{||\mathbf{x}||}(\mathbf{I}_3 - \frac{\mathbf{x}\mathbf{x}^\mathrm{T}}{\mathbf{x}^\mathrm{T}\mathbf{x}})null(\mathbf{x}^\mathrm{T})|_{\mathbf{x}=\tilde{\mathbf{x}}_{ij}^s} \tag{5.14}$$

Equation (5.12) can then be rewritten as:

$$\mathbf{r}_{ij} = \mathbf{H}_{\mathbf{X}i}^j \delta\mathbf{X} + \mathbf{H}_{f_i}^j \delta\mathbf{X}_r^{sj} + \mathbf{n}_{ij} \tag{5.15}$$

where $\delta\mathbf{X}$ is the error-state vector described in the system model and $\mathbf{n}_{ij}$ is the vector of measurement noise. The Jacobian matrices $\mathbf{H}_{\mathbf{X}_i}^j$ and $\mathbf{H}_{f_i}^j$ are given by the following equations:

$$\mathbf{H}_{\mathbf{X}_i}^j = \begin{bmatrix} \mathbf{0} & \mathbf{J}_s^\mathrm{T}(\tilde{\mathbf{x}}_{ij}^s)\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i}\lfloor(\hat{\mathbf{X}}_0^j)\wedge\rfloor & -\mathbf{J}_s^\mathrm{T}(\tilde{\mathbf{x}}_{ij}^s)\mathbf{C}_{\mathcal{E}}^{\mathcal{C}_i}\hat{X}_h^j & \mathbf{0} \end{bmatrix} \tag{5.16}$$

where the dimension of the left and right zeros are $2 \times (15 + 6(i-1))$ and $2 \times 6(M-i)$, with M the number of camera poses augmented in the state vector.

$$\mathbf{H}_{f_i}^j = \mathbf{J}_s^\mathrm{T}(\tilde{\mathbf{x}}_{ij}^s)[\mathbf{I}_3|\mathbf{0}_{3,1}]\mathbf{M}_{\mathcal{C}_i}^{-1} null(\hat{\mathbf{X}}^{sj\mathrm{T}}) \tag{5.17}$$

The residuals $\mathbf{r}_{ij}$ pertaining to the observations of the feature $\mathbf{X}^j$ for $i \in \mathbb{S}_j$ can now be stacked to form a single block vector:

$$\mathbf{r}_j = \mathbf{H}_{\mathbf{X}}^j \delta\mathbf{X} + \mathbf{H}_f^j \delta\mathbf{X}_r^{sj} + \mathbf{n}_j \tag{5.18}$$

where $\mathbf{H}_{\mathbf{X}}^j$ and $\mathbf{H}_f^j$ are the block matrices of $\mathbf{H}_{\mathbf{X}_i}^j$ and $\mathbf{H}_{f_i}^j$, respectively. The measurement noise covariance matrix $\mathbf{R}_j$ is a block diagonal matrix including the individual covariance matrices $\mathbf{R}_{ij}$, $i = 1 \cdots N_j$, given by (Förstner, 2012)

$$\mathbf{R}_{ij} = \mathbf{J}_s^\mathrm{T}(\tilde{\mathbf{x}}_{ij}^s) \begin{bmatrix} \mathbf{Q}_{\mathbf{u}''\mathbf{u}''} & \mathbf{0} \\ \mathbf{0}^\mathrm{T} & 0 \end{bmatrix} \mathbf{J}_s(\tilde{\mathbf{x}}_{ij}^s) \tag{5.19}$$

Finally, the standard measurement model that can be used in a Kalman filter update

is obtained by projecting equation (5.18) on to the left null space of the matrix $\mathbf{H}_f^j$. The reduced equation is

$$\mathbf{r}_{rj} = \mathbf{H}_{rj}\delta\mathbf{X} + \mathbf{n}_{rj} \tag{5.20}$$

where,

$$\begin{aligned} \mathbf{r}_{rj} &= (null(\mathbf{H}_f^{jT}))^{\mathrm{T}}\mathbf{r}_j \\ \mathbf{H}_{rj} &= (null(\mathbf{H}_f^{jT}))^{\mathrm{T}}\mathbf{H}_{\mathbf{X}^j} \\ \mathbf{n}_{rj} &= (null(\mathbf{H}_f^{jT}))^{\mathrm{T}}\mathbf{n}_j \end{aligned} \tag{5.21}$$

Equation (5.20) is a standard linear measurement model that is used in an EKF estimation to update the error-state $\delta\mathbf{X}$, as well as the covariance matrix.

5.4 Summary

In this chapter, a novel approach for the integration of inertial measurements with omnidirectional image features using the multi-state constraint Kalman filtering method has been proposed. Since the measurement model in the conventional MSCKF is only suitable for perspective cameras, a new measurement model on a plane tangent to the unit sphere around each ray has been defined. An advantage of this space is that the distortion of omnidirectional images can be well-defined, removing their effect from the image observations. To find the relationship between points on the unit sphere and their corresponding positions on the image plane, the calibration of these cameras is essential. To do so, a flexible and easy-to-use calibration method has been described in this chapter. The calibration process exploits a polynomial to find the interior parameters of any kind of omnidirectional camera. This work has employed this method to calibrate the camera used in the real experiments. The proposed method is implemented in Matlab and it has been evaluated with various real and simulated datasets.

The presented approach enables motion estimation in 6 DoF. From this perspective,

it is the first time that omnidirectional images have been integrated with inertial measurements in a real-time pose estimation framework. Employing the MSCKF allows the maximum potential that a visible feature creates to be exploited. In addition, utilizing an omnidirectional camera, it is more likely that the greater number of features surrounding the camera will contribute to the motion estimation, resulting in more accurate localization.

Chapter 6
Experiments and Results

6.1 Introduction

THE algorithms described in Chapter 4 and Chapter 5 were tested in a series of experiments. To evaluate the stereo VIO, several KITTI datasets were used. These datasets were collected in urban environments. For evaluation of the OVIO approach, experiments with synthetic data in both indoor and outdoor environments were conducted along with a large-scale real-world experiment. The performance of both approaches as compared with the perspective VIO was also assessed in terms of the accuracy of the reconstructed trajectories. In this chapter, the term "perspective VIO" refers to MSCKF-VIO using a perspective camera and "spherical VIO" refers to spherical MSCKF using an omnidirectional camera named OVIO in this thesis. In all experiments, accuracy was measured by computing the Root Mean Squared Error (RMSE), which represents the difference between the estimated pose and the ground truth. The results presented in this chapter have been published in (Ramezani and Khoshelham, 2018) and (Ramezani et al., 2018).

6.2 Stereo Visual-Inertial Odometry

To evaluate the performance of the VIO approach in the mono and stereo mode, an experiment was conducted using KITTI datasets 0001, 0005, and 0020 (Geiger et al., 2012). Datasets 0001 and 0005 contain 108 and 154 pairs of images, respectively, both captured

on 26 September 2011. Dataset 0020 consists of 330 pairs of images recorded on 30 September 2011. Sample frames from the datasets are shown in Figure 6.1. The images in all datasets have the same size of 1392×512, and were recorded at approximately 0.1 second intervals. The baseline between the two cameras in the static stereo is 54 cm which is sufficient for initial 3D point estimation. Moreover, the datasets include inertial measurements recorded by an OXTS RT3003 inertial and GPS navigation system at a measurement rate of 100 Hz. The RTK GNSS measurements serve as the ground truth in these datasets. It is worth noting that the INS used by KITTI is a high quality inertial sensor and its inertial data has been purified regarding gravity, biases and integration error. Therefore, it is expected that the INS only solution will provide a good motion estimate, competing with the visual-inertial integration.

The entire algorithm was implemented in MATLAB R2015a on an ordinary PC with a 3.4 GHz Intel(R) Core(TM)i7 processor and 8 GB RAM.

6.2.1 Feature Detection, Matching and Tracking

As already mentioned, to extract and match features, the Speeded Up Robust Features (SURF) detector (Bay et al., 2008) was used. In stereo vision, features were matched by applying the approximate matching method described in (Lowe, 2004). Having matched the features across the static stereo, the corresponding pairs of features along the image sequence were tracked by employing the Kanade-Lucas-Tomasi (KLT) tracker (Tomasi and Kanade, 1991). In the mono mode, features were tracked without a cross matching step. Despite the fact that SURF is not the best choice for feature tracking in terms of real-time operation, a comparison of various feature detectors indicated that this detector provided the most robust and well-distributed features. The outliers were discarded by using the M-estimator Sample Consensus (MSAC) approach (Torr and Zisserman, 1998) with a 2-point similarity transform. Figure 6.2 shows an example of tracked features indicated by green plus signs in the left camera, and red circles in the right camera.

Visual examination of feature extraction results showed that the majority of SURF

(a)

(b)

(c)

Figure 6.1: Sample frames from the three tested KITTI datasets. (a) dataset 0001. (b) dataset 0005. (c) dataset 0020.

keypoints were found at the center of the images where there was less motion blur. Also fewer SURF keypoints were found in the stereo mode as some keypoints could not be matched across the stereo image pair.

The number of tracked features varied between the mono and stereo mode. Figure 6.3 shows a boxplot of the number of tracked features in each frame over the three datasets 0001, 0005, and 0020 for both mono and stereo modes. Box plots are a useful statistical tools to depict the distribution of data based on 5 parameters: minimum, 25th percentile, median, 75th percentile and maximum. Pluses, in this Figure, are also outliers.The lower number of features in stereo mode is due to the fact that not all SURF features can be

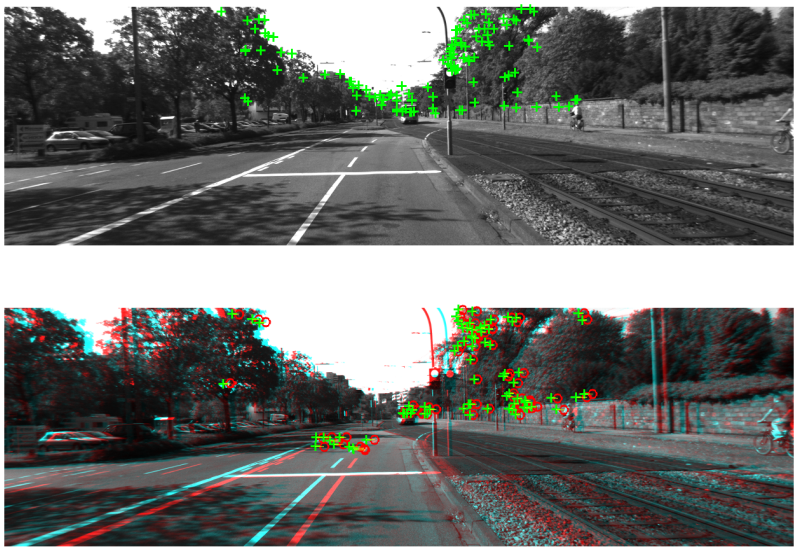


Figure 6.2: Features tracked in mono vision (**Top**) and stereo vision (**Bottom**) from KITTI dataset 0001.

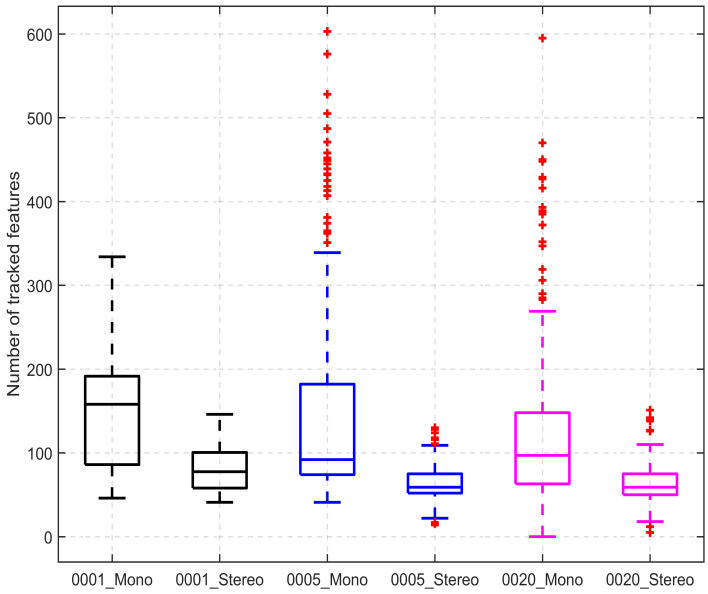


Figure 6.3: Statistics on the number of tracked features in each frame over the three datasets 0001, 0005, and 0020 for both mono and stereo modes.

Table 6.1: Comparison of Tracking Performance in Mono and Stereo on KITTI Traverses 0001, 0005 and 0020.

	Mono			Stereo		
KITTI Traverse	0001	0005	0020	0001	0005	0020
Total Tracked Features	750	1975	3500	300	950	1500
Processing Time(sec)	15	23	95	30	46	115

matched across the stereo pair. However, as can be seen, in the stereo mode the variation in the number of tracked features is significantly less than that in the mono mode, which indicates that the tracking performs more consistently along the image sequence.

Table 6.1 summarizes the average number of all tracked features along the sequence of images in the mono and stereo modes on the datasets employed, as well as their processing time averaged over 10 runs of the program. It is clear that cross matching rejects approximately half of the features tracked along the image sequence. However, the overall time of feature tracking in stereo mode is almost two times more than that in mono mode.

