Large Scale Structure from Motion for Autonomous Underwater Vehicle Surveys

by

Oscar Pizarro

Engineer's Degree, Electronics Engineering
Universidad de Concepcion, Chile (1997)
S.M. O.E./E.E.C.S. Massachusetts Institute of Technology and Woods Hole Oceanographic Institution (2003)

Submitted to the Joint Program in Applied Ocean Science and Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Oceanographic Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2004

Author ....
Joint Program in Applied Ocean Science and Engineering
September, 2004

Certified by..
Hanumant Singh
Associate Scientist, WHOI
Thesis Supervisor

Accepted by..
Mark Grosenbaugh
Associate Scientist, WHOI
Chairman, Joint Committee for Applied Ocean Science and Engineering

Large Scale Structure from Motion for Autonomous Underwater Vehicle Surveys

by

Oscar Pizarro

Submitted to the Joint Program in Applied Ocean Science and Engineering
on September, 2004, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Oceanographic Engineering

Abstract

Our ability to image extended underwater scenes is severely limited by attenuation and backscatter. Generating a composite view from multiple overlapping images is usually the most practical and flexible way around this limitation. In this thesis we look at the general constraints associated with imaging from underwater vehicles for scientific applications – low overlap, non-uniform lighting and unstructured motion – and present a methodology for dealing with these constraints toward a solution of the problem of large area 3D reconstruction.

Our approach assumes navigation data is available to constrain the structure from motion problem. We take a hierarchical approach where the temporal image sequence is broken into subsequences that are processed into 3D reconstructions independently. These submaps are then registered to infer their overall layout in a global frame. From this point a bundle adjustment refines camera and structure estimates.

We demonstrate the utility of our techniques using real data obtained during a SeaBED AUV coral reef survey. Test tank results with ground truth are also presented to validate the methodology.

Thesis Supervisor: Hanumant Singh
Title: Associate Scientist, WHOI

Acknowledgments

My advisor, Hanumant Singh, has been a nonstop fountain of energy and enthusiasm from start to finish. Always willing to give me time to figure things out. I would've never guessed his patience was to last this long. Thanks, Hanu.

My committee members have all shined. John Leonard's passion for robotics has been a constant source of inspiration. I hope Seth Teller's ability to go after the interesting result and then show it clearly somehow appears in this thesis. Marc Pollefeys always had the precise reference to help with a problem. Marc's optimism was critical when we drove from MIT to WHOI through a snow storm in a 86 Trooper with no muffler and a loose hood.

WHOI's Deep Submergence Lab set the bar when it comes to being relentless. Steve Gegg took me under his wing on my first Jason cruise to the Juan de Fuca Ridge. His stories of Woods Hole and the Captain Kidd 'back in the day' inspired many a fun night of carousing. I've had the opportunity to hike, party, sail and dive with Matt Heintz. Thanks for all of it. Jon Howland was willing to part with many books. Ann Stone has been a tireless advocate willing to forgive my inability to grasp concepts such as 'advanced request & authorization to travel'.

The Academic Programs Office at WHOI was truly remarkable in their good will, genuine concern and effectiveness. Marsha and Julia went out the way to help when it was usually (and clearly) a problem of my own making. They feel like dear friends by now. And boy can they dance. John Farrington was willing to believe in a couple of grad students which has led to a cruise to Chile next year.

Building and running Seabed was a challenge and a great opportunity to work with some amazing people at WHOI. Neil McPhee was great in the lab and at sea. Kenny Doherty and Terry Hammar showed the way when designing pressure housings. John and Terry in the machine shop were willing to offer all their vast experience. Esmail and Carlos helped out greatly with Seabed and many last minute requests. You'd never get the answer you were looking for at Al Bradley's lab, only the tools to figure it out yourself. Thanks Al Duester for all the patience with the batteries. Dave Olmstead and the R/V Asterias helped SeaBED give its first steps. The guards at WHOI took good care of us and still think we were building a underwater beer making machine. Marga, you're the best. James Kinsey

has lots of good Karma coming his way and several kegs in beer heaven for his help with experiments at JHU. Louis Whitcomb set many a good example.

My lab buddies Ryan and Chris defined much of grad school. Our styles are very different yet I feel we gained each others' respect. I'm glad I could consistently get Ryan to laugh and that I got to see him blossom into a calm and extremely capable dude. I've been very fortunate to work and live with Chris. His vast array of skills have come in handy countless times. Though you wouldn't guess it from the looks, he knows how to enjoy life and is always willing to share that with friends. Thanks for introducing me to windsurfing. Thanks guys for all the hard work, the trips, and the arguments. It all made the experience richer.

Living at Millfield was many things but never dull. Thanks to Dirk and Jim for good conversations, great dinners, and putting up with my bad movie choices. Thanks Charlie for defending your ideals, for the runs together, and for being Charlie. Mark Johnson to this day keeps stretching the definition of outrageous and intense. Joe Warren set an example on living the grad student life and took me out for my first paddle. Steph offered kind words, good company and gave a feminine touch to Millfield. Mike Jakuba for a wonderful mix of engineering-know-how and knowing how to have fun. Rachel is the best girlfriend of a friend I've ever met.

This place draws some unique people: Dave Lund has the perfect mix of guts, heart, sense of humor and mischievousness. Few friends can pull off being both an accomplice and confessional priest at the same time. Thanks for all the laughs, for all the moral dilemmas, for listening and for sharing all those meals and beers. Fabian, by an accident of nature, is from Chile. To this day I still argue that this had nothing to do with us becoming close friends. Chilean Spanish is rich in expressions and imagery which helped time and again when it came to unloading my latest love problem on Fabian's eternally patient shoulders. His Buddhist-like way of accepting friends and acquaintances just as they are and enjoying life with them fully has been a source of never ending amazement. Lara and Henrik were great hosts and even better friends. Jason defines 'intelligent fun' every time. Thanks Claudia for looking after me and for some amazing cooking.

Tim Prestero got me to enjoy my first year at MIT. Since then a high fraction of the best times happened with Tim nearby. Liz became a warm and gentle friend by her own means. At their wedding I cried more than all the other best men and maids of honor combined. Thanks to Mark and Linda Prestero for feeling like family.

Jeanne and I made it rain after a brief spring. Thanks for believing in me. Florica was a steady friend during some of the rainiest days in Belgium and let me crash on her couch when I needed it most. Marie-Louise was the best travel companion I could ask for. A special thanks to Murray and the Heights. They opened their home to me in Australia for three fabulous weeks. And that was the moment this thesis started to really come together.

This last year Dr. Rich Camilli reminded me that there is life, and mostly good life, after grad school. Rhea kept me sane and warm when it was cold everywhere. She offered affection, friendship and genuine caring that woke up things I thought had been lost for good. Her courage to stay in touch with what matters is the essence of being human. I'll always cherish her smile and carry a bit of her in my heart. Rhea, I'm still learning and I'm very grateful for all of it. Patty and Bo redefined the meaning of feeling welcome. The girls from Katy Hatchs, Linda, Cara and Margaret were good natured, a good laugh, and generous with their space and tea.

My family, though far away for most of this experience, were always close by when it came to offering support and lending an ear. If it weren't for them I would have never dreamed of starting this journey.

This research was funded by NSF (through the Center for Subsurface Sensing Imaging Systems under grant EEC-9986821) and the MIT Presidential Fellowship.

Contents

List of Figures

List of Tables

Chapter 1

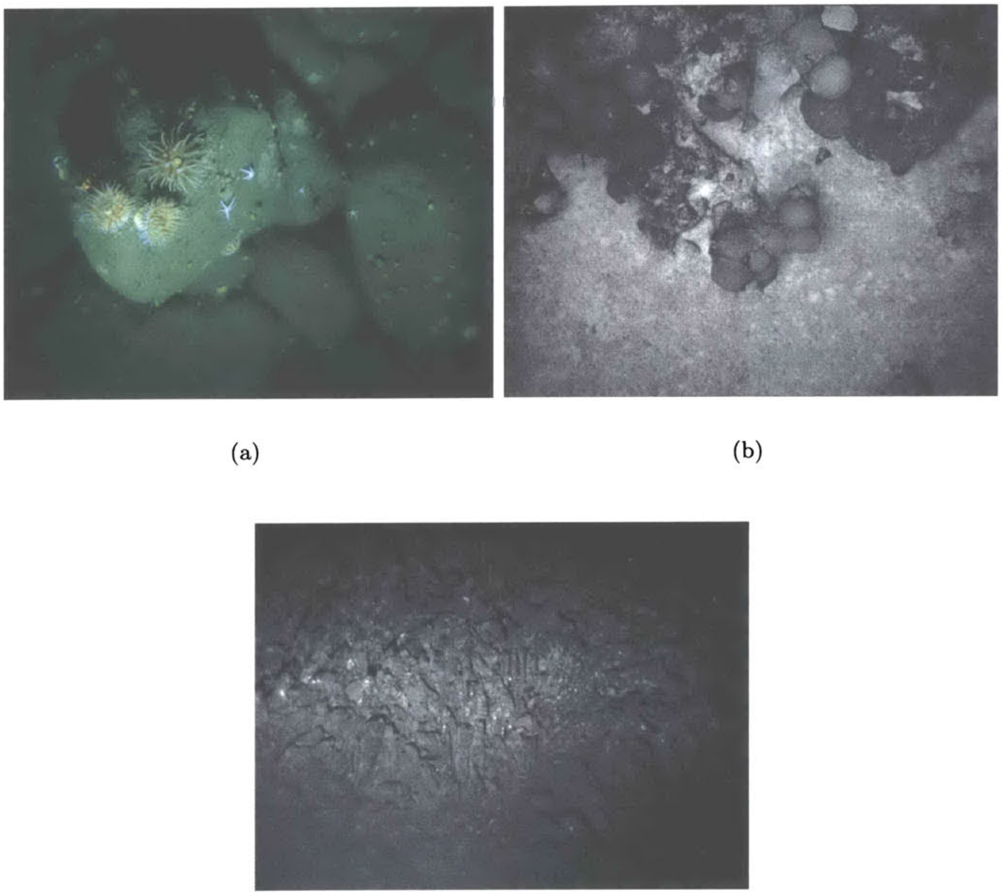
Introduction

1.1 Motivation

Optical imaging of the ocean floor offers scientists high level of detail and ease of interpretation. However, light underwater suffers from significant attenuation and backscatter, limiting the practical coverage of a single image to only a few square meters. For many scientific surveys, however, the area of interest is large, and can only be covered by hundreds or thousands of images acquired from a robotic vehicle or towed sled. Such surveys are required to study hydrothermal vents and spreading ridges in geology [130], ancient shipwrecks and settlements in archeology [4] [5], forensic studies of modern shipwrecks and airplane accidents [46] [79], and surveys of benthic ecosystems and species in biology [113] [32] [100] [111].

Generating a composite view by exploiting the redundancy in multiple overlapping images is usually the most practical and flexible way around this limitation. Recent years have seen significant advances in mosaicing [103] [102] [110] [71] [16] and full 3D reconstruction [25] [69] [95] [124] though most of these results are land based and do not address issues particular to underwater imaging.