6.2.2 Accuracy of Stereo Visual-Inertial MSCKF

To assess the accuracy of the estimated trajectories, the translational RMSE, defined as the Root Mean Square of the three translation errors at each timestep, was used. The translational error at each timestep was calculated from the estimated position and the ground truth. The total translational error was measured as the Euclidean norm of all RMSEs, whereas ARMSE, noted in Table 6.2, was calculated as the average of RMSEs. Since GNSS does not provide the attitude of the vehicle, there was no ground truth available for determination of rotational accuracy. For this reason, rotational accuracy of the stereo VIO has not been considered in the evaluation.

Tracking salient features, the rectified image observations, i.e. features corrected with

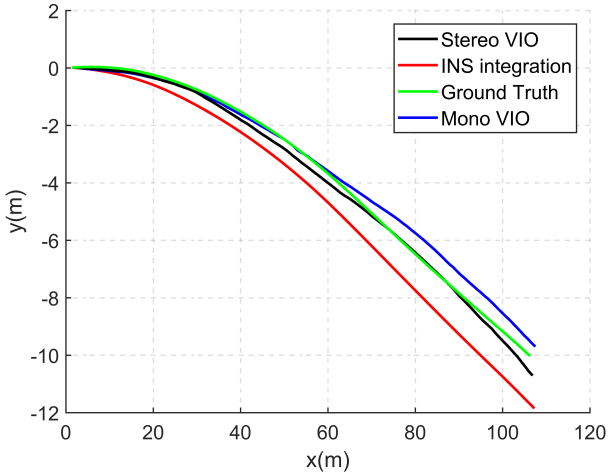
respect to the intrinsic parameters of the camera, were taken as input to estimate the 3D position using Gauss-Newton least-squares minimization (Montiel et al., 2006). In contrast to the mono VIO in which scene features are reconstructed based on observations over a sequence of images, the observations from both the static stereo and multiple images over time were employed. This way, more 3D point estimates achieve sufficient precision to contribute to the EKF update step. In addition, in order to have the best initialization, the initial 3D position was calculated via triangulation between the first and the last frame in which the feature was visible. Figure 6.4 shows the trajectories obtained by the mono and the stereo VIO compared with the INS-only integration, as well as the ground truth. While VIO in both modes performed much better than inertial positioning, the stereo VIO provided a more accurate trajectory than the mono VIO.

Figure 6.5 shows the cumulative distribution of translational errors for the proposed stereo method as compared with the mono and INS-only integration. The performances brought here are for the optimum camera pose window. Stereo VIO performed better than the other methods and reached an average accuracy of 0.9% of the trajectory length compared to 1.3% and 2.5% for the mono and the INS only integration, respectively. Note that the drift of INS-only integration is quite low because the KITTI inertial data was obtained using a precise INS (OCTS RT3003), and the data was filtered for biases and integration errors. Using a low-grade INS, it is anticipated that INS-only integration would have resulted in a larger drift, therefore the benefit of VIO would be more apparent.

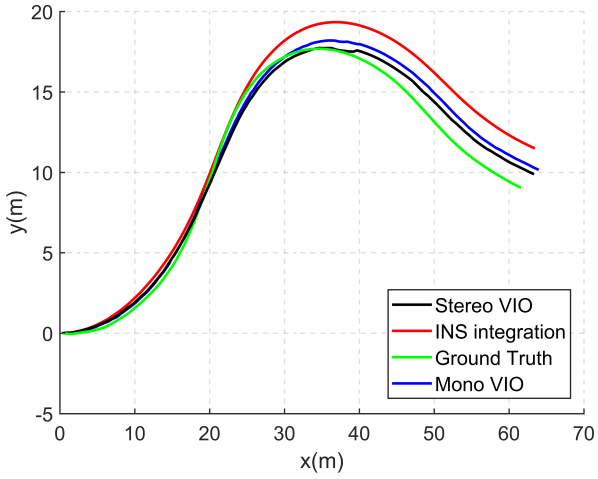
Table 6.2 summarizes the performance of stereo VIO compared with mono and INS on the basis of Translational Average Root Mean Squared Error (ARMSE) and RMSE values over the entire sequence of frames for the tested KITTI datasets.

6.3 Omnidirectional Visual-Inertial Odometry

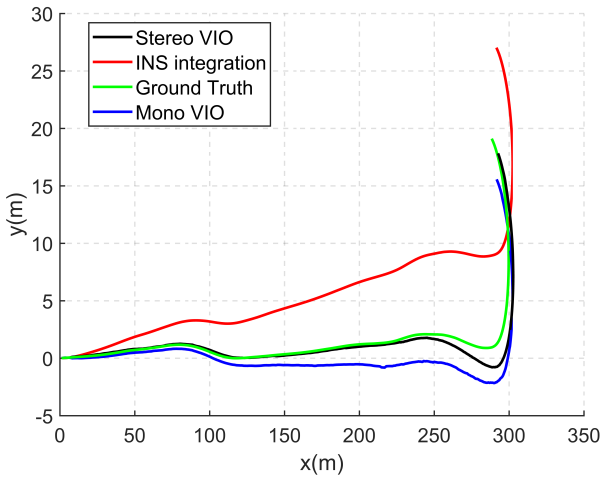
As noted earlier, for the evaluation of the new OVIO approach, both synthetic and real datasets were employed. The purpose of experiments with synthetic images is that exact



(a)

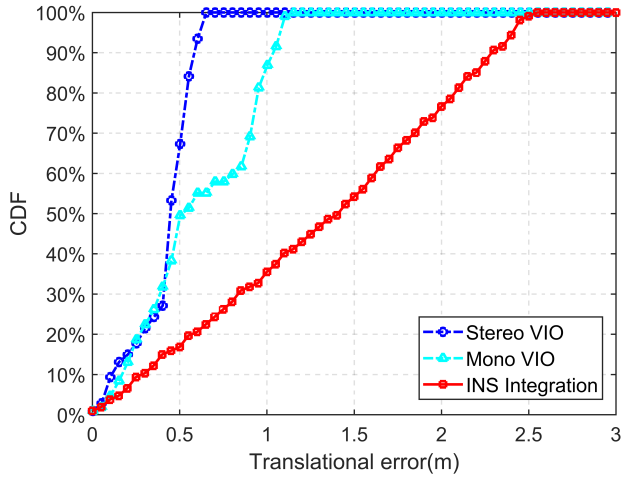


(b)

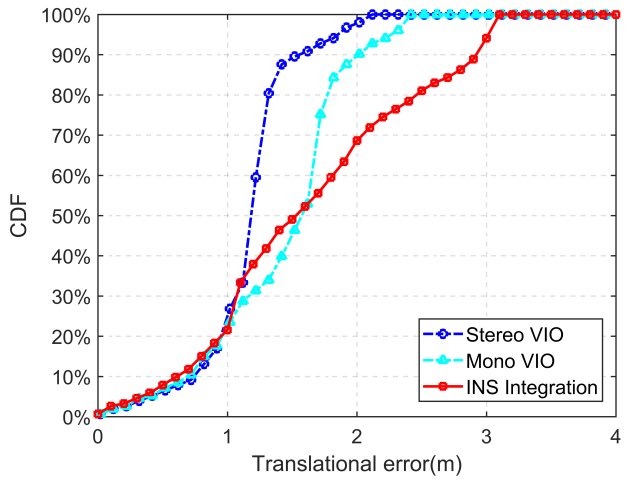


(c)

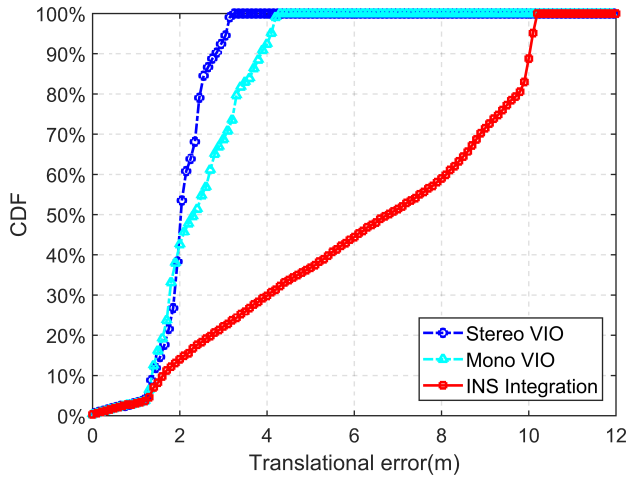
Figure 6.4: Trajectories obtained by the stereo VIO as compared with the mono and INS integration for the three KITTI traverses. (a) dataset 0001. (b) dataset 0005. (c)dataset 0020.



(a)



(b)



(c)

Figure 6.5: Cumulative distribution of errors for the stereo method compared with mono and INS integration for the tested KITTI datasets 0001 (a), 0005 (b) and 0020 (c).

Table 6.2: Comparison of Average Root Mean Square Error (ARMSE) of INS Integration, Mono VIO, and Stereo VIO on KITTI Traverses 0001, 0005 and 0020.

KITTI Traverse		0001	0005	0020
INS Integration	Trans. ARMSE(m)	0.78	0.97	3.65
	Total Trans. Error(m)	2.53	3.11	9.90
Mono VIO	Trans. ARMSE(m)	0.36	0.85	1.44
	Total Trans. Error(m)	1.13	2.44	2.76
Stereo VIO	Trans. ARMSE(m)	0.25	0.70	1.22
	Total Trans. Error(m)	0.66	2.14	2.25

ground truth can be obtained to evaluate the performance of the OVIO. The process of feature extraction, matching and tracking is similar to that described in the previous section, hence it is not explained here. To estimate the global position of features, a method similar to that for stereo-VIO is utilized for the OVIO, with a slight difference in the image observation which is instead of taking image points on the image plane, points on the the unit sphere are considered. Because omnidirectional images have high distortion, calibration of these cameras is essential. Therefore, this section begins with the calibration results of the fisheye camera used in the conducted experiments.

6.3.1 Camera Calibration

As noted in Section 5.2, a pre-calibration of the used fisheye camera was conducted using OcamCalib Toolbox (Scaramuzza et al., 2006b). This Toolbox employs a polynomial model to calculate the calibration parameters of any omnidirectional camera that has a single effective point of view. In this experiment, a checker-board pattern was used and several video streams were captured while the camera was fixed and the checker-board was moving and facing toward the camera. The settings of the camera, including image resolution, exposure time and focal length were selected so as to be similar to the moment of the real experiment, since a change in camera parameters would necessi-

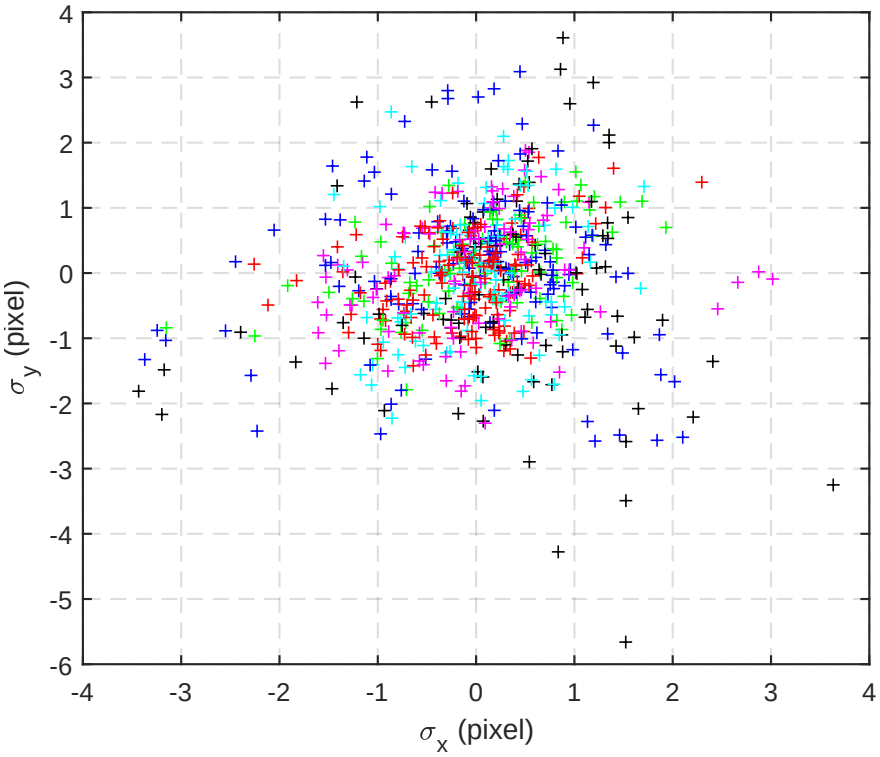


Figure 6.6: Distribution of reprojection errors.

tate camera recalibration. Having taken the video streams, several frames covering the checker-board from various angles, with different orientations were extracted and used in the calibration process. As recommended in (Scaramuzza, 2008b), the adopted order of the polynomial was 4 so as to attain a minimum value for the reprojection error. Figure 6.6 shows the distribution of reprojection errors, indicated with different colors for different images.

Table 6.3 summarizes the RMSE computed between the projected corners of the checker-board and their corresponding reprojected points as well as their standard deviations.The average error, which is the average of RMSEs, was 1.11 pixels, verifying the success of the calibration process. Figure 6.4 shows several sample images used in the calibration. Projected points are marked with plus symbols whereas their corresponding reprojected points are indicated by circles. The white circle shows the center of the image after calibration.