The attenuation lengths (a drop by a $1/e$ factor in intensity) in water for the visible spectrum range from 5 m (red) to 25 m (blue-green) make the use of ambient lighting practical only for the first few tens of meters of depth. Thus most deep ocean vehicles carry out optical imaging surveys using their own light source. Apart from casting shadows that move

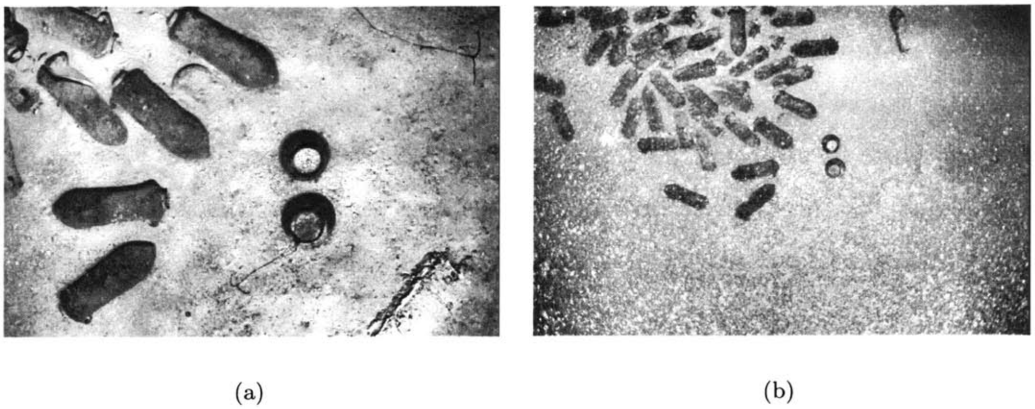


(a)

(b)

(c)

Figure 1-1: Sample images from AUV surveys. The strong falloff in lighting is typical of energy-limited vehicles. (a) Image from a boulder pile in Stellwagen Banks acquired with the SeaBED AUV. (b) Coral reef survey performed by the SeaBED. (c) Lava flow imaged by the Autonomous Benthic Explorer (ABE).



(a) (b)

Figure 1-2: A pair of images taken at different altitudes (a) 3.5 m and (b) 9.5 m. Light absorption and backscatter limit the altitude from which images can be acquired. Covering an area of interest may require hundreds or thousands of images.

across the scene as the vehicle moves, power and/or size limitations lead to lighting patterns that are far from uniform (Figure 1-1). Also with the advent of autonomous underwater vehicles (AUVs) for imaging surveys [130] [111] additional constraints are imposed by their limited energy budgets. AUV surveys are typically performed with strobed light sources rather than continuous lighting, and acquire low overlap imagery in order to preserve power and cover greater distances. Optical imaging from towed sleds can yield imagery with low or uncontrolled overlap, since cable dynamics can induce significant changes in camera position and orientation from frame to frame [46].

High dynamic range cameras are more robust to variable lighting (in terms of extracting and matching features between dark and bright areas). Image processing techniques such as adaptive histogram equalization [135] and specification [22] can partially compensate nonuniform lighting pattern to yield visually appealing images, though their effect on image registration (and similarity measures) is little understood. High power Light Emitting Diodes (LEDs) hold out the promise of efficient lighting through an array of LEDs that control beam pattern and spectral content [23]. Image processing techniques using multiple view points and/or lighting sources have the potential to reduce the effects of backscatter [62] and to exploit shadows to improve matching [98]. Such advances should eventually have a positive impact on underwater optical imaging. This thesis assumes that in its most general form, generating composite views underwater implies imagery acquired with low to moderate overlap, terrain relief, non-uniform lighting and unstructured surveys. Underwater mosaicing, although used successfully in some applications, imposes strong restrictions on scene geometry [36] [91]. Natural scenes are not necessarily planar and light attenuation limits the distance from which the scene can be imaged, such that parallax will be noticeable. As conditions deviate from the planar scene assumption, degradation and errors are often hard to quantify and understand, even though blending can produce a visually appealing representation that hides many inconsistencies. Ultimately, with large induced parallax, matching might fail or the transformed images become obviously distorted (Figure 1-3). In addition, actual measurements such as lengths and areas are desirable for scene features that span multiple images. Mosaicing cannot provide accurate metrology as is necessary to account for the scene relief when estimating camera poses and feature locations.

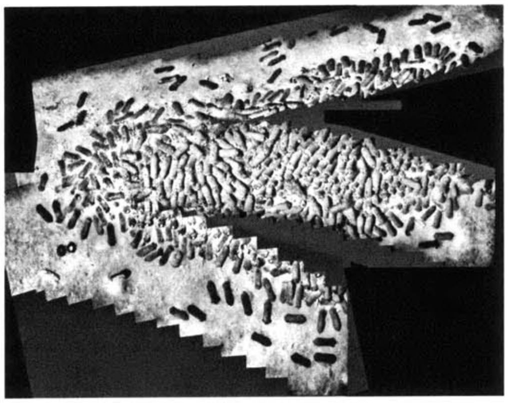


Figure 1-3: Example of the inability of mosaicing (under the assumption of planarity) to account for significant structure. The site of intereset is a Phoenician shipwreck composed of approximately 300 identical amphorae in a central mound that slopes of toward the edges. The 3D structure of the mound causes the amphorae to appear of different sizes when mosaiced.

Underwater vehicles are performing optical surveys of areas with significant structure [131] [111] [5]. There is also a growing interest in generating accurate and self-consistent composite views for measurement purposes and for tracking change through time [9]. There have been some efforts at 3D reconstructions [82] but they remain limited to small areas or artificial environments.

Prior to the work proposed herein, there was no practical, robust, and repeatable way of generating a reconstruction from underwater that combines hundreds to thousands of images acquired with moderate overlap, poor lighting, and possibly in an unstructured fashion. This thesis demonstrates large-area underwater 3D reconstruction by addressing all these issues with an effective image registration technique in a local to global framework.

1.2 Context

This thesis brings together aspects of underwater vehicle technology and of structure from motion. While the following chapters focus on the details of our approach, this chapter briefly reviews some background and context.

1.2.1 Underwater Imaging

In comparison to air and space, water is fairly opaque to light. The typical absorption and scattering lengths are on the order of a few meters [76], usually much less than the dimensions of the area to be imaged optically. There has been significant interest in understanding the behavior of light underwater, starting with Duntley's work [20] which collected over twenty years of experimental data concerning attenuation, scattering, radiance and irradiance as a function of wavelength, depth and water masses for both sunlight and collimated light. The unscattered residual radiant power P_r at distance r is given by

$$P_r = P_0 e^{-\alpha r} \tag{1.1}$$

where P_0 is the flux of the beam at the source. The spectral volume attenuation coefficient α has units of natural log per unit length and is frequency dependent. It is usually easier to visualize the *attenuation length* $1/\alpha$ at which $\approx 63\%$ of a beam of light has been attenuated. Light underwater is attenuated significantly through two main processes – absorption and scattering – such that $\alpha = a + s$, where a is the volume absorption coefficient and s is the total volume scattering coefficient. Absorption represents mainly the conversion of photon energy into heat and is a wavelength dependent phenomenon. In pure water the maximum attenuation lengths are on the order of 28 meters for wavelengths of 480 nm (blue-green) while the red end of the visible spectrum has attenuation lengths of 3-5 meters. Scattering is mostly independent of wavelength since it is produced mainly by particles that are large relative to the wavelength. For clear ocean water, scattering represents at most 60% of the attenuation (for blue-green light). For other wavelengths absorption plays a predominant role. In practice attenuation varies with location and depth, since changes in temperature, salinity, and biological activity significantly affect the properties of the medium.

Underwater optical imaging systems are significantly limited by the properties of the medium. Recent advances in hardware and image processing have allowed some improvements. Jaffe [50] [51] classified underwater imaging systems in terms of their effective range, camera to light separation and factor limiting the range. In general, configurations associated with a camera and a light-source nearby can acquire images at ranges of 1-2 attenuation lengths, limited by backscatter. By increasing camera to light separation the common volume of water between the field of view and the illumination source is reduced, reducing backscatter and extending the effective range to 2-3 attenuation lengths. Beyond that it is necessary to use more advanced systems that can sample more finely in time (range gated light pulses) or in space (synchronous scans). These approaches tend to be limited in power or contrast.

Even without attenuation the illumination toward the edges of an image drops by the fourth power of the off-axis angle ϕ [72]. This effect can be broken down into three contributing factors: Spherical spreading of a light from a point source increases the area in proportion to the square of the distance and the radiance diminishes in inverse proportion. Since the range at the off-axis angle ϕ is $1/\cos\phi$ greater than to the center, the radiance is reduced by a $\cos^2\phi$ factor. An additional factor of $\cos\phi$ comes from the foreshortening of the circular lens aperture as seen from ϕ off-axis. And the final $\cos\phi$ factor comes from the oblique angle at which off-axis rays strike the focal plane (Lambert's law). For example, for a 45° field of view lens, the illumination at the edge is reduced to 25% of the value at the center.

1.2.2 Vehicles

Underwater vehicles serve multiple purposes including surveying for scientific, commercial and military purposes, sample collection, underwater construction and inspection. Vehicles act as a sensor platform that brings the sensors within range for measurements of the ocean floor or water column. Manned vehicles such as Alvin, the MIRs, Nautilus and Shinkai carry humans to ocean depths and allow for direct observation and manipulation. Their mission duration is limited by life support systems and energy. An Unmanned vehicle can be tethered to the surface (usually to a ship) as a remotely operated vehicle (ROV) or untethered as an autonomous underwater vehicle (AUV). ROVs such as JASON II, Ventana and ROPOS are both powered and controlled from the surface. They offer fine positioning and the capability to manipulate their surroundings. As a sensing platform they are stable and can carry a large suite of sensors and lights since power is provided externally. They are best suited to work on areas of a few tens to hundreds of square meters since covering larger areas requires the support ship to move. AUVs such as ABE, Seabed, REMUS, Odyssey, FAU Explorer, Hugin, tend to be specifically designed as sensing platforms for surveys. They cover larger distances and follow fairly simple survey patterns, either sampling the water column or moving over the bottom while sensing (including cameras). They are limited by the energy they can carry in their batteries, which usually limits the power that goes into lighting.

This thesis uses data gathered with the WHOI's Seabed AUV, which was designed specifically for optical imaging of the ocean floor. Seabed is described in more detail in §1.2.4.

1.2.3 Navigation

The ability to estimate pose (position and orientation) underwater is critical in many tasks performed by underwater vehicles. Since electromagnetic waves do not penetrate beyond a few meters it is not possible to rely on fixes from a Global Positioning System (GPS) receiver. It is possible, however, to use sound in water to estimate position as well as a host of other navigation sensors such as inertial sensors, depth sensors, heading references, magnetometers, tilt sensors, and velocity logs.

For deployments where repeatability is important or where bounded error estimates are required regardless of deployment length, it is usually necessary to use an acoustic positioning system [74]. These systems can be classified according to the size of the baseline relative to wavelength. Long baseline (LBL) systems require deploying transponders at a scale comparable to the survey area. The travel times from the vehicle transponder to the LBL net are measured and the position of the vehicle can then be triangulated based on the (known) transponder positions. Ultra Short Baseline (USBL) and Short Baseline (SBL) systems are used primarily to track a vehicle (with an array mounted on the ship) or to home the vehicle onto a beacon (with an array mounted on the vehicle).

Often it is inconvenient to deploy an LBL net, and navigation must be dead-reckoned. Precise velocity measurements relative to the bottom are available from an acoustic Doppler Velocity Log (DVL) and Acoustic Doppler Current Profilers (ADCP) [99]. These instruments measure velocity along several acoustic beams based on the Doppler shift caused from backscatter elements in the water column and sea-floor. The velocities along the beams are expressed as sensor frame velocities. The conversion to world frame velocities requires rotating the velocities using orientation information and then integrating them to produce position estimates. Though navigation estimates produced by such systems tend to drift, the noise is normally very small and the drift is mostly due to unmodeled biases and heading errors.

Another enabling class of sensors comprises precise depth sensors based on the oscillations of a crystal subject to pressure. This helps constrain LBL solutions and can be merged into the DVL estimates.

1.2.4 The Seabed AUV

The Seabed AUV acquired the field data used in this thesis (Figure 1-4).The vehicle was designed for underwater imaging in mind. Seabed is capable of maneuvering at slow speed and passively stable in pitch and roll. The vehicles specifications are summarized in Table 1.1. The data used in this thesis was collected by Seabed using survey patterns preprogrammed as a mission and executed in dead-reckoning mode (xy position from integrating velocities of the DVL).

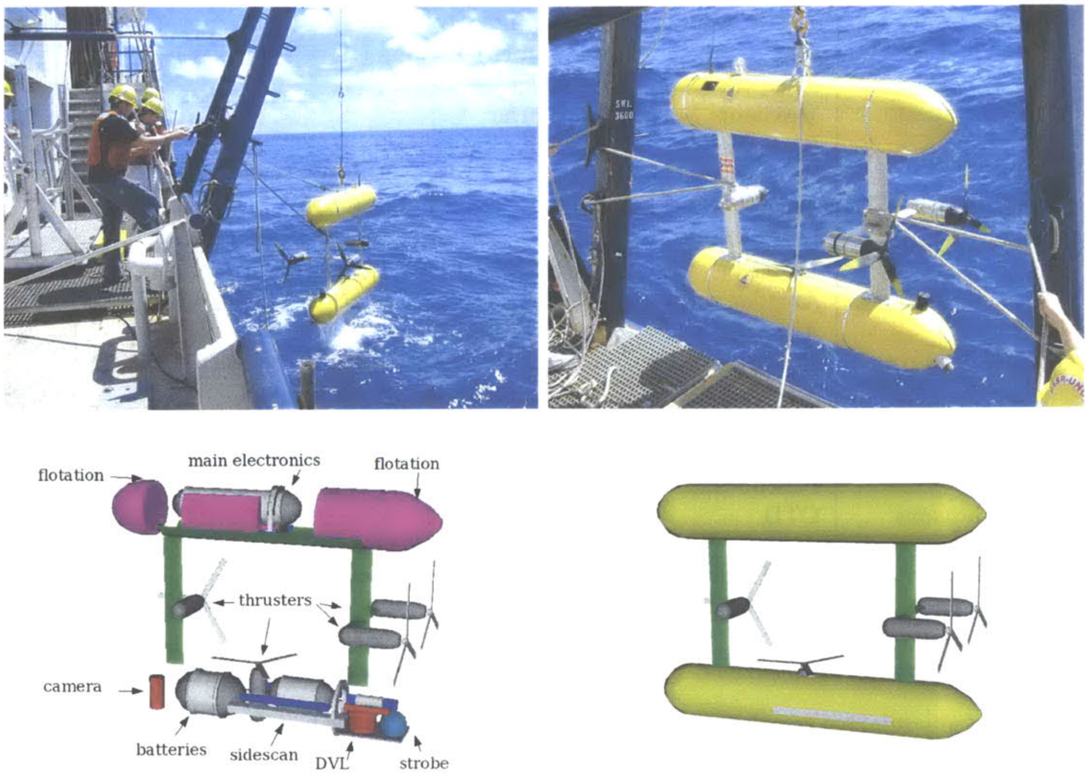


Figure 1-4: (top) The Seabed vehicle in the Bermuda 2002 cruise. (bottom) CAD views showing the vehicle with and without shells.

1.3 An overview of related work

We briefly present the context for this thesis in the fields of computer vision, mobile robotics and underwater mosaicing. Throughout the dissertation these and other references will be discussed in detail as the need arises.

1.3.1 Structure from Motion (SFM)

Given a scene viewed by a moving camera (or multiple cameras), structure from motion (SFM) attempts to recover the scene structure and the camera poses from the multiple views of the scene. The last decade has seen significant advances in the theoretical and practical understanding of multi-view geometry (for comprehensive treatments see the textbooks by Hartley and Zisserman [39] and Faugeras and Long [25]) which has led to several successful

Vehicle	Depth rating	2000 meters
	Size	2.0 m (L) × 1.5 m (H) × 1.5 m (W)
	Mass	200 kg
	Maximum Speed	1.2 m/s
	Batteries	2 kWh Li-ion pack
	Propulsion	four 150 W brushless DC thrusters
Navigation	Attitude+Heading	Tilt ±0.5°, Compass ±2°
	Depth	Paroscientific pressure sensor, 0.01%
	Velocity	RDI Navigator ADCP $\pm 1 - 2mm/s$
	Angular rates	Crossbow 3-axis gyro
	Altitude	RDI Navigator
Optical Imaging	Camera	Pixelfly 12bit 1280×1024 color or BW CCD
	Lighting	one 200 Ws strobe
	Separation	1m between camera and light
Acoustic Imaging	Sidescan sonar	MST 300 kHz (300 m depth rating)
	Pencilbeam sonar	Imagenex 881 675 kHz
Other Sensors	CTD	Seabird 37SBI

Table 1.1: Summary of the Seabed AUV specifications.

implementations of vision-based reconstructions. Vision systems can produce a wealth of measurements relative to other sensors. One challenging issue is to reliably relate two images that view the same scene. A key development has been the adoption of robust estimation techniques such as Random Sample Consensus (RANSAC) [28] that can automatically classify data points into inliers and outliers based on their ability to explain the rest of the data. Recently, several feature descriptors suitable for wide-baseline matching [127] [73] [7][68] have enabled SFM solutions to challenging image sets and are relevant to underwater applications.

Beardsley et al. [8] introduced a practical sequential structure from motion algorithm that has served as inspiration for many later improvements. Pollefeys demonstrated a complete system for SFM recovery from video sequences [93] and Fitzgibbon and Zisserman [29] addressed loop closure and error drift by dividing the input sequence of images into short subsequences, in a local to global framework.

The optimal SFM solution attempts to solve for all camera poses and all 3D features simultaneously. Given the nonlinear projection of 3D features into image measurements, this problem is solved as a large nonlinear minimization known as bundle adjustment [126]. The SFM problem is sparse in the sense that each measurement (projection of a 3D feature point onto an imaging plane) depends only on a 3D feature point and on the camera viewing it. This sparsity can be exploited to significantly reduce computational complexity.

Typically, SFM does not rely on motion information to produce estimates of structure and motion. While concentrating on the potential of image-based reconstructions, pure SFM will suffer from loss of scale (due to projection) and is prone to ambiguities that are not always resolved by image data alone [121].

1.3.2 Simultaneous Localization and Mapping (SLAM)

SLAM seeks to recover an estimate of the environment (map) and robot motion by use of both sensors (such as laser rangefinders and sonars) and motion instruments (inertial units, heading references, odometry, GPS receivers, etc) [114] [27]. SLAM should run in realtime on robots, which suggests a recursive filtering approach. Most direct implementations rely on the Kalman Filter as a framework for estimating the state of the robot and the environment [114]. Limitations in the scalability of the representation of state and uncertainty have lead to approximations and alternative representations such as submaps and hierarchical methods [14] [3], covariance intersection [54], particle filters [77], and sparse extended information filters [119].

Typically features in the environment are sensed with both range and bearing information, though bearing or range only SLAM [19][83] and vision based SLAM [18] have recently drawn some attention.

1.3.3 Underwater Mosaicing

Vision-based navigation and station keeping close to the sea-floor [30] [31] [81] [36] has served as a motivation for underwater mosaicing. A limited form of global alignment is considered in [30] [31] through the 2D topology of image positions and a 'smoothing' stage to distribute errors in the placement of images in the mosaic. Real-time constraints force the homographies that relate overlapping imagery to be pure translations, sufficient for local navigation but inadequate for an accurate rendition of the sea-floor. Registration is based on matching borders of zones with different textures. Underwater imaging implies changing lighting conditions that destroy the brightness constancy constraint (BCC), which is the key assumption in most direct (intensity based) registration methods [42]. In [81] a modified BCC approximates light attenuation underwater but this method has not been proved for low overlap imagery and for unstructured terrain.

Gracias and Santos-Victor [36] presented a global alignment solution for underwater mosaicing with excellent results for video-based imagery over an area of approximately 50 m^2. At video rates the relatively slow speed of underwater vehicles yields high overlap, narrow baseline imagery. This simplifies the matching stage by assuming that translation is the dominant motion between consecutive frames (correlation is used to match feature points described by a window of fixed size). Even though their global mosaic is constructed with a subset of images with significant inter-image motion, the feature matching is performed with high overlap (the homography between two images with low overlap is calculated as the composition of video rate homographies). It is not clear how this technique would fare when only low overlap imagery is available. In addition, their method does not account for lens distortion, which can have a significant impact at larger scales. Given that the main objective of these approaches is vision-based navigation, distortions in the mosaic are not critical as long as the vehicle is able to register its current view to the mosaic. Pizarro and Singh [91] addressed the large-area mosaicing problem with low overlap under the assumption of planarity. In the presence of 3D structure unavoidable distortions occur.

Mosaicing assumes constraints on camera motion or scenery to merge several images into a single view, effectively increasing the camera field of view without sacrificing resolution. The key assumption for a distortion free mosaic is that the images come from an ideal camera (with compensated lens distortion) rotating around its projection center, or that the scene is planar [116] [25]. In either case, camera motion will not induce parallax; therefore no 3D effects are involved and the transformation between views can then be correctly described by a 2D homography. These assumptions often do not hold in underwater applications since light attenuation and backscatter rule out the traditional land-based approach of acquiring distant, nearly orthographic imagery. Underwater mosaics of scenes exhibiting significant 3D structure do not satisfy the assumptions for mosaicing and usually contain obvious distortions.

1.3.4 Underwater 3D Reconstruction

In contrast to mosaicing, the information from multiple underwater views can be used to extract structure and motion estimates using ideas from SFM and photogrammetry [112]. We propose that when dealing with a translating camera over non-planar surfaces, recovering 3D structure is the proper approach to providing a composite global view of an area of interest. The same challenges seen in mosaicing underwater apply to SFM underwater with the added requirement that scene points must be imaged at least twice to produce a roughly uniform distribution of reconstructed feature points through triangulation (50% overlap in the temporal image sequence). These techniques are considerably more complex than mosaicing: even for land-based applications (with high overlap, structured motion and uniform lighting) consistency at large scales can not be guaranteed unless other sensors are available. Some promising work has gone into 3D image reconstruction underwater [80] using a stereo-rig with high overlap imagery in a controlled environment.