Following the calibration experiment, a test was conducted to assure that residuals described in equation (5.12) follows Gaussian distribution, since it is a fundamental assumption in the use of a Kalman filtering method (Groves, 2013). In this experiment, a check-board was set to be fixed in front of the camera and the camera was moving while facing toward the check-board. Observations were bundles of rays and the calibration parameters, estimated from the previous experiment, were employed to calculate the coordinates of image points on the unite sphere. The exterior orientation and the residuals were estimated through a bundle adjustment. Figure 6.8 shows the camera position and its orientation indicated by three coloured axes representing the x,y and z axes of the fisheye camera.

In Figure 6.9 (**Left**), the histogram of all 6000 residuals is plotted and a normal density function, green curve, fitted on the histogram is shown. This Figure verifies that residuals followed a Gaussian distribution and because their expectation is around zero, therefore they can be considered as unbiased. In addition, Figure 6.9 (**right**) displays a normal probability of all residuals along a reference line. This figure, also, confirms the normality of residuals since the majority appear along the reference line and their distribution is symmetric around this line.

6.3.2 Experiments with Synthetic Data

Indoor Synthetic Dataset

To compare the performance of the proposed OVIO with VIO in an indoor environment, an experiment was carried out in a synthetic corridor through the use of Blender [1], which is a professional open-source 3D content creation suite. Photo-realistic synthetic images were generated along a ground truth trajectory for both perspective and fish-eye cameras. To render the perspective images, the built-in perspective module in Blender was used, and for the omnidirectional images an open-source patch (Zhang et al., 2016), which is

[1] http://www.blender.org/

Table 6.3: RMSE of reprojection and the corresponding standard deviations.

Image	RMSE (pixel)	σ (pixel)
1	0.53	0.27
2	0.77	0.35
3	0.81	0.40
4	1.04	0.59
5	1.10	0.57
6	1.28	0.64
7	1.00	0.48
8	1.08	0.59
9	1.10	0.63
10	2.26	1.17
11	0.96	0.63
12	1.02	0.48
13	2.19	0.73
14	0.55	0.30

based on the Scarramuzza's model (Scaramuzza et al., 2006a), was applied. To stimulate further research in vision-based indoor positioning, the authors publicly released this dataset, which consists of synthetic images rendered with the perspective and fish-eye cameras [2]. The images were recorded at 30 Hz with the resolution of 640×480 pixels, looking in the forward direction. The FoV considered for the perspective images is 90° and for the fish-eye images 180°. Figure 6.10 shows sample images of this dataset and Figure 6.11 shows an example of feature matching results in both the perspective and fish-eye views.

Ground truth comprising 3D position and 4D quaternion data resampled at 300 Hz was used to generate the INS measurements. A constant error and normalized random noise to simulate the static and dynamic bias of the accelerometers and gyroscopes were added to the measurements.

A top view of the estimated trajectories compared to the ground truth is shown in Figure 6.12.

It can bee seen that the trajectories estimated by the OVIO and the perspective VIO closely follow the ground truth whereas the INS trajectory exhibits a significant drift. Fig-

[2] http://www.researchgate.net/profile/Milad_Ramezani/contributions

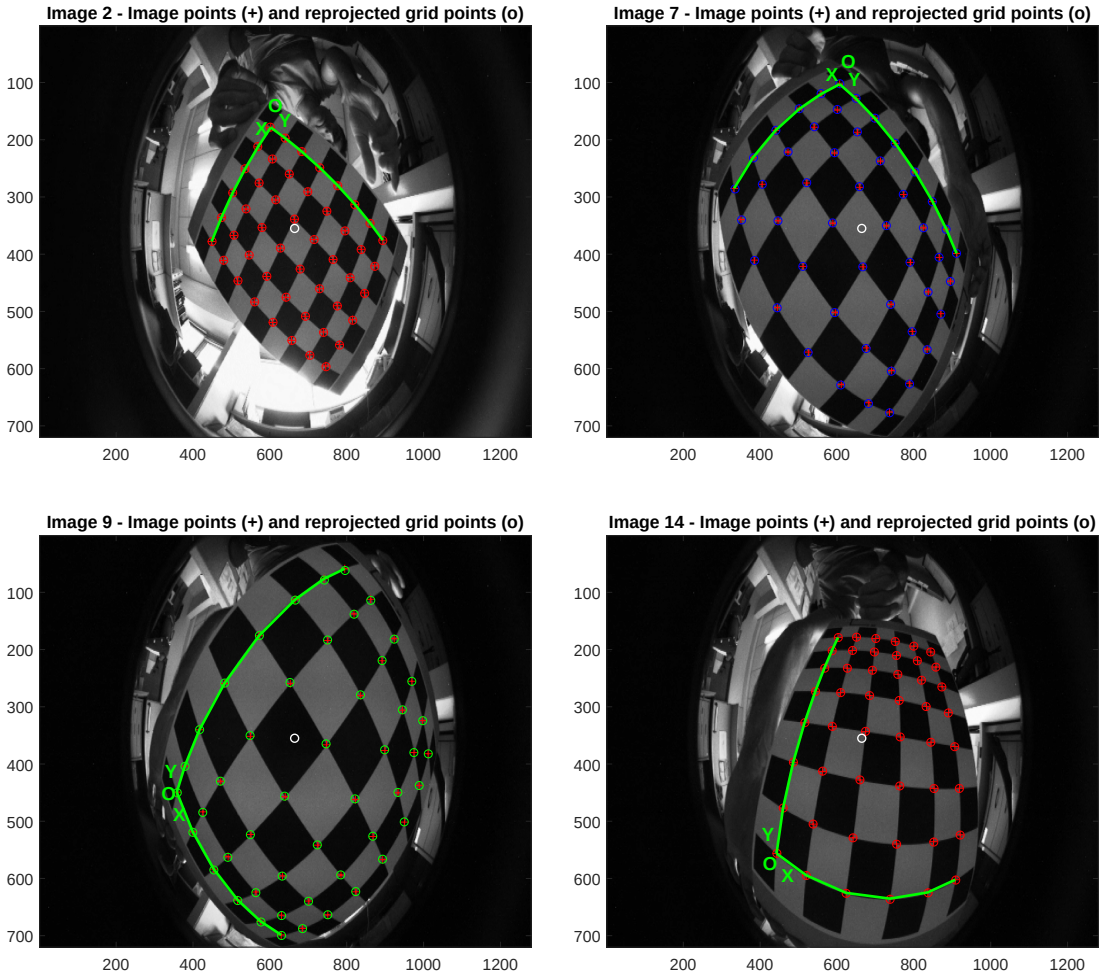


Figure 6.7: Sample images used in the calibration including the projected points and their corresponding reprojected points.

ure 6.13 shows the cumulative distribution of the translation and rotation errors for the perspective and the omnidirectional VIO, as well as for the INS-only integration. Clearly, the OVIO outperforms the perspective VIO and results in more accurate motion estimation. The difference in the performance of VIO and OVIO is more significant in terms of rotational accuracy, where OVIO gives rotation estimates that are 3 to 4 times more accurate than those obtained by VIO. The total length of the trajectory that the camera travelled was about 47 meters. Comparing the length of the trajectory and the total error for translation, the drifts for the INS-only integration, perspective and omnidirectional

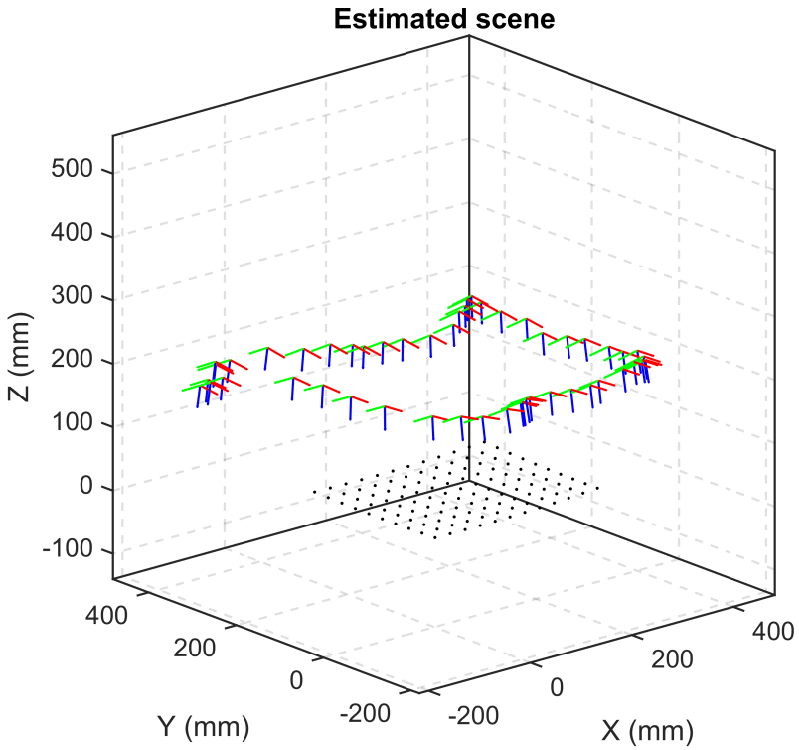


Figure 6.8: Estimated scene including checker-board corners (black dots) and the camera pose (coloured axes).

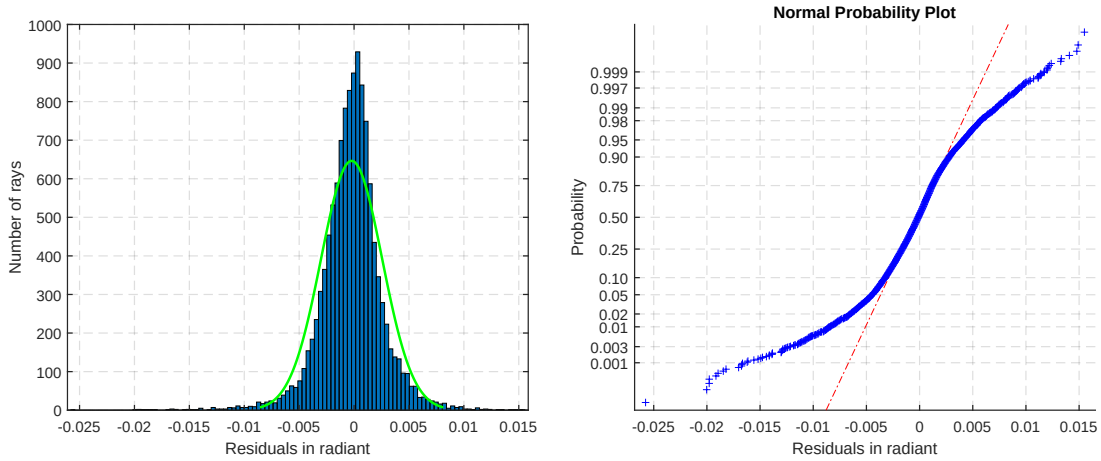


Figure 6.9: **Left**: Histogram of estimated residuals. **Right**: Normal probability plot of estimated residuals

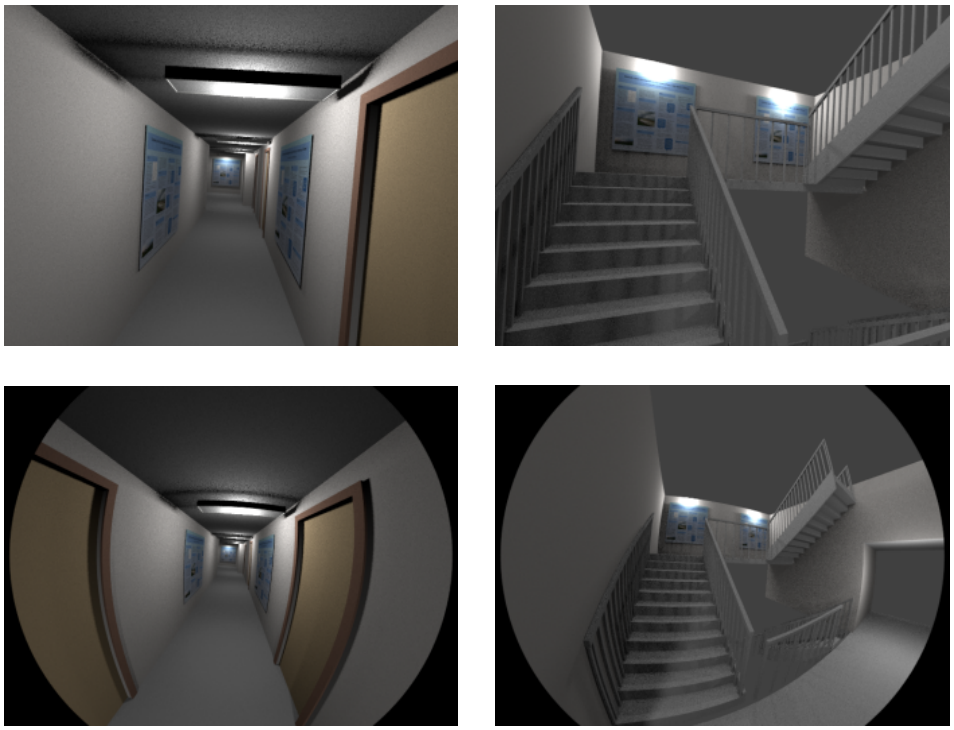


Figure 6.10: Example images from the sequence recorded using a simulated perspective camera (**top**) and a simulated fish-eye camera (**bottom**).