1.3.5 Relation to thesis

Underwater vehicle technology is advancing at a rapid pace. Although it is possible to rely on external references for positioning (such as triangulation by long baseline acoustic networks) there is a growing interest in performing surveys without positioning networks, in order to simplify deployments, enable fast exploration and reduce costs. Currently these surveys are performed by dead reckoning, which can result in deviations from the intended survey due to accumulated errors and small biases. This thesis focuses on generating a 3D reconstruction from imagery and navigation data acquired during a dead reckoned survey. We assume that all data is available and that we can use batch processing techniques. The temporal sequence is used as an ordering device and to extract relative pose information between successive cameras. Our algorithm constructs an initial guess of the layout of cameras and structure that can be optimized to best explain the image and instrument-based measurements. The general problem of mapping and localizing a robot can be addressed in a Simultaneous Mapping and Localization (SLAM) framework where the vehicle improves upon dead-reckonned estimates by sensing the environment and estimating both its pose and the state of the environment. The focus of SLAM is to enable robots to operate in an initially unknown environment, which leads to real-time requirements and recursive filtering implementations. Although this thesis uses some SLAM concepts, it concentrates on basic challenges facing the realization of robust underwater SFM algorithms.

1.4 Thesis Statement

Robust wide-baseline, feature-based relative pose approaches combined with local-to-global mapping techniques that use navigation information can recover scene structure and camera pose from a large set of underwater images, and provide uncertainty estimates for structure and motion.

1.4.1 Objectives

The basic thesis objective is to enable large area 3D reconstruction from underwater imagery acquired with robotic vehicles. More precisely, given a sequence of calibrated and lens distortion-compensated images acquired from a robotic vehicle, we seek techniques to generate a 3D reconstruction using a sound theoretical foundation that can

- reliably extract and match features from underwater imagery,
- use navigation and sensor data to aid and constrain the reconstruction,
- generate motion and structure estimates that are globally consistent,
- provide uncertainty estimates for motion and structure,
- scale to hundreds or thousands of images and larger areas,
- employ largely automatic processing, and
- yield additional benefits such as providing calibration information for other vehicle instruments.

1.4.2 Contributions

The main contributions of this thesis are:

- This is the first demonstration of large area 3D reconstructions of underwater environments. This thesis demonstrates the integration of several techniques in computer vision and SLAM to provide reliable estimates of large underwater scenes.

- We present a robust two view and three view egomotion estimation method for calibrated and instrumented imaging platforms. At the core of the structure from motion algorithm is a robust essential matrix estimation and resection that takes advantage of camera calibration and pose sensors to constrain matching, to provide priors for optimization of pose and structure, and to disambiguate vision-based estimates.

- We validated our results and approach with ground truth for pose and structure. Large scale results are self-consistent, and are shown to be close to ground truth where it is available.

- We also present a compensation procedure for sensor bias. Our methodology relies on self-consistency in the reconstruction to identify and compensate for sensor bias.

1.4.3 Assumptions and Restrictions

This thesis is grounded within current oceanographic AUV technology. This implies several assumptions and restrictions that shaped our choices and priorities throughout this work:

- Simple camera and lighting configuration. We assume an imaging configuration of one calibrated monochrome camera and one light source. The field of view (FOV) is limited to approximately 45° due to attenuation and lighting. We assume lighting can vary significantly for power-limited surveys and thus require a similarity measure that is robust to changes in lighting. This lighting assumption was relaxed for daytime shallow water surveys (significant ambient light) and for the tank tests where two lights minimized the effect of shadows.

- Calibrated camera. This allows us to work with normalized coordinates and to define the Essential matrix (5 DOF) rather than the Fundamental Matrix (7 DOF). When imaging a planar scene the fundamental matrix has infinite solutions and therefore cannot guide the correspondence search. The essential matrix has only up to three possible solutions in the case of a planar scene, simplifying the decision process.

- Calibrated imaging platform. Position and angular offsets of navigation sensors are known well enough that their measurements provide a useful prior to the image matching stage. For instance, an initial essential matrix (and associated uncertainty) can be estimated from navigation and attitude data (and their uncertainty). This translates into a search along an epipolar band for correspondences. Prior knowledge of scene depth limits the search to a segment of the band. The navigation-based prior can also constrain refinements of pose and structure when used in maximum a posteriori estimation, in particular providing scale information that would otherwise be lost by the image formation process. In addition pose priors are used to disambiguate situations where multiple structure and motion solutions explain the imagery.

- Large Area Survey. A set of images that covers an area of hundreds of square meters. Given the limitations of optical imaging underwater (attenuation, backscatter, lighting, FOV) this translates into a set of hundreds to thousands of images.

- Unstructured Survey. A large area survey is typically performed as a 'mow the lawn' pattern in the horizontal plane. While surveying the vehicle controls its depth to keep an approximately constant distance from the bottom. A survey consists of a sequence of overlapping images acquired along multiple parallel tracklines. Image overlap along a trackline is set by the camera FOV, vehicle altitude, speed, and strobe rate. This overlap has to be at least 50% in order to have image features in more than one image and therefore allow triangulation. Overlap between parallel tracklines is set by camera FOV, vehicle altitude, and navigation precision. Overlap between tracklines only has to be sufficient to recognize corresponding points, in practice 10-20%. A structured survey presents two distinct matching situations:

  - Along trackline. In temporally adjacent images a feature should be matched with angular displacements of approximately 20° (less than half the FOV). Similarity-based matching can provide enough correct putative matches for robust estimation (using some form of RANSAC). Navigation data and scene structure can act as rough priors to constrain possible matches.

- Across tracklines. Spatially but not temporally adjacent images can present matching features with changes in viewpoint almost up to the FOV (40° approximately). Relative pose uncertainty can be significantly larger than in the temporal sequence and it is necessary to propose putative matches under wide baseline conditions and significant lighting changes.

However, dead-reckoned navigation often results in surveys in which the actual trajectory is far from the preprogrammed pattern (*i.e.*, an unstructured survey). The reconstruction algorithm must be able to recognize loop closures and overlapping sections even if these are not initially suggested by the navigation estimates.

1.4.4 Outline of Methods

Our methodology (Figure 1-5) takes a local-to-global approach inspired by mosaicing [47] and the work of Fitzgibbon and Zisserman [29], and Zhang and Shan [134] but takes advantage of navigation and attitude information. Local subsequences are derived independently, then registered in a global frame for bundle adjustment. Our approach seems more suitable than pure sequential methods [8] because in an underwater survey each 3D feature appears only in a few images making the global solution more like a series of weakly correlated local solutions.

The core of the algorithm is a robust estimator of relative pose from a pair and triplets of images. Prior pose uncertainty and scene depth constrain possible correspondences, and affine invariant descriptors propose putative matches that are then refined into inliers and outliers using a six-point algorithm for essential matrix estimation.

We generate local structure, the submap, by using sequential methods on the temporal sequence. We show that it is computationally advantageous to keep submap sizes limited to a bounded number of features.

To initialize bundle adjustment we require an estimate of the global poses of the cameras (by determining the global poses of the submaps). The problem is cast as a graph, where nodes in the graph correspond to submap local coordinate frames and edges in the graph correspond to the relative transformation between submap frames. Most of the work is

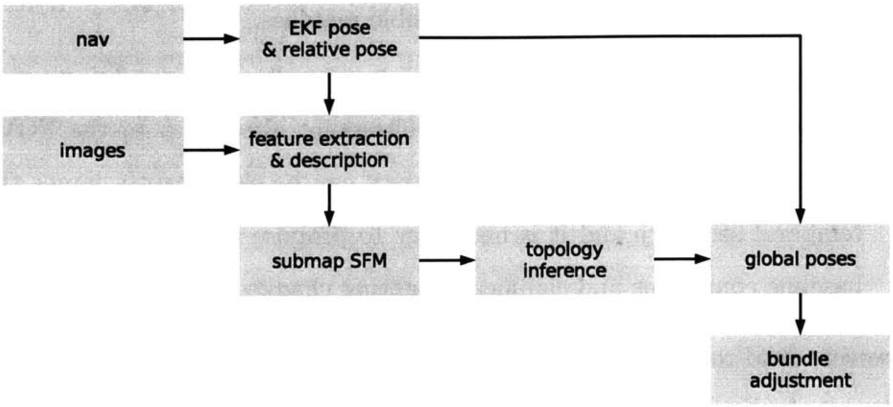


Figure 1-5: Flowchart of structure and motion recovery from underwater imagery. An image sequence is processed into short submaps of structure and motion aided by navigation information. Submaps are then matched to infer and refine additional spatial constraints (such as loop closures and parallel tracklines). An initial guess of poses and structure in a global frame is then used to perform a final bundle adjustment.

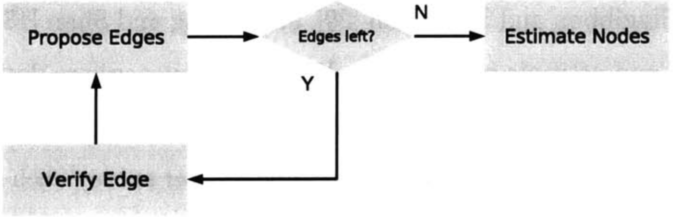


Figure 1-6: Consistent estimates of nodes (the submap frames in a global reference frame) depend on establishing additional edges between nodes. These can be proposed and verified entirely in 'relative space' based on the composition of edges before calculating the node frames.

done using relative transformations, delaying the representation of poses in a global frame (Figure 1-6). This is similar to the Atlas framework [13] and offers increased robustness by avoiding an early commitment to a particular topology.

We frame global pose estimation as an optimization problem, where we determine the poses that best explain all the relative pose measurements and are close to the navigation. New edges are proposed by using the accumulated uncertainty over multiple paths to decide which edge to verify next.

1.5 Dissertation Structure

The rest of this dissertation presents the theory, methods, results and validation of the thesis. The following chapters cover feature extraction and description (Chapter 2), robust two view relative pose estimation (Chapter 3), submap generation (Chapter 4), topology exploration and local to global registration (Chapter 5). The last part of the thesis (Chapter 6) presents results from a coral reef survey. This framework is validated by tank experiments with ground truth. Finally (chapter 7) we offer concluding thoughts and suggestions for future work.

Chapter 2

Image matching based on similarity

2.1 Overview

Identifying common scene elements in two or more images forms the basis of many computer vision tasks such as object recognition, tracking, pose estimation and structure recovery [33]. The image of a scene is dependent on the pose of the camera (Appendix A) and multiple images of the same scene can provide information on the relative motion of the camera as well as of the scene structure. *Image registration* attempts to bring images into alignment by identifying common elements and yielding a transformation that maps an image (or parts of it) onto another image. Images can be related to each other by utilizing the entire image (direct methods) or by concentrating on specific regions that hold information (feature based methods).

Direct methods [10] [49] align images based on discrepancies in overall intensities, assuming that some form of the Brightness Constancy Constraint (BCC) [42] holds. Direct methods are unsuitable to underwater applications because of moving, non-uniform lighting effects and moving shadows.

Feature-based methods [38] [107] abstract regions of interest into projections of geometric entities such as points and lines which can then be matched across images. Such approaches provide a greater degree of robustness to occlusion, changes in illumination and effects associated with large parallax [122]. In addition, structure and motion estimates can be formulated relatively simply from the projection of geometric features, which leads to

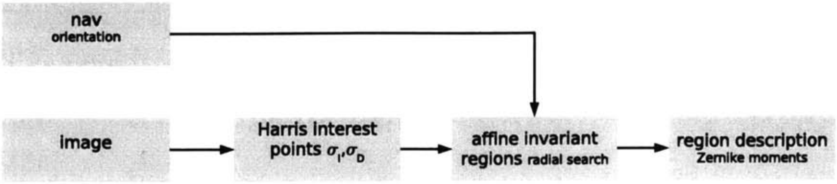


Figure 2-1: Overview of the feature extraction process

efficient robust estimation algorithms. Feature-based methods require matching features to relate images. Typically interest points are detected and the image region around the point is used to describe the feature under the assumption that the same interest point viewed in another image will lie within a similar neighborhood. Matching in narrow baseline applications is traditionally performed with an intensity-based similarity measure between fixed-shape window image patches centered around feature points. In its simplest and most common form the description of the feature point is the image patch around it, and the similarity measure is usually some variant of the sum of squared differences or cross-correlation[8]. This approach is effective when inter-image motion is small relative to the depth of the scene.

Under more general imaging conditions, changes in view point will result in the neighborhood boundaries deforming under perspective projection (*e.g.* a circle in one image will appear as an ellipse in another). In wide-baseline situations the local image deformations cannot be realistically approximated by translation, rotation and scaling. These changes to feature appearance can often be modeled locally as affine transformations. Matching features in the presence of such changes requires compensating for the motion (with prior knowledge) or using a description and similarity measure that is invariant to such transformations. Our approach uses a mixture of compensation and invariance to represent features, by using attitude sensor data to compensate for changes in orientation and extracting features in an affine invariant manner (Figure 2-1).

Our approach to relating images has four distinct stages:

- Feature detection. We relate images using a feature-based approach under wide-baseline imaging conditions with changing illumination and unknown scene structure. A modified Harris corner detector [38] yields interest points by selecting local maxima of the smaller eigenvalue of the second moment matrix.

- Feature extraction. We determine a neighborhood around each interest point that is invariant to affine geometric transformations using a modified version of the method proposed by Tuytelaars [127]. In essence, we sample the image neighborhood along lines radiating from the interest point. For each line we select the extrema of an affine invariant function (maximum difference in intensities between the interest point and points along the ray). The set of these maximal points defines the boundary of a region that can be extracted under affine geometric transformations. This region is approximated with an elliptical neighborhood which is then mapped onto the unit circle. To increase discriminating power, a second neighborhood twice as large as the first is also mapped onto a unit circle. These circular patches are normalized for affine photometric invariance (demeaned and normalized by their energy content so that linear changes in the intensity values do not affect the normalized appearance of a patch).

- Feature description. Moment-based descriptors [75] have shown promise in describing image regions for matching purposes. We chose to use Zernike moments as descriptors as they are compact (generated from an orthogonal complex polynomials) and highly discriminating [60]. Typical applications use only the magnitude of Zernike moments as this provides rotational invariance, but we pre-compensate for orientation using attitude sensors, and can therefore utilize the full complex moments.

- Feature matching. We derive the proper weighting of the Zernike moments such that the dot product of the vector of weighted moments approximates the correlation score for the original patches (warped into a disc).

2.2 Matching requirements

A typical survey configuration for SFM uses a downward-looking camera from an approximately level platform with a field of view of $\sim 45^\circ$; for images acquired along the temporal sequence feature points have to be imaged at least twice for two view reconstruction or three

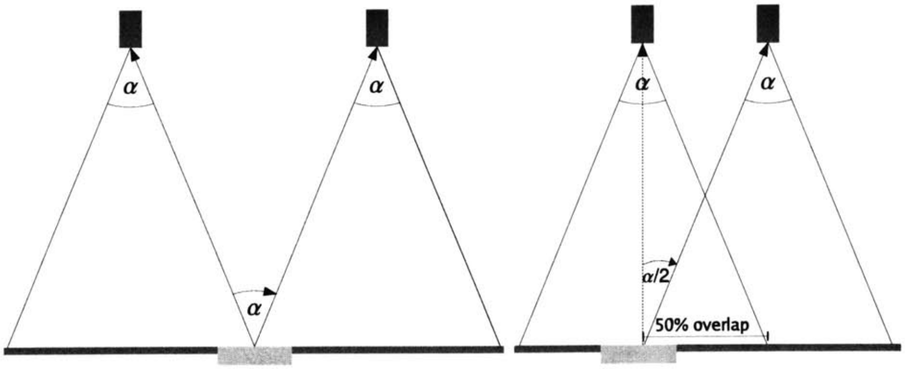


Figure 2-2: Range of viewing angles for an image patch (light gray) given the camera field of view (α).

times for resection. This implies image overlap of at least 50% and 67% respectively (Figure 2-2). Under these conditions image-based reconstruction falls somewhere in between the extremes of narrow-baseline and general wide-baseline. For a field of view of 45°, the view point to a surface patch will change by 22.5° and 15° for temporally adjacent images which is close to the breakdown point of reliable matching based on correlation windows (§2.6). To match images that are not temporally adjacent (*e.g.* images across tracklines), the worst case scenario implies that the view point to the surface patch will change by 45°. This requires detecting and extracting features in a manner robust to view point changes. We note that utilizing full affine invariants, even though appropriate, comes at the price of added computational cost and lower stability.

2.3 Interest points

Describing features in a way that is invariant to expected geometric and photometric variations is important for successful matching. Even more important perhaps is the ability to localize a consistent set of interest points in two overlapping images. We choose to detect features with a modified version of the Harris interest point detector [38] since it has been shown to be effective in detecting the same interest point in the presence of rotation and moderate scale changes [107]. Figure 2-3 shows a typical set of Harris feature points for an overlapping pair of images.

Interest points are determined from the second moment matrix, describing the curvature of the autocorrelation function in the neighborhood of a point $\mathbf{x}$:

$$\mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 g(\sigma_I) * \begin{bmatrix} L_x^2(\mathbf{x}; \sigma_D) & L_x L_y(\mathbf{x}; \sigma_D) \\ L_x L_y(\mathbf{x}; \sigma_D) & L_y^2(\mathbf{x}; \sigma_D) \end{bmatrix} \tag{2.1}$$

where $*$ is a convolution operator, σ_I is the integration scale, σ_D the derivative scale, $g(\sigma)$ the Gaussian of standard deviation σ, and L the image I smoothed by a Gaussian, such that L_x represents a smoothed derivative:

$$L_x(\mathbf{x}; \sigma_D) = \frac{\partial}{\partial x} g(\sigma_D) * I(\mathbf{x}) \tag{2.2}$$

The second moment matrix offers a description of local image structure at a given scale – an interest point or a corner point will have a μ with two positive eigenvalues (significant changes in intensity in any direction); an ideal edge will present one positive and one zero eigenvalue while a perfectly uniform area will have two zero eigenvalues.

The σ_I and σ_D define the scale at which features are extracted. A characteristic scale can be associated with features by processing with multiple values of σ and looking for the extrema of the scale-function (such as the Laplacian of Gaussian) [107] [63]. For our application we assume that vehicles control their altitude to be approximately constant. Thus the same feature should be observed at roughly the same scale and a multi-scale approach is not necessary. To summarize, our approach extracts regions in an affine invariant manner which offers robustness to modest scale changes.

2.4 Feature Extraction

In order to match interest points based on their appearance, we extract the neighborhood around the interest point as the feature. For narrow baseline applications a region of fixed size is extracted around an interest point and a similarity based measure (e.g. correlation) is used to compare features. This is acceptable because the shape of the region does not change significantly between similar viewpoints.

For wide baseline situations, in which the view point changes noticeably, the shape of the neighborhood around an interest point will also change projectively and it becomes

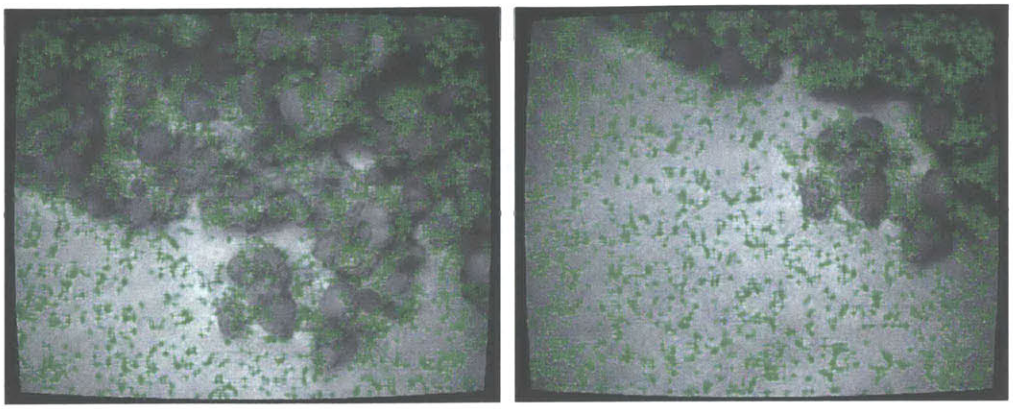


Figure 2-3: Two overlapping images and the Harris interest points for $\sigma_I = 4$ and $\sigma_D = 2$. 2000 interest points detected per image. To have a roughly uniform distribution the image is subdivided into 25 non overlapping regions (a 5×5 grid) and in each region the 80 interest points with highest curvature are selected.

important to extract the neighborhood in a manner that is invariant (or neraly invariant) with respect to the viewpoint.

In recent years several approaches have been proposed to invariant feature extraction. Some are specifically tailored for planar surfaces [97] [118]. These are not particularly useful for unstructured natural terrain. Baumberg [7] showed a practical approach that iteratively modifies the shape of the region to make the second moment matrix isotropic. Several modifications have been proposed to this idea [73] that optimize over closely coupled parameters of scale, shape and localization.

Matas [68] proposed the use of extremal regions, which are connected regions that have a persistent boundary when varying an intensity threshold (*e.g.* regions with a border or in which the intensity is significantly different than the surrounding intensities). This method extracts regions independently of interest points and is particularly suitable for images with multiple, distinct objects and high contrast.

Tuytelaars and Van Gool [127] proposed finding the region border by detecting the extrema of an affine invariant function along rays emanating from points of local extremum of intensity. They use these samples along the border to fit an ellipse that defines the region to be extracted.