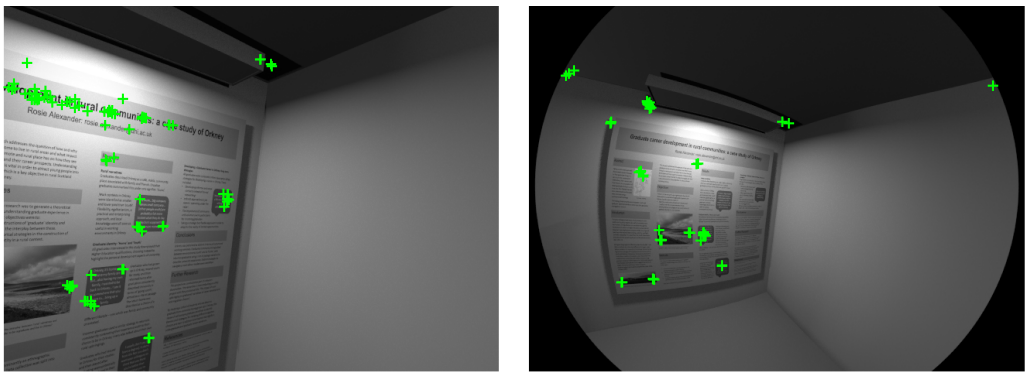


Figure 6.11: Features after the tracking process in a perspective view (**left**) and fisheye view (**right**). The green pluses show the features tracked along the image sequence.

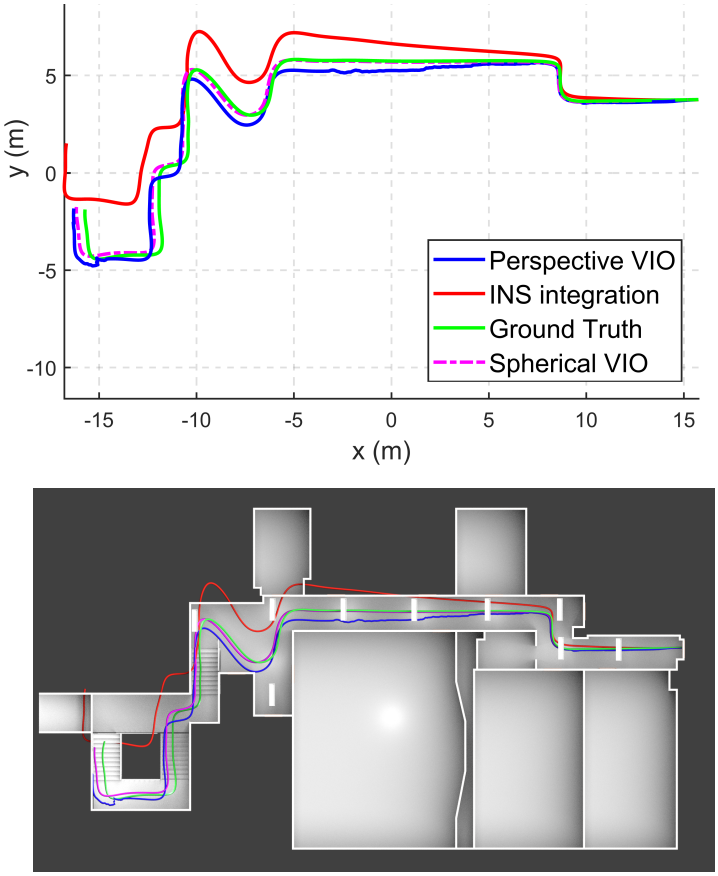


Figure 6.12: Top view of the estimated trajectory for perspective VIO (blue), and omni-directional VIO (magenta) compared with the INS-only integration (red) and the ground truth (green).

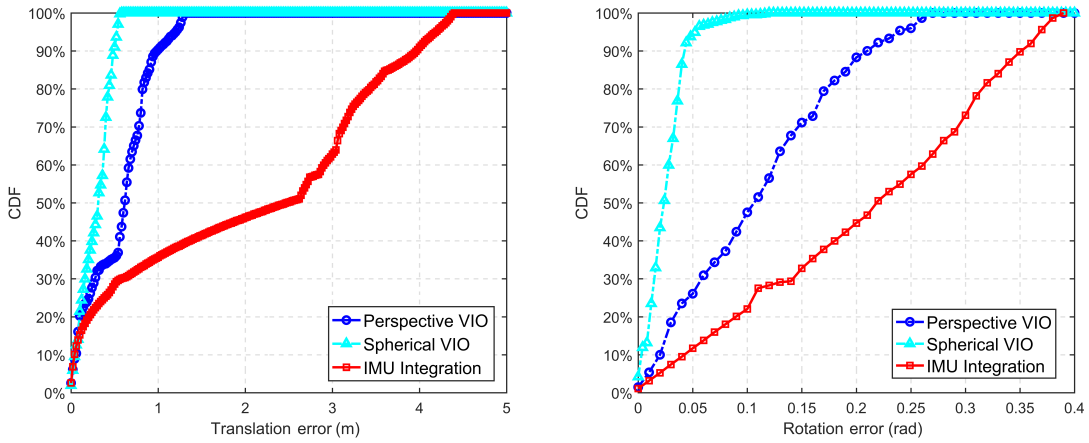


Figure 6.13: **Left**: Cumulative distribution of translational error for the perspective VIO, the omnidirectional VIO and the INS-only integration. **Right**: Cumulative distribution of rotational error for the perspective VIO, the omnidirectional VIO and the INS-only integration.

VIO are 9%, 1.3% and 1.1%, respectively.

Outdoor Synthetic Dataset

To examine the effect of larger FoVs on the performance of VIO in an outdoor environment, photorealistic synthetic images of the urban canyon dataset generated by Zhang et al. (2016) were used. Three sequences of 1500 images captured by the simulated perspective, fisheye and catadioptric cameras, along a trajectory of approximately 295 m, were selected. Because the catadioptric images were rendered in an upward-looking orientation, compared to the perspective and fisheye images rendered in a front-looking orientation, it is possible to observe the effect of different camera orientations. However, this synthetic dataset did not include INS measurements. Therefore, INS readings were generated from the ground truth trajectory resampled at 300 Hz in a similar fashion to that described in the previous section. Figure 6.14 shows sample images of the urban canyon dataset.

Figure 6.15 (left) shows boxplots of the number of tracked features in each frame for the perspective, fisheye and catadioptric views. The minimum number of tracked

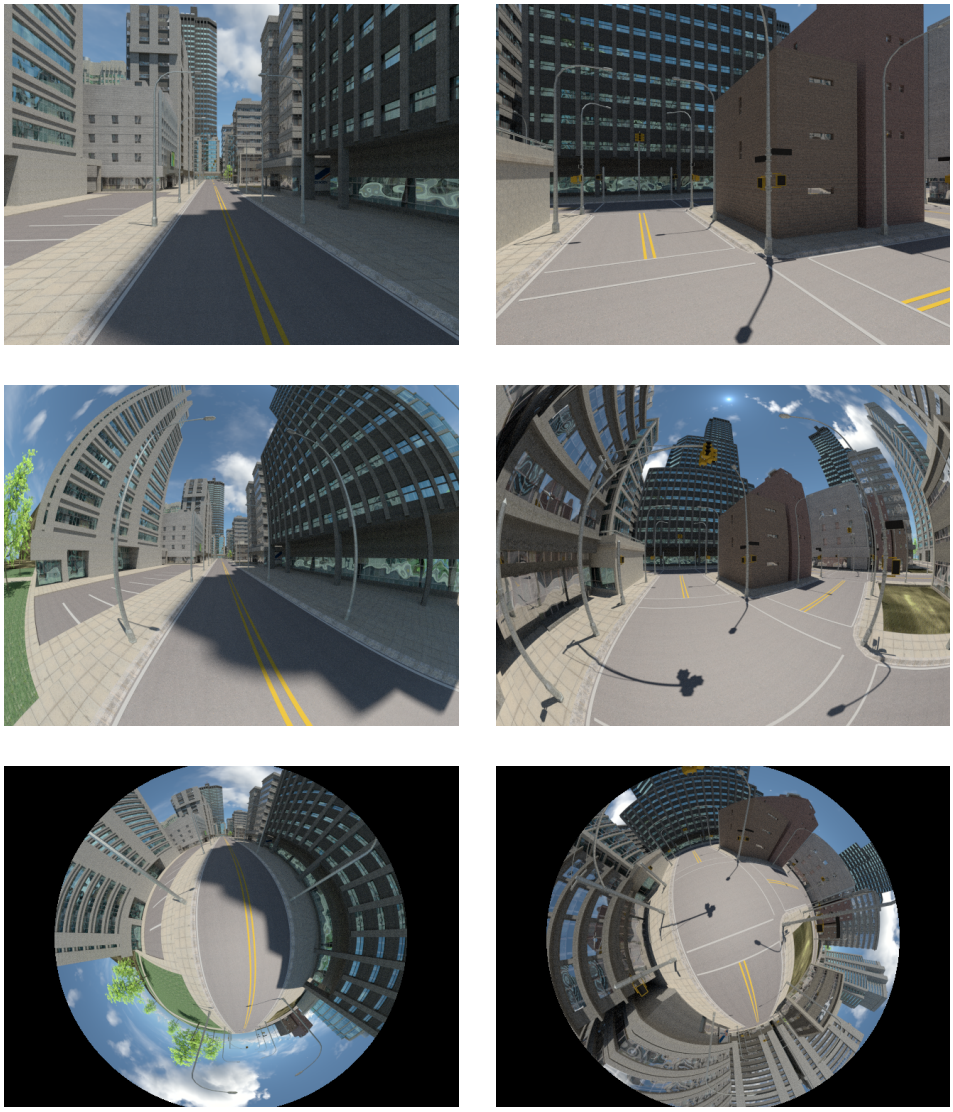


Figure 6.14: Sample images of the urban canyon dataset simulated by Zhang et al. (2016) in different FoV cameras namely perspective (**top**), fish-eye (**middle**) and catadioptric (**bottom**).

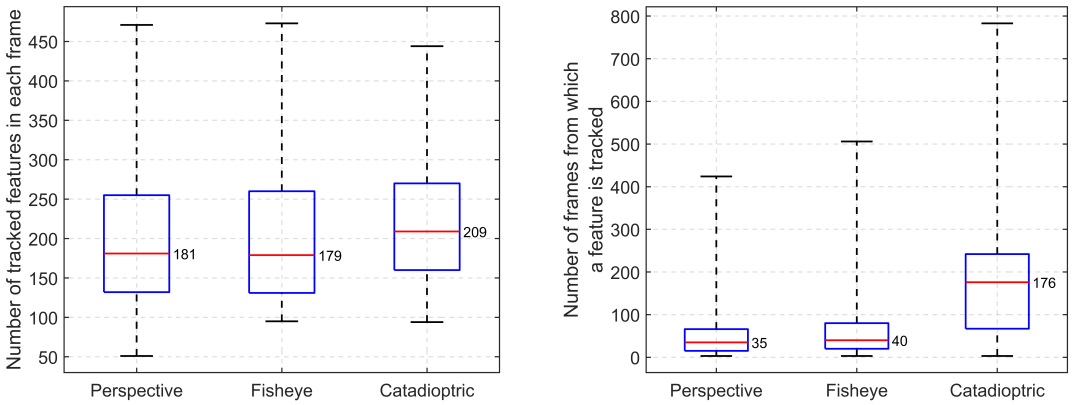


Figure 6.15: **Left**: Boxplot of the number of tracked features in each frame for perspective, fisheye and catadioptric views. **Right**: Boxplot of the number of frames from which each feature is observed.

features for the fisheye and the catadioptric cases is the same, but is twice as large as the minimum number of tracked features in the perspective view. However, the median values show no significant increase in the number of tracked features in each frame for the greater FoVs indicating that larger FoV does not necessarily provoke weaker features.

Figure 6.15 (right) shows boxplots of the number of frames from which each feature is observed. Here the catadioptric camera with an FoV of 360° presents a much longer feature tracking window as compared to the perspective and fisheye cameras. This can also be attributed to the upward-looking orientation of the catadioptric camera resulting in features being visible over longer sequences.

The 3D trajectories estimated from different data integrations are shown in Figure 6.16. As can be observed, while the drift of the INS-only integration is significant, the reconstructed trajectories of VIO and OVIO closely follow the ground truth. In the following, only the performance of the proposed OVIO with the perspective VIO is compared.

By computing the cumulative error distribution curves for rotational and translational errors measured by RMSE at each time-step, as shown in Figure 6.17, it can be seen that OVIO yields more accurate position and rotation estimates than perspective VIO. In particular, a comparison of the translational errors of the two methods reveals that

the spherical VIO results in a maximum error that is only half that generated by the perspective VIO. Also, from the rotational graph it can be seen that the spherical VIO results in a larger percentage of smaller errors compared to the perspective VIO. Comparing the translation errors for fisheye and catadioptric, there is no meaningful improvement when a FoV larger than 180° is employed. This can be because of the lower angular resolution that catadioptric cameras have, which can affect the translation accuracy. However, the result of rotation errors shows an improvement in the attitude estimation when a catadioptric camera is employed. This improvement is likely the effect of feature tracking over a 360° FoV resulting in more accurate estimation of attitude.