The approaches proposed by Matas and by Tuytelaars have some similarities, though Matas' method can yield more complex regions, and the Tuytelaars method can distinguish

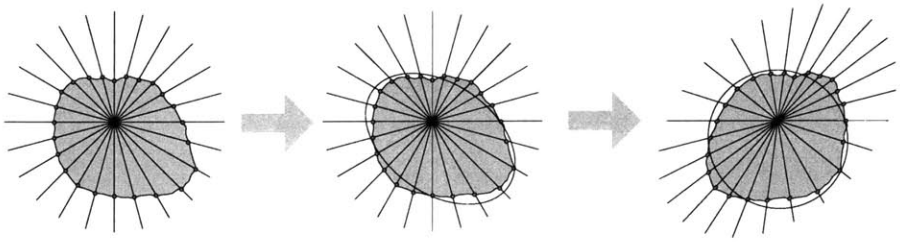


Figure 2-4: Steps in determining an affine invariant region. From left to right, boundary points determined along rays, an ellipse approximates the boundary points using the method of moments, and the elliptical region is mapped onto a unit disc.

regions with less contrast.

2.4.1 Tuytelaars's affine invariant neighborhood

We determine a neighborhood around each interest point that is invariant to affine geometric transformations using a modified version of the method proposed by Tuytelaars [127] [128] (Figure 2-4). The original method defines an affine invariant region around an intensity extreme point by determining affine invariant points along rays radiating from the intensity extremum. The boundary point associated with a ray corresponds to the extremum of an affine invariant function that can be related to the presence of a boundary (Figure 2-5 and 2-6). The boundary points along the rays $r_{invariant}(\theta)$ are given by

$$r_{invariant}(\theta) = \arg_r \max |f(r, \theta) - f_o| \tag{2.3}$$

where f_o is the extremum of intensity and $f(r, \theta)$ are the image values considered in polar coordinates. This region is extracted in an affine invariant manner in the sense that an affine transformation will 'stretch' the individual rays but the boundary points should remain recognizable since points that form a ray remain in a ray when affinely transformed (by definition an affine transformation preserves colinearity and we assume that the any translation is accounted for by keeping track of the interest point).

For natural scenes few interest points correspond to sharp corners of planar surfaces. Instead they are generally blob-like features at different scales. By using rays radiating from the interest point instead of an intensity extremum, the matching procedure is simplified since the feature is well localized. In essence, we sample the neighborhood along

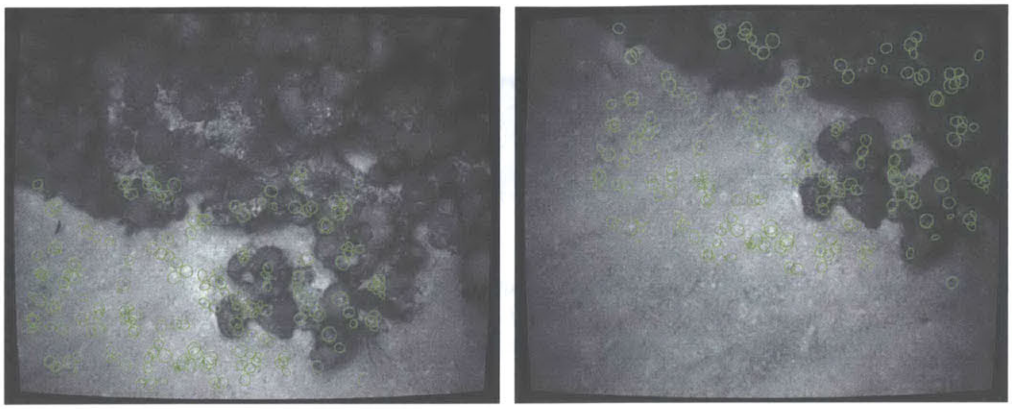


Figure 2-5: Affine invariant regions extracted using a modified version of the method proposed by Tuytelaars. Only regions that are found in correspondence are shown.

lines radiating from the interest point. Our current implementation uses a radius of 25 pixels and samples every 6 degrees (for a total of 60 lines). For each line the boundary point corresponds to the maximum difference in intensities between the intensity extremum nearest to the interest point and points along the ray.

The set of maximal points is approximated with an elliptical neighborhood by using the method of moments where the samples along the boundary are placed on an ellipse that has the same second moment matrix as the original samples. This elliptical region is then mapped onto the unit circle W. In practice the polar representation used to determine the boundary is resampled so that the boundary points have the same radius instead of applying a 2D affine transformation to the region. The canonical form of the region is stored as a polar representation with resampled radii. This representation is particularly convenient when the description of the region is based on Zernike moments since the basis functions are presented more compactly in polar form (§2.5.1).

The sampling along the ray provides robustness to changes in scale, since the boundary point should still be detectable as long as it falls within the search radius. Tuytelaars and Van Gool [127] suggest that the actual ellipse used be twice the size of the one derived in this manner. This increases the discriminating power by including image information outside the region. However, this comes at a cost since there is a greater chance that the appearance of the expanded ellipse might change significantly due to non-planarity.

To obtain some robustness to changes in lighting, prior to calculating descriptors of the

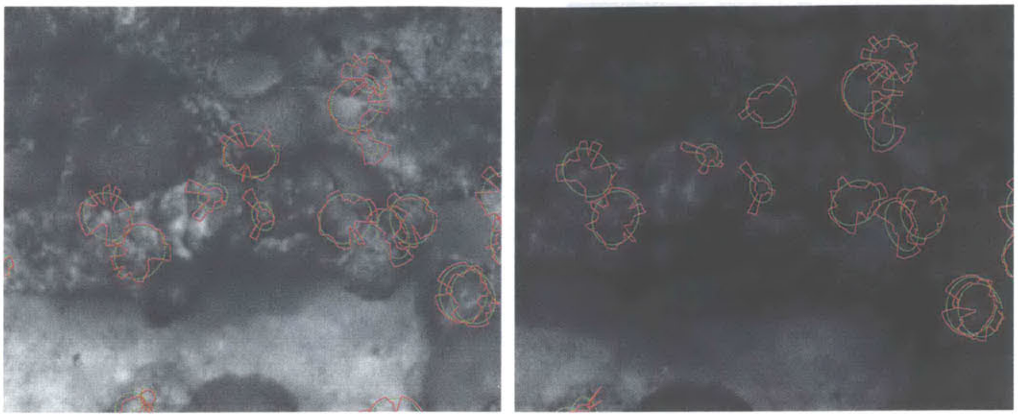


Figure 2-6: Detail of some of the extracted regions in Figure 2-5. The actual border samples are connected with red lines. The elliptical region that approximates the border samples is shown in green.

patch W the resampled intensities $f(x,y)$ are de-meaned and normalized by the energy content over the patch:

$$N(f(x,y)) = \frac{(f(x,y) - \bar{f}_W)}{\sqrt{\sum_{i,j}(f(x+i,y+j) - \bar{f}_W)^2}} \tag{2.4}$$

where $\bar{f}_W$ is the mean of $f(x,y)$ over the patch W. The normalized patch satisfies

$$N(f(x,y)) = N(af(x,y) + b) \tag{2.5}$$

effectively providing invariance to affine changes in intensity. Figure 2-6 illustrates several matches despite significant lighting changes between extracted regions.

2.4.2 Orientation normalization

The navigation instruments provide attitude information that can simplify the description and matching of features. For example, normalized correlation as a feature point similarity metric fails in the presence of modest rotations (more than a few degrees) between an image pair I and I'. It is possible to use descriptors that are invariant to rotations at the price of less discrimination. However, knowledge of 3D orientation for camera frames c and c' in a fixed reference frame w allows for normalization of orientation viewpoint effects via a

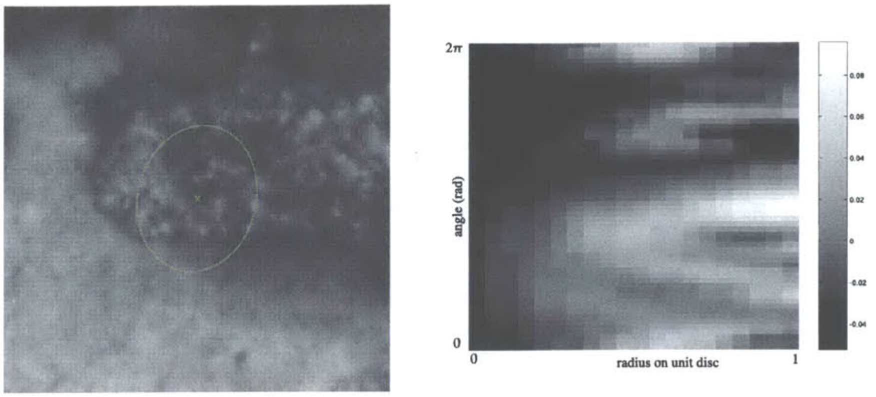


Figure 2-7: An example of a feature and its normalized polar representation (angle vs radius).

homography.

The infinite homography, H_∞, defined as [39]

$$\mathrm{H}_\infty = \mathrm{K} \cdot {}^{b}_{a}\mathrm{R} \cdot \mathrm{K}^{-1} \tag{2.6}$$

where ${}^{a}_{b}\mathrm{R}$ is the orthonormal rotation matrix from frame b to frame a and K is the camera calibration matrix (A.1), warps an image taken from camera orientation a into an image taken from camera orientation b. This warp is exact and independent of scene structure; there is no scene induced parallax between viewpoints a and b, because a and b share the same projective center.

Given 3D camera rotation matrices ${}^{w}_{c}\mathrm{R}$ and ${}^{w}_{c'}\mathrm{R}$ generated from vehicle orientation measurements, we can warp images I and I' each into a canonical viewpoint coordinate frame oriented parallel with frame w (e.g. the warped images correspond to a camera coordinate frame x, y, z oriented with North, East, Down).

2.5 Feature Description

Instead of directly comparing intensities of the affine-invariant image patches, we transform the patches into a compact descriptor vector. By describing patches with small vectors the cost of comparing patches and the storage requirements are significantly reduced. Typically a patch will have $\mathcal{O}(1000)$ pixels while a descriptor vector will have one or two orders of magnitude fewer terms. High frequency terms are not captured by the low order coefficients which provides a representation with some robustness to noise.

Image patches have been described by differential [106] [105] and moment invariants [117] [48]. Differential invariants are constructed from combinations of intensity derivatives that are constant to some geometric and radiometric transformations such as translation, rotation, scaling and affine brightness changes. Moment invariants can be constructed from nonlinear combinations of geometric moments (the projection of the image patch $f(x, y)$ onto the basis set of monomials $x^p y^q$). Since the basis set for this projection is not orthogonal, these invariant moments contain redundant information and have a limited ability to represent an image in the presence of noise. Orthogonal moments based on orthogonal polynomials such as Zernike moments have been shown to be invariant to some linear operations, have superior reconstruction capabilities in the presence of noise, and low redundancy compared to other moment representations [117] [58] [60].

2.5.1 Zernike Moments

Zernike moments are derived from Zernike polynomials, which form an orthogonal basis over the interior of the unit circle, i.e. $x^2 + y^2 = 1$ [58]. If we denote the set of polynomials of order n and repetition m by $V_{nm}(x, y)$, then since these polynomials are complex, and their form is usually expressed as:

$$V_{nm}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{jm\theta} \tag{2.7}$$

with n a positive or zero integer, m an integer such that $n - |m|$ is even, and $|m| \leq n$. We've also defined polar coordinates $\rho = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$. Note $V^*_{nm}(\rho, \theta) = V_{n,-m}(\rho, \theta)$.

The radial polynomial $R_{nm}(\rho)$ is real and of degree $n \geq 0$, with no power of ρ less than $|m|$.