Table 6.4 summarizes the performance of the spherical and perspective VIO as measured by the Total RMSE, defined as the root mean square of RMSEs, and Average RMSE, defined as the average of RMSEs along the trajectory. It can be seen that the spherical VIO outperforms the perspective VIO in terms of both Average and Total RMSE. Considering the length of the trajectory and the total error for translation, the OVIO using the fisheye images achieved a translational accuracy of 0.78% of the trajectory length and 0.84% for

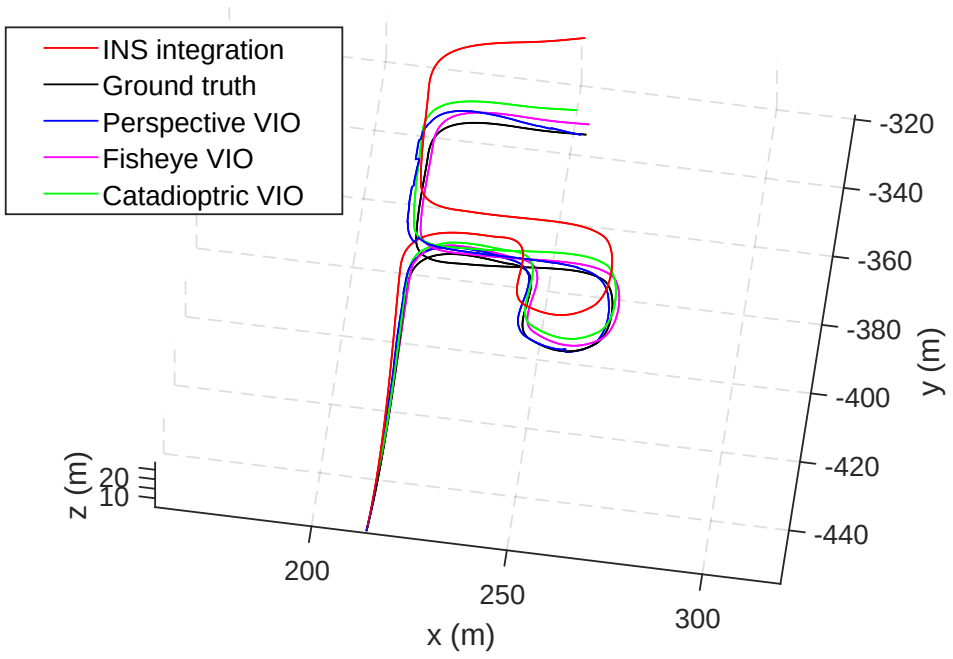


Figure 6.16: 3D estimated trajectories from different integration in the global frame.

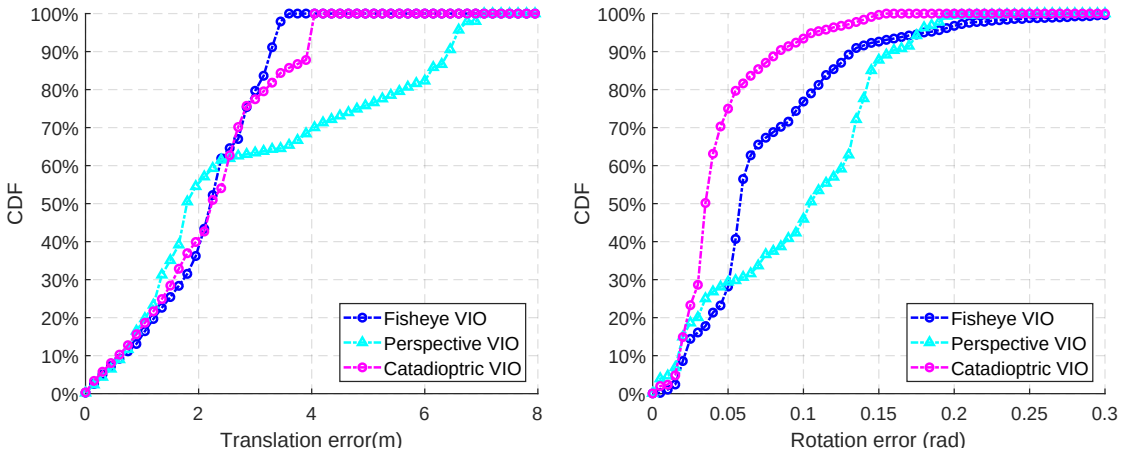


Figure 6.17: **Left**: Cumulative distribution of the translational errors for the urban canyon synthetic dataset. **Right**: Cumulative distribution of the rotational errors on the same trajectory.

Table 6.4: Comparison of Average and Total RMSE of the proposed spherical VIO on fisheye and catadioptric VIOs against perspective VIO.

	Fisheye	Catadioptric	Perspective
Average Translational RMSE (m)	1.31	2.1	2.86
Average Rotational RMSE (rad)	0.073	0.044	0.095
Total Translational RMSE (m)	2.3	2.5	3.5
Total Rotational RMSE (rad)	0.08	0.052	0.11

the catadioptric, compared to 1.2% for the perspective. This indicates that through use of the proposed method a positional accuracy of better than 1 meter is achieved along a 100 meter trajectory. In comparison with similar translational errors submitted on the KITTI website [3], the achieved accuracy by the OVIO is among the top four reported approaches, which is a significant result.

[3] http://www.cvlibs.net/datasets/kitti/eval_odometry.php

6.3.3 Real-World Experiments

To test the OVIO algorithm in a real environment, a sensor board developed for the research and shown in Figure 6.18 was strapped onto the top of a car and driven through a built up area near the Parkville campus of the University of Melbourne. The sensor board included a usb3 fish-eye camera providing 185° FoV, installed to be upward-looking. This way, the negative effect of mobile objects (e.g. vehicles, pedestrians), whose presence is inevitable in urban environments, was removed. Moreover, tracking features from both sides of the motion trajectory could provide better geometry and thus more accurate position and rotation estimates. It is also worth noting that an upward-looking camera ensures sufficient difference between the motion heading and the camera optical axis, resulting in an imaging geometry which is more suitable for vision-based positioning (Montiel et al., 2006). However, using a camera facing upwards can decrease the number of tracked features in areas with low-level buildings, even though the camera is equipped with a fish-eye lens.

A sample image from the real dataset, along with an image showing the tracked features from both sides of the road between two consecutive images, are shown in Figure 6.19. The resolution of the images is 1280×720 pixels, recorded at about 20 Hz. In total, 5260 sequential images were captured along a 1.3 km-long trajectory. For feature tracking, the same algorithm as employed in the synthetic environments was again used. On average, features were tracked across about 30 sequential images. The maximum number of camera poses in the state vector was thus set as 30. This value is appropriate for employment of constraints between the multitude of camera poses in the state vector of the MSCKF.

The inertial measurements, consisting of accelerations and angular velocity in 3 dimensions, were obtained at a frequency of 100 Hz by utilizing MMQG-INS [4]. GNSS observations were collected and used for the initialization of the INS. However, raw inertial

[4] http://www.unmannedsystemstechnology.com/company/systron-donner-inertial/mmqg-ins-gps/, Accessed: 2017-02-20

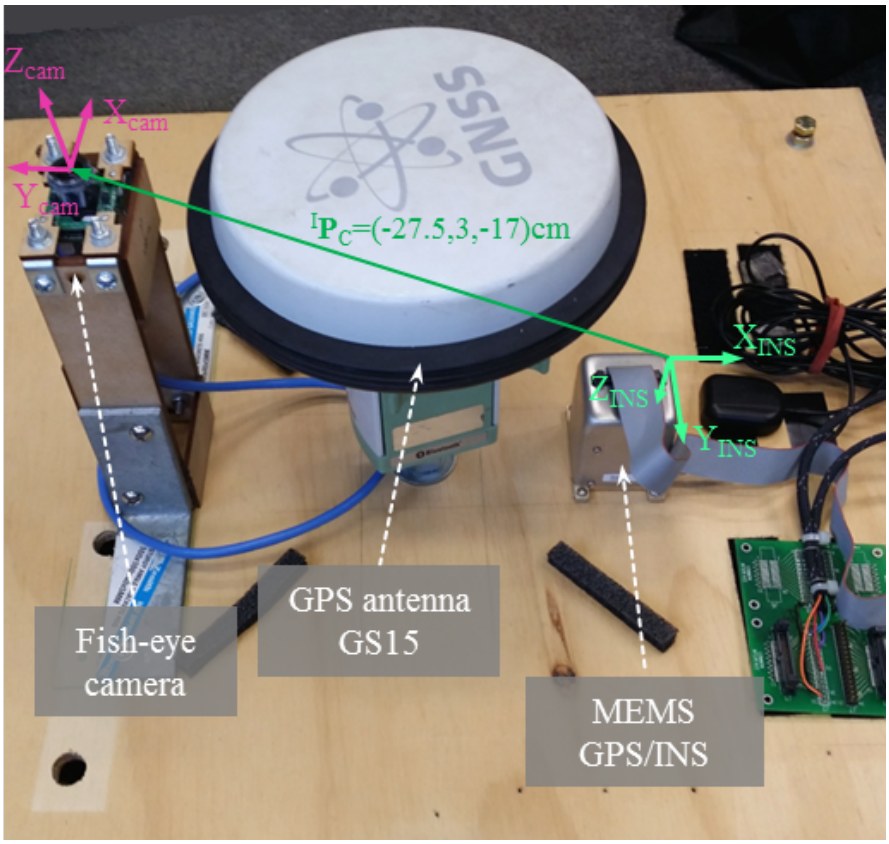


Figure 6.18: The sensors used in the real experiment, these being installed on a board that was attached to the roof of a car using rubber magnets.

outputs were stored without being corrected by GNSS. In order to synchronize the inertial readings with the images, the local time-stamp of the images was recorded and precisely converted to GPS time. The difference between Coordinated Universal Time (UTC) and local time, as well as the leap seconds value, which was 17 on the date of the experiment, were accounted for. This allowed a comparison of the absolute accuracy of the vehicle motion estimated by the proposed OVIO against the ground truth obtained by the surveying grade GNSS antenna, GS 15 (See in Figure 6.18). The frequency of the ground truth was 1 Hz, resulting in 242 points with centimetre level accuracy for the evaluation.

In Figure 6.20, top left, a top view of the estimated trajectory overlaid on Google Earth is shown. The trajectory estimated by the OVIO and the ground truth for the area

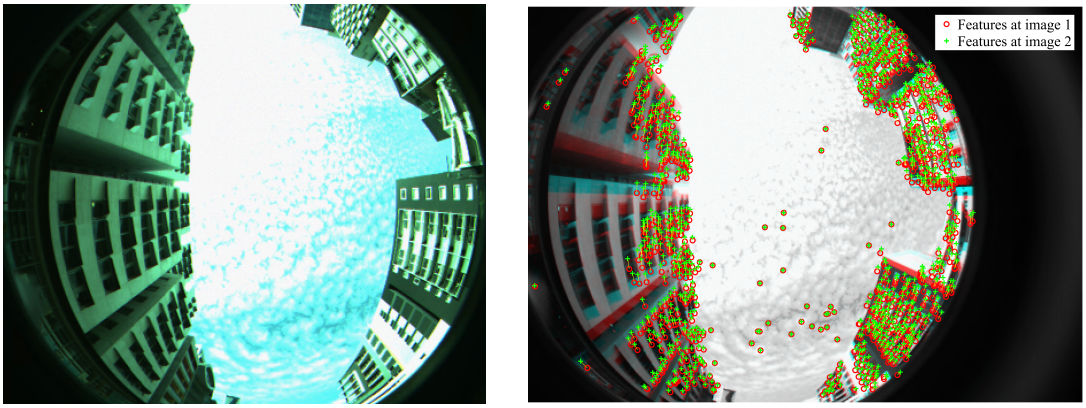


Figure 6.19: **Left**: A raw image captured by the fish-eye camera looking upward. **Right**: Features tracked between two consecutive images.

where precise GNSS was available is also shown in Figure 6.20, top right. The total RMSE over this part of the trajectory, which was 600 m long, was about 2 m. Thus, a relative accuracy of 0.33% was achieved, illustrating that the positioning can be kept to better than 0.5 m within a 150 meter GNSS outage. This accuracy is appropriate for positioning requirements of many location-based services in both indoor and outdoor environments.

As expected, two key reasons contribute to the development of the localization using omnidirectional cameras. First, tracked features on a sequence of omnidirectional images are more likely to be well-distributed surrounding the environment resulting in a more robust geometry of the triangulation and the bundle adjustment. This can resemble the position of the satellites for GNSS. The better distribution of the involved satellites over the user, the more accurate positioning of the GNSS would be achieved. The second reason, as seen in the experiments, the omnidirectional features are tracked over a longer window of camera poses leading to more robust and accurate localization.