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s (n-s)!}{s!(\frac{n+|m|}{2} - s)!(\frac{n-|m|}{2} - s)!} \rho^{n-2s} \tag{2.8}$$

The Zernike moment of order n with repetition m corresponding to the projection of an image function $f(x,y)$ (in the unit circle) is given by:

$$A_{nm} = \frac{n+1}{\pi} \int\int_{x^2+y^2 \le 1} f(x,y) V^*_{nm}(x,y)\, dx\, dy \; . \tag{2.9}$$

Note that A_{nm} is complex and $A^*_{nm} = A_{n,-m}$. In the case of a discrete image $f[x,y]$ the moments can be approximated as

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f[x,y] V^*_{nm}(x,y) \; , \; x^2 + y^2 \le 1 \; . \tag{2.10}$$

The orthogonality relation for V_{nm} permits reconstruction of an image from its Zernike moments.

$$\int\int_{x^2+y^2 \le 1} V_{nm}(x,y) V^*_{pq}(x,y)\, dx\, dy = \frac{\pi}{n+1} \delta_{np} \delta_{mq} \tag{2.11}$$

so that

$$\hat{f}(x,y) = \sum_{n=0}^{\infty} \sum_m A_{nm} V_{nm}(x,y), x^2 + y^2 \le 1 \tag{2.12}$$

The magnitude of Zernike moments are rotationally invariant, i.e. corresponding Zernike coefficients of two image patches that differ only by a rotation have the same magnitude, and their phase difference is related to the angle of rotation. For two images that differ by a rotation ϕ

$$g(r,\theta) = f(r, \theta + \phi) \; , \tag{2.13}$$

their Zernike moments are related by

$$A_{nm}(g) = A_{nm}(f) e^{jm\phi} \; . \tag{2.14}$$

Note that the recovery of the rotation angle using moments with $|m| \neq 1$ is non-trivial [59] because any rotation α, ($g(r,\theta) = f(r,\theta + \phi)$) of the form

$$\alpha = \phi + \frac{2\pi k}{m}, k = 0, ..., m-1; m > 0 \tag{2.15}$$

will produce the phase difference $m\phi$ between $A_{nm}(f)$ and $A_{nm}(g)$.

In a related context, Badra *et al.* [2] calculate Zernike moments for a disk around matching points and determines rotation and scaling factors that relate the images directly from the relationship between Zernike moments. The scaling relationship between moments is only approximate and is shown to hold for the images they consider. Translation is dealt with by using phase correlation once images are corrected for rotation and scaling.

2.5.2 Similarity measure

A vector of moments can be used directly as the descriptor for an image feature. Similarity between features can then be expressed as a distance between vectors. The problem with this approach is that the distances between vectors of moments do not necessarily have an obvious meaning. Using training data it is possible to derive a distance metric [106] [7] but this requires relearning the metric if the training set no longer represents the imagery. Instead, we determine that the cross-correlation between image patches can be expressed conveniently by weighted Zernike moments and form a feature descriptor from appropriately weighted moments.

We express the cross correlation between image patches f and g in terms of their moments

$$\begin{aligned} S(f,g) &= \int\int_{x^2+y^2\leq 1} f(x,y)g(x,y)\,dx\,dy && (2.16)\\ &\approx \int\int_{x^2+y^2\leq 1} \sum_n\sum_m A_{nm}(f)V_{nm}(x,y)\sum_p\sum_q A_{pq}(g)V_{pq}(x,y)\,dx\,dy && (2.17)\\ &= \sum_n\sum_m\sum_p\sum_q A_{nm}(f)A_{pq}(g)\int\int_{x^2+y^2\leq 1} V_{nm}(x,y)V_{pq}(x,y)\,dx\,dy && (2.18)\\ &= \sum_n\sum_m A_{nm}(f)A^*_{nm}(g)\frac{\pi}{n+1} && (2.19)\end{aligned}$$

where * denotes the complex conjugate.

This result suggests that we construct a vector of descriptors from all Zernike moments up to order n and repetition m by concatenating the coefficients $\sqrt{\frac{\pi}{n+1}}A_{nm}$ for all considered n and m into a vector s. We can then define the similarity score $d_{f,g}$ (based on Zernike moments of up to order n and repetition m) for the preliminary match as

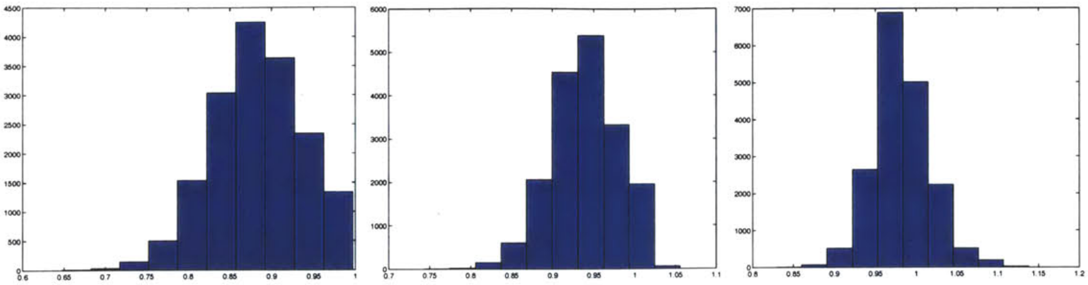


Figure 2-8: Distribution of self-similarity scores for multiple image patches. The similarity measure based on weighted Zernike moments only approximates the cross-correlation score. From left to right: n=12, n=16, n=20. The self-similarity score should be unity. With more terms the distribution is tighter and closer to unity.

$$d_{f,g} = \mathbf{s}(f)^\top \cdot \mathbf{s}(g)^* = \sum_{nm} \sqrt{\frac{\pi}{n+1}} A_{nm}(f) \cdot \sqrt{\frac{\pi}{n+1}} A^*_{nm}(g) \tag{2.20}$$

To obtain the exact correlation score requires evaluating an infinite sum. In practice only a few coefficients suffice to approximate image patches reasonably well. Figures 2-10 and 2-12 show that the reconstruction quality improves as the order (n) is increased from 8 to 24. The quality of the reconstruction depends on the number of terms used and the frequency content of the image patch. For smoothly varying patches fewer coefficients are sufficient for a close approximation. To determine the number of coefficients required we conducted a simple test based on the self-similarity of the descriptors for multiple (over 18000) patches from typical imagery. Since such a measure should approximate the autocorrelation we expect the values to be close to unity. Figure 2-8 shows the distributions of self-similarity scores for $n = 12, 16, 20$.

In addition, to test the performance of the descriptors for other values of correlation score we generated a synthetic sequence of image patches where each image is a small random perturbation of the previous one. This yields patches that are highly correlated with nearby patches in the sequence but uncorrelated with those that are distant. The true correlation score between patches is shown in plot (a) of Figure 2-9. The rest are the similarity scores based on the descriptor vectors. The same information is summarized as curves of similarity score versus true correlation for different order of descriptors in Figure 2-10. A sample patch, its polar representation and the polar reconstructions for different orders of coefficients are shown in Figures 2-11 and 2-12. The frequency content of the synthetic patches was adjusted so that the autocorrelation scores approximated those observed in typical underwater imagery. Overall, we chose to use all moments up to order $n = 16$ as a compromise between quality of approximation and compactness.

A 51-pixel diameter patch requires multiplying 2041 ($\pi D^2/4$) pixel values in the disc to calculate the correlation directly while the similarity measure that approximates the cross correlation requires multiplying 153 ($n \leq 16$ and all valid repetitions m) weighted moments.

2.6 Performance evaluation

To evaluate the performance of our method, the affine invariant region extraction and moment-based descriptor was compared to a fixed-window correlation-based match on a sequence of underwater imagery. We conducted our test for a diverse range of baseline magnitudes by matching each of 67 images to the next six images in a test sequence (for a total of 351 two view matches). The details of the robust two view matching technique we used are described in the next chapter. We used it here as a means to compare similarity-based measures over many correspondences by determining which proposed matches are consistent with the epipolar geometry.

Navigation sensors provide an image-independent estimate of baseline magnitude $|\mathbf{t}|$ and altitude z, which allows us to formulate a normalized baseline magnitude $|\mathbf{t}|/z$. This is the relevant quantity for induced parallax and allows us to plot the number of correct matches against a growing baseline (Figure 2-13). In addition, for pairs that could be matched reliably and for which the camera pose could be calculated accurately, the change in viewing angle to a feature can be calculated from the camera poses and from the rays in correspondence (Figure 2-14).

The fixed-window feature method failed to match 122 of the 351 pairs, typically for large baselines. This can be seen in Figure 2-13 for normalized baseline magnitudes above 0.45. The affine-invariant regions failed on only 44 pairs, with the degradation in matching performance for increasing baseline far more gradual (2-14) for the shape-adaptive regions. This can also be seen in the 2D histograms of the ratio of inliers to proposed matches as a

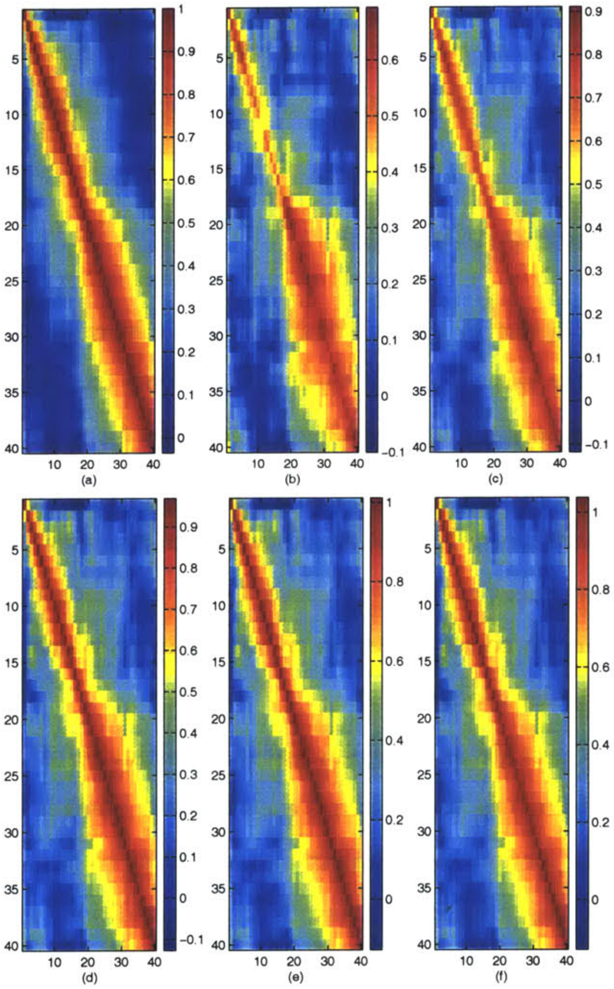


Figure 2-9: Similarity measure between 40 synthetic image patches. (a) Actual correlation score. Similarity scores for n=8 (b), n=12 (c), n=16 (d), n=20 (e) and n=24 (f). The approximation improves as more terms are added, in particular for high correlations.