To assist the reader in understanding the motion characteristics, other estimated parameters are also shown in Figure 6.20. In the middle left side, the motion velocity is shown with respect to frame $\{\mathcal{E}\}$. On the right side, middle, the estimated attitude is shown in the form of axis angle resolved in the local navigation frame for visualization purposes only. The local navigation frame is defined with its origin located at the mobile

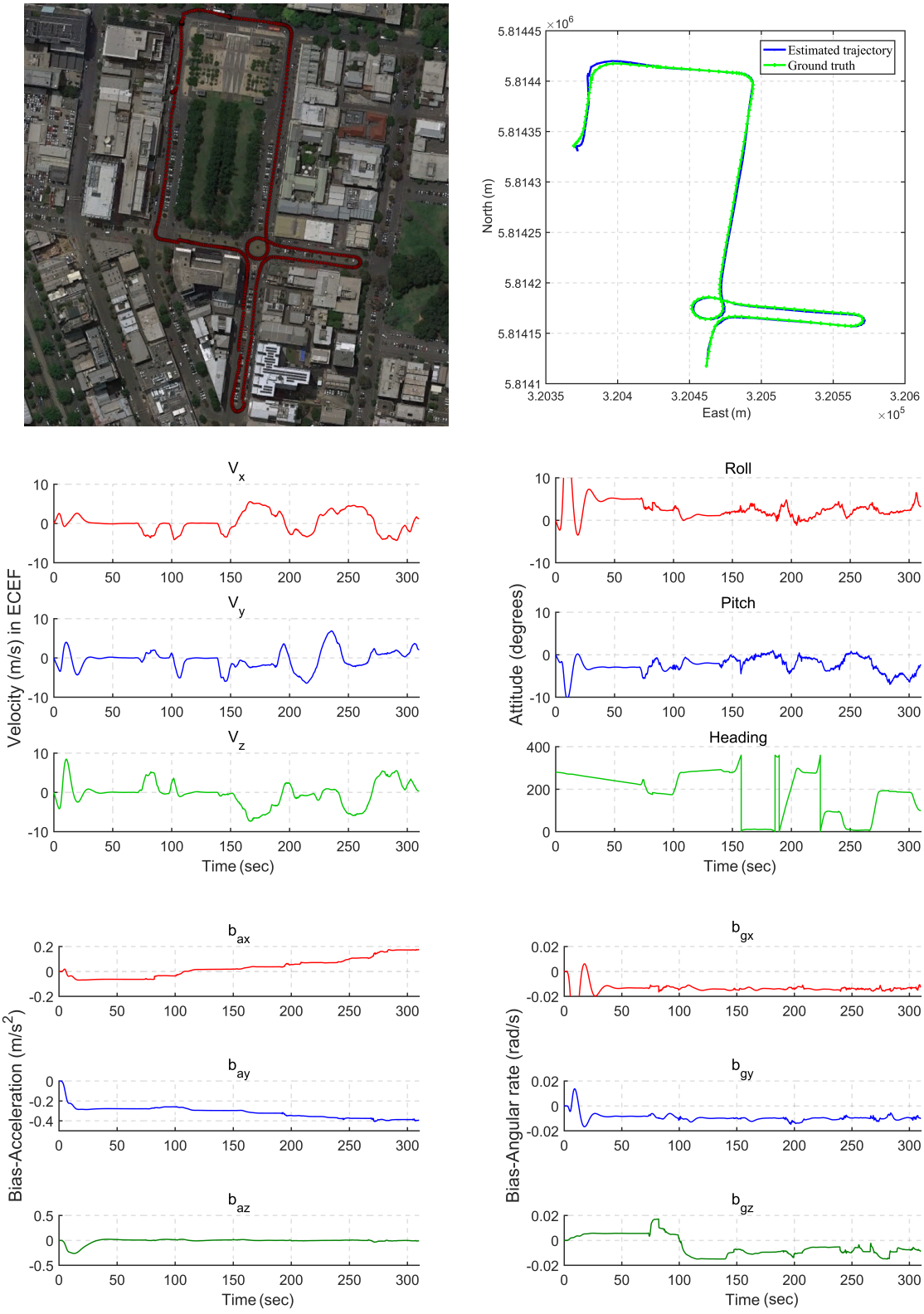


Figure 6.20: **Top**: 2D top view of the estimated trajectory in frame $\{\mathcal{E}\}$ (**left**), and the comparison between the estimated trajectory and the GNSS ground truth (**right**). **Middle**: Estimated velocity with respect to frame $\{\mathcal{E}\}$ (**right**), and attitude with respect to local navigation frame (**left**). **Bottom**: Acceleration and angular rate biases with respect to frame $\{\mathcal{B}\}$.

navigation system, the z axis being vertically downwards, and the x and y axes point to north and east, respectively. The estimated INS biases associated with acceleration and gyro outputs, with respect to frame $\{\mathcal{B}\}$, are also depicted at the bottom of Figure 6.20.

6.4 Summary

In this chapter, two VIO modules, presented in Chapter 4 and 5, have been evaluated. For stereo VIO, several datasets, collected by KITTI in an urban environment, were used. As noted in Chapter 4, a key advantage of the stereo VIO is that, in addition to using image observations along a single sequence of images, the epipolar constraint across the static stereo pairs of images, captured by a stereo camera system, can be exploited. This was carried out by matching the features across the stereo images using the approximate matching method. Furthermore, for the 3D reconstruction, observations from both static stereo and multiple images in time are used. This way, more robust features contribute to the update step of the EKF resulting in a more accurate localization. Experimental evaluation using three KITTI datasets showed that the stereo VIO outperformed the mono VIO using the MSCKF, despite the fact that the number of measurements in mono VIO is generally larger than that in the stereo VIO. This indicates that features tracked from both static and temporal stereo are more robust, therefore contributing to more accurate estimation of parameters.

The second VIO module, introduced as OVIO, was experimentally applied on synthetic datasets in both indoor and outdoor environments, so validated the new measurement model projecting residuals onto the tangential plane around each ray. These experiments highlighted the superior performance of the OVIO over mono VIO using a perspective camera. However, the experiment using synthetic images taken from a catadioptric camera showed that position estimation is not necessarily improved when the FoV increases to more than 180°, likely due to a reduction in angular resolution. On the other hand, it can be seen that a significant improvement in attitude estimation was

achieved, likely due to having features distributed over a larger area of the scene around the system. The OVIO approach utilizing a real-world video stream captured by an upward looking 180° fish-eye camera on top of a car was also experimentally evaluated. Results showed a small 0.33% drift of the OVIO from the ground truth over the travelled trajectory. In developing the proposed system, it was aimed to achieve a positional accuracy of better than 1 metre over 100 metre GNSS outages. The results of the experiments with synthetic and real data have illustrated both that this goal was realised, and that the performance is in accordance with design expectations.

Chapter 7
Conclusion and Future Work

7.1 Conclusion

THE thesis began with a literature review of existing localization approaches. Based on the selection of a coordinate system attached either to the environment or to the moving platform, the localization approaches were classified as either absolute or relative localization. In another classification, systems which exploit only on-board sensors for a localization purpose were grouped as sensor-fused localization, whereas systems which employ sensors on-board and off-board were grouped as cooperative localization. Studying the advantages and disadvantages of each method, the integration of vision with inertial readings was selected to achieve localization which is real-time, accurate and continuous. Having investigated various linear and non-linear methodologies of visual-inertial integration, the MSCKF method was selected due to the fact that this technique endeavours to utilize the maximum constraints that a visible feature creates from a sequence of images, while the computation remains linear in the number of features compared to non-linear methodologies in which the computation is at least quadratic. The mathematics of this approach was fully described in Chapter 3 and performance of the MSCKF was tested via a simulation.

In Chapter 4, an enhanced VIO method to robustly track visual features in a sequence of stereo images was described. The MSCKF algorithm, described in Chapter 3, was further improved to benefit from visual observations captured from a stereo camera instead of a single camera. This way, additional constraints across the stereo images contribute to

the localization resulting in better accuracy. Moreover, due to exploitation of the 3D scene reconstruction from a stereo image, the system enables a fine initialization shortly before motion is initiated. Also, combining static with temporal stereo images, the VIO system benefits from the reconstruction of micro and macro-scale environments simultaneously providing better localization.

In Chapter 5, a novel VIO approach was introduced that unlike other previous VIOs, centred upon estimation of localization parameters including position, velocity and attitude through the integration of an omnidirectional camera with inertial measurements. Since omnidirectional cameras have high distortion, the image plane is not a suitable space in which to describe the measurement model. Therefore, the measurement model was introduced on the unit sphere whose centre is coincident with the center of the camera lens. To link 3D scene points to their 2D projections on the image plane and their 3D corresponding points on the lens, in dioptric cameras, or on the mirror, in catadioptric cameras, a calibration method, described in Section 5.2, was used. By exploiting omnidirectional images, instead of a perspective camera, features from a larger portion of the scene being imaged can contribute to the localization, thus improving pose estimation. In addition, it is less likely that images captured by a camera with a FoV larger than 180° will fully covered by moving entities such as vehicles and pedestrians.

Finally, in Chapter 6, the novel new stereo-VIO and the OVIO methods were evaluated. The performance of the stereo-VIO was evaluated using three KITTI datasets. As seen, despite the fact that the number of measurements in mono-VIO is generally larger than in the stereo-VIO, the stereo-VIO outperformed the mono MSCKF. This indicated that features tracked from both static and temporal stereo are more robust, therefore contributing to a more accurate estimation motion and localization of parameters. The performance of the OVIO was experimentally evaluated and validated on synthetic datasets in both indoor and outdoor environments. These experiments showed the superiority of the OVIO over an mono-VIO system utilizing a perspective camera. The OVIO approach utilizing a real-world video stream captured by an upward looking 185° fish-eye camera

on top of a car was also successfully tested experimentally. Results of impressive robustness and accuracy were achieved; these showing a small 0.33% drift over the OVIO from the ground truth over the travelled trajectory.

7.2 Summary of Contributions

In aiming for the provision of a real-time, accurate and seamless localization system, this thesis has addressed problems of existing localization systems, particularly the vision-aided navigations systems. Towards this aim, the thesis made the following principal contributions:

- A novel stereo-VIO system based on the MSCKF approach, which integrates inertial measurements with visual observations extracted from both static and temporal stereo pairs through the multi-state constraint Kalman filtering algorithm. The exploitation of a stereo camera and the MSCKF approach results in a more robust localization system compared to the conventional MSCKF, which utilizes a mono camera, and other stereo-VIOs employing a non-MSCKF approach. In fact, on the one hand, the proposed stereo-VIO improves the robustness of feature tracking by cross-matching features in static stereo pairs and on the other hand, additional constraints across the stereo pairs are fully employed for a visible feature through the MSCKF approach, resulting in more accurate estimation.

- A novel omnidirectional VIO based on the MSCKF approach, which integrated inertial measurements with visual observations extracted from omnidirectional images through the multi-state constraint Kalman filtering algorithm. This method, which was introduced as OVIO, is one of the first vision-aided navigation systems that estimate 6 DoF ego-motion from the integration of inertial measurements with image observations taken by an omnidirectional camera. Moreover, because the MSCKF method is used as an estimator, a large window of camera poses from which a feature is visible can be augmented in the state vector which results in

more accurate localization. Also, omnidirectional cameras capture features from a greater proportion of the operating environment, thus providing better distribution of features around the system which improves the geometry to a certain degree.

7.3 Perspectives

Although the experiments demonstrated the superiority of the proposed new systems compared to conventional VIO systems, there exist several aspects that can be taken into account to achieve better performance.

- Because incremental drift is inevitable in the mechanism of VIO, to completely remove the drift over time, the use of external references is necessary. For instance, the exploitation of information extracted from a Building Information Model (BIM) or City Information Model (CIM) for indoor and outdoor localization, respectively, could remove the accumulation error. Another source of external reference could be from visual landmarks deployed in indoor or outdoor environments. Visual landmarks such as traffic signs or markings on roads with known coordinates with respect to a global frame could be beneficial for drift reduction.

- Since the presented VIOs enable the provision of localization with an accuracy of better than 1 meter over 100 meters, and usually in cities such as Melbourne one can expect to reach an intersection in 100 meters where a sufficient number of satellite observations can be received, the integration of the VIOs with the available GNSS observations at road intersections could assist in correcting drift.

- So far, existing VIOs take advantage of only metric information extracted from the cameras. Object recognition, known as semantic information, could be associated with geometric information, resulting in a more reliable localization. Moving toward autonomous navigation, robot operability will be significantly improved if robot decisions are based not only on metric information, but also on the functional

attributes and interconnections of the entities in different environments. To utilize semantic information, the use of algorithms such as deep learning and reinforcement learning are necessary.

- Apart from using a stereo camera or an omnidirectional camera, the integration of inertial measurements with other types of cameras could be beneficial. One of these types of cameras is the event-based camera, which is only sensitive to movement and can capture images at a high frequency. Through integration of this type of cameras with inertial readings, the precise localization of very dynamic platforms such as drones will be quite possible.

Appendix A

Appendix

A.1 3D Scene Reconstruction

In order to estimate the position of the tracked feature, f_j, MSCKF uses an inverse-depth least-squares Gauss-Newton optimization. The input of the optimization is the rectified image coordinates of feature f_j and the corresponding camera poses.

If $\mathcal{C}_f$ is the first frame in which the feature f_j is visible, the position of the feature with respect to the camera frame $\{\mathcal{C}_i\}$ is given by:

$$^{\mathcal{C}_i}\mathbf{p}_{f_j} = \mathbf{C}_{\mathcal{C}_f}^{\mathcal{C}_i}\ {}^{\mathcal{C}_f}\mathbf{p}_{f_j} + {}^{\mathcal{C}_i}\mathbf{p}_{\mathcal{C}_f} = (\mathbf{C}_{\mathcal{C}_f}^{\mathcal{C}_i}\begin{pmatrix} {}^{\mathcal{C}_f}X_j \\ {}^{\mathcal{C}_f}Y_j \\ {}^{\mathcal{C}_f}Z_j \end{pmatrix} + {}^{\mathcal{C}_i}\mathbf{p}_{\mathcal{C}_f}) \tag{A.1}$$

By considering $\alpha_j = \frac{{}^{\mathcal{C}_f}X_j}{{}^{\mathcal{C}_f}Z_j}$, $\beta_j = \frac{{}^{\mathcal{C}_f}Y_j}{{}^{\mathcal{C}_f}Z_j}$, and $\gamma_j = \frac{1}{{}^{\mathcal{C}_f}Z_j}$, (A.1) can be re-cast as:

$$^{\mathcal{C}_i}\mathbf{p}_{f_j} = \frac{1}{\gamma_j}\begin{bmatrix} h_{i1}(\alpha_j, \beta_j, \gamma_j) \\ h_{i2}(\alpha_j, \beta_j, \gamma_j) \\ h_{i3}(\alpha_j, \beta_j, \gamma_j) \end{bmatrix} \tag{A.2}$$

where h_{i1}, h_{i2}, and h_{i3} are scalar functions of α_j, β_j, and γ_j, respectively. Substituting (A.2)

into (3.27), the model for a perspective camera is expressed as:

$$\mathbf{z}_i^{(j)} = \frac{1}{h_{i3}} \begin{bmatrix} h_{i1} \\ h_{i2} \end{bmatrix} + \boldsymbol{\eta}_i^{(j)} \tag{A.3}$$

The vector of camera measurement error used in the least-squares system is then defined as:

$$\mathbf{r}_i^{(j)}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \hat{\mathbf{z}}_i^{(j)} - \frac{1}{h_{i3}(\hat{\alpha}, \hat{\beta}, \hat{\gamma})} \begin{bmatrix} h_{i1}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \\ h_{i2}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \end{bmatrix} \tag{A.4}$$

Computing the entire residuals of feature f_j, the estimated value of α, β, and γ can be obtained and finally the 3D position of feature f_j can be estimated by Gauss-Newton least-squares minimization as:

$${}^{\mathcal{E}}\hat{\mathbf{p}}_{f_j} = \frac{1}{\hat{\rho}_j} \mathbf{C}_{\mathcal{C}_f}^{\mathcal{E}} \begin{bmatrix} \hat{\alpha_j} \\ \hat{\beta}_j \end{bmatrix} + {}^{\mathcal{E}}\hat{\mathbf{p}}_{\mathcal{C}_f} \tag{A.5}$$

Bibliography

A.Arnold, J. and Persaud, R. Mapping technology assessment for connected vehicle highway network applications. U.S. Department of Transportation, Federal Highway Administration. Available at: https://www.gps.gov/cgsic/states/2012/seattle/arnold.pdf, 2012.

Abdulrahim, K., Hide, C., Moore, T., and Hill, C. Aiding mems imu with building heading for indoor pedestrian navigation. In *Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2010*, Pp. 1–6. IEEE, 2010.

Alam, N. and Dempster, A. G. Cooperative positioning for vehicular networks: Facts and future. *IEEE transactions on intelligent transportation systems*, 14(4):1708–1717. IEEE, 2013.

Alt, H., Efrat, A., Rote, G., and Wenk, C. Matching planar maps. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, Pp. 589–598. Society for Industrial and Applied Mathematics, 2003.

Aulinas, J., Petillot, Y. R., Salvi, J., and Lladó, X. The slam problem: a survey. *CCIA*, 184(1):363–371, 2008.

Barreto, J. P. and Araujo, H. Issues on the geometry of central catadioptric image formation. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, Pp. II–II. IEEE, 2001.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359. Elsevier, 2008.

Bayard, D. S. and Brugarolas, P. B. An estimation algorithm for vision-based exploration of small bodies in space. In *American Control Conference, 2005. Proceedings of the 2005*, Pp. 4589–4595. IEEE, 2005.

Benslimane, A. Localization in vehicular ad hoc networks. In *Systems Communications, 2005. Proceedings*, Pp. 19–25. IEEE, 2005.

Bisnath, S. and Gao, Y. Current state of precise point positioning and future prospects and limitations. In *Observing our changing earth*, Pp. 615–623. Springer, 2009.

Biswas, J. and Veloso, M. Depth camera based indoor mobile robot localization and navigation. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, Pp. 1697–1702. IEEE, 2012.

Brakatsoulas, S., Pfoser, D., Salas, R., and Wenk, C. On map-matching vehicle tracking data. In *Proceedings of the 31st international conference on Very large data bases*, Pp. 853–864. VLDB Endowment, 2005.

Caruso, D., Engel, J., and Cremers, D. Large-scale direct slam for omnidirectional cameras. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, Pp. 141–148. IEEE, 2015.

Davison, A. J. Real-time simultaneous localisation and mapping with a single camera. In *null*, P. 1403. IEEE, 2003.

Deuker, A. Del 11.2: Mobility and lbs. *FIDIS Deliverables*, 11(2). Citeseer, 2008.

Diel, D. D., DeBitetto, P., and Teller, S. Epipolar constraints for vision-aided inertial navigation. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 2, Pp. 221–228. IEEE, 2005.

dos Santos, D. R., Basso, M. A., Khoshelham, K., de Oliveira, E., Pavan, N. L., and Vosselman, G. Mapping indoor spaces by adaptive coarse-to-fine registration of rgb-d data. *IEEE Geoscience and Remote Sensing Letters*, 13(2):262–266. IEEE, 2016.

Drawil, N. M. and Basir, O. Intervehicle-communication-assisted localization. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):678–691. IEEE, 2010.

Eade, E. and Drummond, T. Scalable monocular slam. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 1*, Pp. 469–476. IEEE Computer Society, 2006.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., and Burgard, W. An evaluation of the rgb-d slam system. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, Pp. 1691–1696. IEEE, 2012.

Engel, J., Schöps, T., and Cremers, D. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, Pp. 834–849. Springer, 2014.

Forster, C., Carlone, L., Dellaert, F., and Scaramuzza, D. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. Georgia Institute of Technology, 2015.

Förstner, W. Minimal representations for testing and estimation in projective spaces. *Photogrammetrie-Fernerkundung-Geoinformation*, 2012(3):209–220. E. Schweizerbart'sche Verlagsbuchhandlung, 2012.

Furgale, P., Rehder, J., and Siegwart, R. Unified temporal and spatial calibration for multi-sensor systems. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, Pp. 1280–1286. IEEE, 2013.

Ganganath, N. and Leung, H. Mobile robot localization using odometry and kinect sensor. In *Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on*, Pp. 91–94. IEEE, 2012.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, Pp. 3354–3361. IEEE, 2012.

Geyer, C. and Daniilidis, K. A unifying theory for central panoramic systems and practical implications. *Computer Vision—ECCV 2000*, Pp. 445–461. Springer, 2000.

Greenfeld, J. S. Matching gps observations to locations on a digital map. In *81th annual meeting of the transportation research board*, volume 1, Pp. 164–173, 2002.

Grejner-Brzezinska, D. A., Toth, C. K., Sun, H., Wang, X., and Rizos, C. A robust solution to high-accuracy geolocation: Quadruple integration of gps, imu, pseudolite, and terrestrial laser scanning. *IEEE Transactions on instrumentation and measurement*, 60(11):3694–3708. IEEE, 2011.

Groves, P. D. *Principles of GNSS, inertial, and multisensor integrated navigation systems*. Artech house, 2013.

Gu, F., Kealy, A., Khoshelham, K., and Shang, J. Efficient and accurate indoor localization using landmark graphs. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41, 2016a.

Gu, F., Khoshelham, K., Shang, J., and Yu, F. Sensory landmarks for indoor localization. In *Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS), 2016 Fourth International Conference on*, Pp. 201–206. IEEE, 2016b.

Gu, Y., Lo, A., and Niemegeers, I. A survey of indoor positioning systems for wireless personal networks. *IEEE Communications surveys & tutorials*, 11(1):13–32. IEEE, 2009.

Hannah, M. J. Computer matching of areas in stereo images. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1974.

Hanning, G., Forslöw, N., Forssén, P.-E., Ringaby, E., Törnqvist, D., and Callmer, J. Stabilizing cell phone video using inertial measurement sensors. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Pp. 1–8. IEEE, 2011.

Hartley, R. and Zisserman, A. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Hawkins, A. J. Ford says this sensor will make its driverless cars the most advanced in the world. The Verge. Available at: http://www.theverge.com/2016/1/5/10717040/ford-driverless-carvelodyne-ultra-puck-lidar-ces-2016, 2016.

Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663. SAGE Publications Sage UK: London, England, 2012.

Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. *Global positioning system: theory and practice*. Springer Science & Business Media, 2012.

Hofmann-Wellenhof, B. and Moritz, H. *Physical geodesy*. Springer Science & Business Media, 2006.

Hummel, B. Map matching for vehicle guidance. *Dynamic and Mobile GIS: Investigating Space and Time*, Pp. 437–438. CRC Press, 2006.

Huster, A., Frew, E. W., and Rock, S. M. Relative position estimation for auvs by fusing bearing and inertial rate sensor measurements. In *OCEANS'02 MTS/IEEE*, volume 3, Pp. 1863–1870. IEEE, 2002.

Ingram, S., Harmer, D., and Quinlan, M. Ultrawideband indoor positioning systems and their use in emergencies. In *Position Location and Navigation Symposium, 2004. PLANS 2004*, Pp. 706–715. IEEE, 2004.

Javanmardi, M., Javanmardi, E., Gu, Y., and Kamijo, S. Towards high-definition 3d urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery. *Remote Sensing*, 9(10):975. Multidisciplinary Digital Publishing Institute, 2017.

Jia, C. and Evans, B. L. Probabilistic 3-d motion estimation for rolling shutter video rectification from visual and inertial measurements. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, Pp. 203–208. IEEE, 2012.

Kaijaluoto, R., Kukko, A., and Hyyppä, J. Precise indoor localization for mobile laser scanner. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Pp. 1–6, 2015.

Kannala, J. and Brandt, S. S. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340. IEEE, 2006.

Karpenko, A., Jacobs, D., Baek, J., and Levoy, M. Digital video stabilization and rolling shutter correction using gyroscopes. *CSTR*, 1:2, 2011.

Khoshelham, K. Direct 6-dof pose estimation from point-plane correspondences. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, Pp. 1–6. IEEE, 2015.

Khoshelham, K. Closed-form solutions for estimating a rigid motion from plane correspondences extracted from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:78–91. Elsevier, 2016.

Khoshelham, K., Dos Santos, D., and Vosselman, G. Generation and weighting of 3d point correspondences for improved registration of rgb-d data. *Proc. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, 5:W2, 2013.

Khoshelham, K. and Elberink, S. O. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454. Molecular Diversity Preservation International, 2012.

Kim, J., Jeong, J., Shin, Y.-S., Cho, Y., Roh, H., and Kim, A. Lidar configuration comparison for urban mapping system. In *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*, Pp. 854–857. IEEE, 2017.

Kim, J. and Sukkarieh, S. Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, 55(1):62–71. Elsevier, 2007.

Klein, G. and Murray, D. Parallel tracking and mapping on a camera phone. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, Pp. 83–86. IEEE, 2009.

Kneip, L., Scaramuzza, D., and Siegwart, R. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, Pp. 2969–2976. IEEE, 2011.

Kontkanen, P., Myllymaki, P., Roos, T., Tirri, H., Valtonen, K., and Wettig, H. Topics in probabilistic location estimation in wireless networks. In *Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004. 15th IEEE International Symposium on*, volume 2, Pp. 1052–1056. IEEE, 2004.

Kottas, D. G. and Roumeliotis, S. I. Exploiting urban scenes for vision-aided inertial navigation. In *Robotics: Science and Systems*, 2013.

Kouba, J. and Héroux, P. Precise point positioning using igs orbit and clock products. *GPS solutions*, 5(2):12–28. Springer, 2001.

Krumm, J., Horvitz, E., and Letchner, J. Map matching with travel time constraints. Technical report, SAE Technical Paper, 2007.

Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., and Burgard, W. g 2 o: A general framework for graph optimization. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, Pp. 3607–3613. IEEE, 2011.

Lacroix, S., Mallet, A., Chatila, R., and Gallo, L. Rover self localization in planetary-like environments. In *Artificial Intelligence, Robotics and Automation in Space*, volume 440, P. 433, 1999.

Langelaan, J. W. State estimation for autonomous flight in cluttered environments. *Journal of Guidance Control and Dynamics*, 30(5):1414. AMERICAN INST OF AERONAUTICS AND ASTRONAUTICS, 2007.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., and Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334. SAGE Publications Sage UK: London, England, 2015.

Levinson, J., Montemerlo, M., and Thrun, S. Map-based precision vehicle localization in urban environments. In *Robotics: Science and Systems*, volume 4, P. 1, 2007.

Li, L., Liu, Y.-H., Wang, K., and Fang, M. A simple algorithm for position and velocity estimation of mobile robots using omnidirectional vision and inertial sensors. In *Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on*, Pp. 1235–1240. IEEE, 2014.

Li, M., Kim, B. H., and Mourikis, A. I. Real-time motion tracking on a cellphone using inertial sensing and a rolling-shutter camera. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, Pp. 4712–4719. IEEE, 2013.

Li, M. and Mourikis, A. I. High-precision, consistent ekf-based visual-inertial odometry. *The International Journal of Robotics Research*, 32(6):690–711. SAGE Publications Sage UK: London, England, 2013.

Li, M. and Mourikis, A. I. Vision-aided inertial navigation with rolling-shutter cameras. *The International Journal of Robotics Research*, 33(11):1490–1507. SAGE Publications Sage UK: London, England, 2014.

Li, X., Li, M., Shu, W., and Wu, M. A practical map-matching algorithm for gps-based vehicular networks in shanghai urban area. IET, 2007.

Lin, K.-H., Chang, C.-H., Dopfer, A., and Wang, C.-C. Mapping and localization in 3d environments using a 2d laser scanner and a stereo camera. *Journal of information science and engineering*, 28(1):131–144. Institute of Information Science Academia Sinica, 2012.

Lindeberg, T. Image matching using generalized scale-space interest points. *Journal of Mathematical Imaging and Vision*, 52(1):3–36. Springer, 2015.

Liu, H., Darabi, H., Banerjee, P., and Liu, J. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080. IEEE, 2007.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110. Springer, 2004.

Ma, X., Chen, X., and Refai, H. H. Performance and reliability of dsrc vehicular safety communication: a formal analysis. *EURASIP Journal on Wireless Communications and Networking*, 2009(1):969164. Springer, 2009.

Ma, Y., Soatto, S., Kosecka, J., and Sastry, S. S. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012.

Manandhar, D. and Shibasaki, R. Vehicle-borne laser mapping system (vlms) for 3-d gis. In *Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International*, volume 5, Pp. 2073–2075. IEEE, 2001.

Mei, C. and Rives, P. Single view point omnidirectional camera calibration from planar grids. In *Robotics and Automation, 2007 IEEE International Conference on*, Pp. 3945–3950. IEEE, 2007.

Mervart, L., Lukes, Z., Rocken, C., and Iwabuchi, T. Precise point positioning with ambiguity resolution in real-time. In *Proceedings of ION GNSS*, Pp. 397–405, 2008.

Micušık, B. *Two-view geometry of omnidirectional cameras*. PhD thesis, Czech Technical University, 2004.

Montiel, J. M., Civera, J., and Davison, A. J. Unified inverse depth parametrization for monocular slam. volume 9, P. 1. Robotics: Science and Systems, 2006.

Moravec, H. P. Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1980.

Moreno-Noguer, F., Lepetit, V., and Fua, P. Accurate non-iterative o (n) solution to the pnp problem. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*, Pp. 1–8. IEEE, 2007.

Mourikis, A. I. A multi-state constraint kalman filter for vision-aided inertial navigation. Technical report, Dept. of Computer Science and Engineering, University of Minnesota. Available at: www.cs.umn.edu/$\sim$mourikis/tech_reports/TR_MSCKF.pdf, 2006.

Mourikis, A. I. and Roumeliotis, S. I. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Robotics and automation, 2007 IEEE international conference on*, Pp. 3565–3572. IEEE, 2007.

Muja, M. and Lowe, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.

Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163. IEEE, 2015.

Newcombe, R. A., Fox, D., and Seitz, S. M. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Pp. 343–352, 2015.

Newton, I., Motte, A., and Chittenden, N. W. *Newton's principia: The mathematical principles of natural philosophy*. Geo. P. Putnam, 1850.

Ni, K. and Dellaert, F. Multi-level submap based slam using nested dissection. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, Pp. 2558–2565. IEEE, 2010.

Nistér, D. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770. IEEE, 2004.

Nistér, D., Naroditsky, O., and Bergen, J. Visual odometry. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, Pp. I–I. IEEE, 2004.

Nistér, D., Naroditsky, O., and Bergen, J. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1):3–20. Wiley Online Library, 2006.

Nistér, D. and Stewénius, H. A minimal solution to the generalised 3-point pose problem. *Journal of Mathematical Imaging and Vision*, 27(1):67–79. Springer, 2007.

Nüchter, A., Lingemann, K., Hertzberg, J., and Surmann, H. 6d slam—3d mapping outdoor environments. *Journal of Field Robotics*, 24(8-9):699–722. Wiley Online Library, 2007.

Olson, C. F., Matthies, L. H., Schoppers, H., and Maimone, M. W. Robust stereo ego-motion for long distance navigation. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, Pp. 453–458. IEEE, 2000.

Ong, L., Ridley, M., Kim, J.-H., Nettleton, E., Sukkarieh, S., et al. Six dof decentralised slam. In *Australasian Conference on Robotics and Automation*, Pp. 10–16, 2003.

Oskiper, T., Zhu, Z., Samarasekera, S., and Kumar, R. Visual odometry system using multiple stereo cameras and inertial measurement unit. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, Pp. 1–8. IEEE, 2007.

Parker, R. and Valaee, S. Vehicular node localization using received-signal-strength indicator. *IEEE Transactions on Vehicular Technology*, 56(6):3371–3380. IEEE, 2007.

Patwari, N., Ash, J. N., Kyperountas, S., Hero, A. O., Moses, R. L., and Correal, N. S. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal processing magazine*, 22(4):54–69. IEEE, 2005.

Patwari, N., Hero, A. O., Perkins, M., Correal, N. S., and O'dea, R. J. Relative location estimation in wireless sensor networks. *IEEE Transactions on signal processing*, 51(8):2137–2148. IEEE, 2003.

Pirchheim, C., Schmalstieg, D., and Reitmayr, G. Handling pure camera rotation in keyframe-based slam. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, Pp. 229–238. IEEE, 2013.

Prasithsangaree, P., Krishnamurthy, P., and Chrysanthis, P. On indoor position location with wireless lans. In *Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on*, volume 2, Pp. 720–724. IEEE, 2002.

Prazenica, R. J., Watkins, A., Kurdila, A. J., Ke, Q. F., and Kanade, T. Vision-based kalman filtering for aircraft state estimation and structure from motion. In *AIAA Guidance, Navigation, and Control Conference*, volume 5, P. 3, 2005.

Qu, X. *Landmark based localization: Detection and update of landmarks with uncertainty analysis*. PhD thesis, Université Paris-Est, 2017.

Quddus, M. A., Ochieng, W. Y., Zhao, L., and Noland, R. B. A general map matching algorithm for transport telematics applications. *GPS solutions*, 7(3):157–167. Springer, 2003.

Ramezani, M. and Khoshelham, K. Vehicle positioning in gnss-deprived urban areas by stereo visual-inertial odometry. *IEEE Transactions on Intelligent Vehicles*. IEEE, 2018.

Ramezani, M., Khoshelham, K., and Fraser, C. Pose estimation by omnidirectional visual-inertial odometry. *Robotics and Autonomous Systems*. Elsevier, 2018.

Rosen, D. M., DuHadway, C., and Leonard, J. J. A convex relaxation for approximate global optimization in simultaneous localization and mapping. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, Pp. 5822–5829. IEEE, 2015.

Roumeliotis, S. I., Johnson, A. E., and Montgomery, J. F. Augmenting inertial navigation with image-based motion estimation. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 4, Pp. 4326–4333. IEEE, 2002.

Saha, A. K. and Johnson, D. B. Modeling mobility for vehicular ad-hoc networks. In *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, Pp. 91–92. ACM, 2004.

Saha, S., Chaudhuri, K., Sanghi, D., and Bhagwat, P. Location determination of a mobile device using ieee 802.11 b access point signals. In *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, volume 3, Pp. 1987–1992. IEEE, 2003.

Salahat, E. and Qasaimeh, M. Recent advances in features extraction and description algorithms: A comprehensive survey. In *Industrial Technology (ICIT), 2017 IEEE International Conference on*, Pp. 1059–1063. IEEE, 2017.

Scaramuzza, D. Omnidirectional vision: from calibration to robot motion estimation. *ETH Zurich, PhD Thesis*, 17635, 2008a.

Scaramuzza, D. *Omnidirectional vision: from calibration to robot motion estimation*. PhD thesis, ETH Zurich, 2008b.

Scaramuzza, D. and Friedrich, F. Visual odometry: Part i: The first 30 years and fundamentals. *IEEE Robotics & Automation Magazine*, 2011.

Scaramuzza, D., Martinelli, A., and Siegwart, R. A flexible technique for accurate omnidirectional camera calibration and structure from motion. In *Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on*, Pp. 45–45. IEEE, 2006a.

Scaramuzza, D., Martinelli, A., and Siegwart, R. A toolbox for easily calibrating omnidirectional cameras. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, Pp. 5695–5701. IEEE, 2006b.

Shao, Z., Li, W., Wu, Y., and Shen, L. Multi-layer and multi-dimensional information based cooperative vehicle localization in highway scenarios. In *Communication Technology (ICCT), 2010 12th IEEE International Conference on*, Pp. 567–571. IEEE, 2010.

Shepard, S. *RFID: radio frequency identification*. McGraw Hill Professional, 2005.

Sibley, G., Matthies, L., and Sukhatme, G. Sliding window filter with application to planetary landing. *Journal of Field Robotics*, 27(5):587–608. Wiley Online Library, 2010.

Siddiqi, S., Sukhatme, G. S., and Howard, A. Experiments in monte-carlo localization using wifi signal strength. In *Proceedings of the International Conference on Advanced Robotics*, Pp. 210–223, 2003.

Snavely, N., Seitz, S. M., and Szeliski, R. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, Pp. 835–846. ACM, 2006.

Snavely, N., Seitz, S. M., and Szeliski, R. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210. Springer, 2008.

Soatto, S., Frezza, R., and Perona, P. Motion estimation via dynamic vision. *IEEE Transactions on Automatic Control*, 41(3):393–413. IEEE, 1996.

Soatto, S. and Perona, P. Recursive 3-d visual motion estimation using subspace constraints. *International Journal of Computer Vision*, 22(3):235–259. Springer, 1997.

Sola, J. Quaternion kinematics for the error-state kf. *Laboratoire d'Analyse et d'Architecture des Systemes-Centre national de la recherche scientifique (LAAS-CNRS), Toulouse, France, Tech. Rep*, (hal-01122406), 2012.

Strasdat, H., Montiel, J., and Davison, A. J. Scale drift-aware large scale monocular slam. *Robotics: Science and Systems VI*, 2. Universidad de Zaragoza Zaragoza, Spain, 2010.

Strelow, D. and Singh, S. Motion estimation from image and inertial measurements. *The International Journal of Robotics Research*, 23(12):1157–1195. SAGE Publications, 2004.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. A benchmark for the evaluation of rgb-d slam systems. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Pp. 573–580. IEEE, 2012.

Taguchi, Y., Jian, Y.-D., Ramalingam, S., and Feng, C. Point-plane slam for hand-held 3d sensors. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, Pp. 5182–5189. IEEE, 2013.

Talaya, J., Alamus, R., Bosch, E., Serra, A., Kornus, W., and Baron, A. Integration of a terrestrial laser scanner with gps/imu orientation sensors. In *Proceedings of the XXth ISPRS Congress*, Pp. 1049–1055, 2004.

Tardif, J.-P., George, M., Laverne, M., Kelly, A., and Stentz, A. A new approach to vision-aided inertial navigation. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, Pp. 4161–4168. IEEE, 2010.

Tardif, J.-P., Pavlidis, Y., and Daniilidis, K. Monocular visual odometry in urban environments using an omnidirectional camera. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, Pp. 2531–2538. IEEE, 2008.

Tomasi, C. and Kanade, T. Detection and tracking of point features. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

Topaz. Topaz local positioning solution. website: http://www.tadlys.com/.

Torr, P. and Zisserman, A. Robust computation and parametrization of multiple view relations. In *Computer Vision, 1998. Sixth International Conference on*, Pp. 727–732. IEEE, 1998.

Ubisense. Ubisense: Location intelligence solutions and software platform. website: http://www.ubisense.net.

Urban, S., Leitloff, J., and Hinz, S. Improved wide-angle, fisheye and omnidirectional camera calibration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:72 – 79. http://www.sciencedirect.com/science/article/pii/S0924271615001616, 2015.

Vectornav. Inertial measurement unites and inertial navigation. website: https://www.vectornav.com/support/library/imu-and-ins.

Vossiek, M., Wiebking, L., Gulden, P., Weighardt, J., and Hoffmann, C. Wireless local positioning-concepts, solutions, applications. In *Radio and Wireless Conference, 2003. RAWCON'03. Proceedings*, Pp. 219–224. IEEE, 2003.

Williams, B., Klein, G., and Reid, I. Real-time slam relocalisation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Pp. 1–8. IEEE, 2007.

Wübbena, G., Schmitz, M., and Bagge, A. Ppp-rtk: precise point positioning using state-space representation in rtk networks. In *Proceedings of ION GNSS*, volume 5, Pp. 13–16, 2005.

Yang, L., Zhang, L., Dong, H., Alelaiwi, A., and El Saddik, A. Evaluating and improving the depth accuracy of kinect for windows v2. *IEEE Sensors Journal*, 15(8):4275–4285. IEEE, 2015.

Yin, J., ElBatt, T., Yeung, G., Ryu, B., Habermas, S., Krishnan, H., and Talty, T. Performance evaluation of safety applications over dsrc vehicular ad hoc networks. In *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*, Pp. 1–9. ACM, 2004.

You, S. and Neumann, U. Fusion of vision and gyro tracking for robust augmented reality registration. In *Virtual Reality, 2001. Proceedings. IEEE*, Pp. 71–78. IEEE, 2001.

Youssef, M. and Agrawala, A. Handling samples correlation in the horus system. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2, Pp. 1023–1031. IEEE, 2004.

Youssef, M. A., Agrawala, A., and Shankar, A. U. Wlan location determination via clustering and probability distributions. In *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on*, Pp. 143–150. IEEE, 2003.

Zhang, Z., Rebecq, H., Forster, C., and Scaramuzza, D. Benefit of large field-of-view cameras for visual odometry. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, Pp. 801–808. IEEE, 2016.

Zhao, L., Huang, S., and Dissanayake, G. Linear slam: A linear solution to the feature-based and pose graph slam based on submap joining. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, Pp. 24–30. IEEE, 2013.

ZigBeeTMAlliance. Network specification version 1.0. website: http://www.zigbee.org.

Zumberge, J., Heflin, M., Jefferson, D., Watkins, M., and Webb, F. H. Precise point positioning for the efficient and robust analysis of gps data from large networks. *Journal of Geophysical Research: Solid Earth*, 102(B3):5005–5017. Wiley Online Library, 1997.