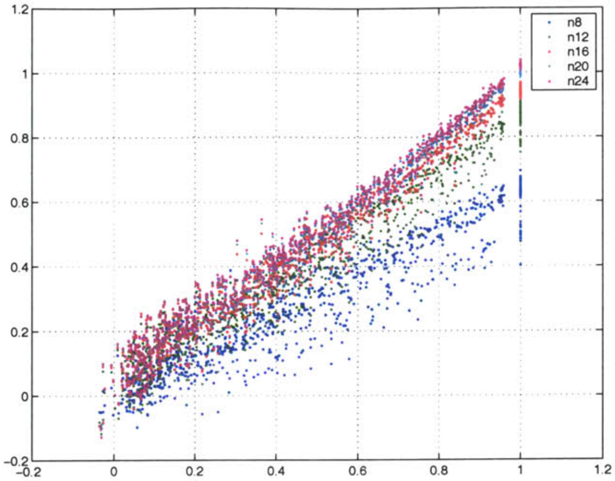


Figure 2-10: Similarity score vs. actual correlation score for varying number of coefficients. The approximation improves as more terms are added, in particular for high correlations.

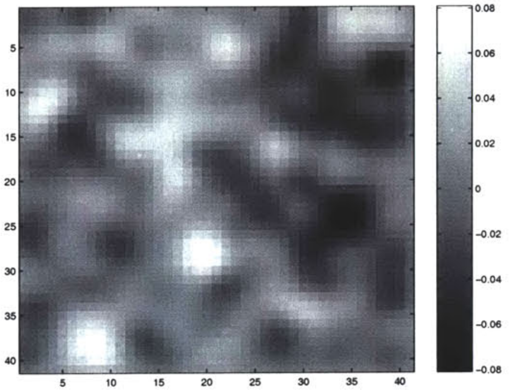


Figure 2-11: Sample image patch for which correlation and similarity scores are calculated.

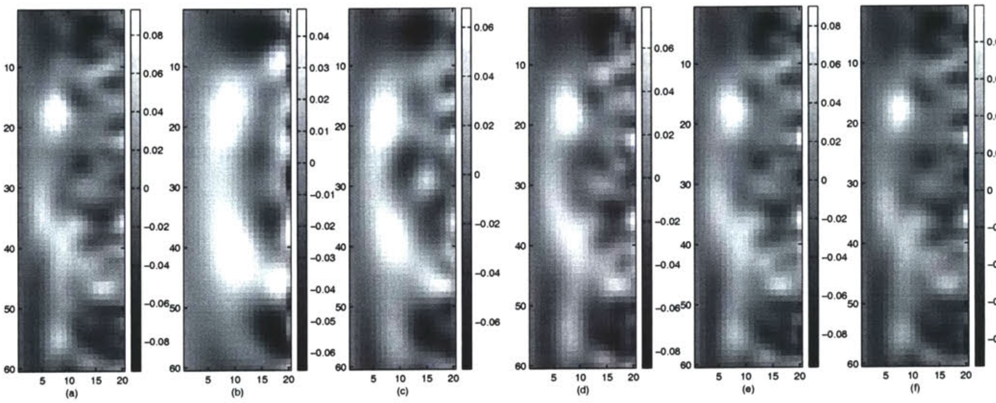


Figure 2-12: Polar representation of the patch in figure 2-11. (a) Exact polar representation. Reconstructions for (b) n=8, (c) n=12 , (d) n=16, (e) n=20 and (f) n=24.

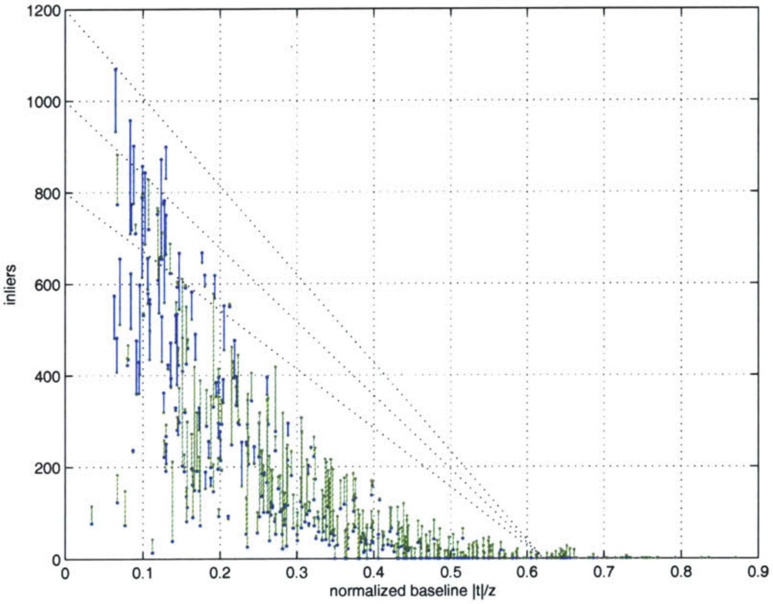


Figure 2-13: Inliers for a fixed window (blue) and shape-adapting window (green) versus normalized baseline magnitude. The vertical lines connecting the corresponding two view match under fixed window and shape-adapting window are colored according to which region provides more inliers. The affine-invariant region outperforms the fixed window as the baseline increases. As the baseline increases there is less overlap and the total number of inliers should decrease linearly even for perfect matching (under assumptions of pure translation and uniformly distributed features). The dotted lines show the expected trend for an image with 1200, 1000 and 800 features assuming a field of view of 34.5° which is the FOV of the SeaBED camera in the direction of motion. There are some inliers beyond the point where overlap would be possible. This is probably due to heading changes and also to the normalized baseline being only an approximation (based on navigation sensors, and ignoring any relief on the ocean floor).

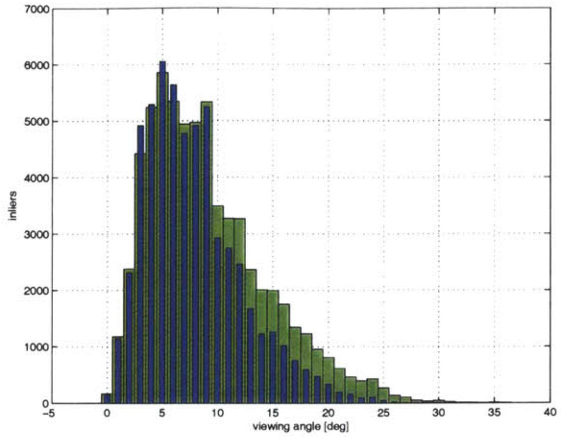


Figure 2-14: For the matches classified as inliers it is possible to calculate the viewing angle change between cameras viewing the feature. For all matches, across all pairs in the trial, we show the number of inliers as a function of viewing angle. For narrow-baseline conditions (angles of 10° or less) both regions behave similarly. For larger viewing angles the affine invariant region (green) outperforms the fixed window method (blue).

function of baseline magnitude in Figure 2-15.

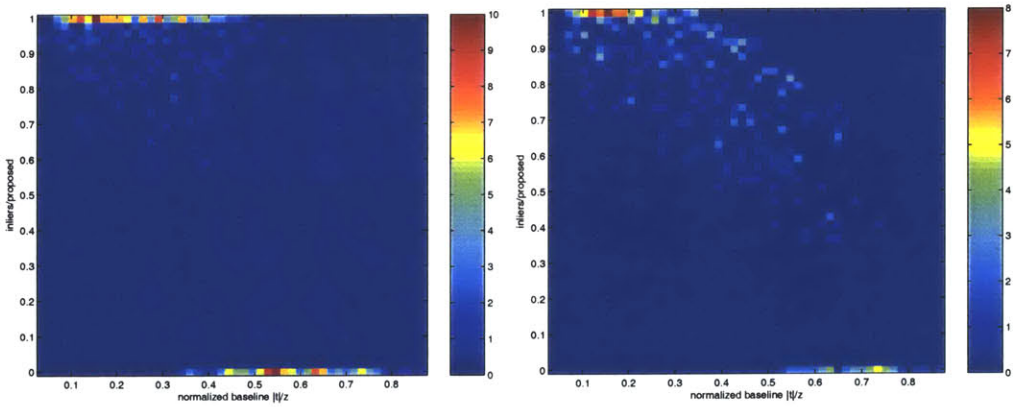


Figure 2-15: (Left) The distribution of the ratio of inliers to proposed matches against baseline magnitude for the 351 test pairs under fixed-window matching. For narrow baseline most proposals are inliers but for large enough baseline this abruptly changes to a low ratio. (Right) For the affine-invariant region, the degradation is gradual and inliers are detected for wider baselines.

Chapter 3

Relative Pose Estimation

3.1 Overview

Robust simultaneous estimation of the two view relation [28] lies at the core of most successful structure and motion algorithms. Recent efforts have focused mainly on the uncalibrated case with no prior knowledge of pose [94], resulting in greater applicability of these techniques as well as a greater understanding of the underlying problem. However, in practice it is often the case that robotic vehicles carry calibrated cameras as well as pose sensors [1] [17]. In this chapter we seek to exploit prior pose knowledge to simplify and improve the reliability of the components used in estimating relative pose from images.

A 'standard' feature-based framework for relative pose estimation comprises three main components: correspondence proposal, robust two-view relation estimation with outlier rejection, and final pose refinement. This chapter presents an equivalent framework for instrumented and calibrated platforms where two view matching forms the core of a structure and motion estimation algorithm.

Following [90] we use prior pose knowledge to limit the search for correspondences to regions consistent with the camera motion (and its uncertainty). The constrained search increases the reliability of our feature matching stage, which is particularly important when dealing with wide-baseline imagery where inter-image motion may be large. We also introduce a new six-point algorithm used within the context of RANSAC to robustly estimate the essential matrix [24] and a consistent set of correspondences. We take advantage of calibration to directly estimate the essential matrix rather than the fundamental matrix since the latter cannot be estimated from planar scenes [121]. Our solution to the essential matrix is simpler than the minimal five point solutions [45] [125] and, unlike the linear 6-point solution [89], it does not fail in the presence of planar scenes. In parallel, Nister developed an efficient solution to the minimal five-point case [85]. The complexity of his algorithm is similar to ours, though because it uses the minimal set of points there are ten possible consistent motions which his algorithm must then chose among.

The first part of this chapter reviews a correspondence search procedure based on prior pose knowledge in a calibrated camera system. Using a pose and calibration dependent two-view point transfer model, we carry forward the uncertainty in our pose and calibration to expand this point transfer to a region. This region is then used to restrict the interest point matching to a set of candidate correspondences.

The second part of this chapter describes an essential matrix estimation algorithm that is used to determine inliers and outliers in a RANSAC framework. By enforcing the constraints specific to an essential matrix, it is possible to solve utilizing six point correspondences. The solution uses the singular value decomposition (SVD) of a system of four equations followed by the solution of a sixth degree polynomial.

3.1.1 Assumptions

The assumptions under which we formulate our solution include

- Wide baseline imaging conditions. Image overlap from 50% to 67% with changes in viewing angle of 15° to 45°.

- A prior on relative pose must be available. We assume that the vehicle navigation system produces an estimate of the vehicle trajectory and that it is possible to extract relative pose and its uncertainty from the trajectory.

- Calibrated camera and sensor frames. The connection between pose priors and interest point locations is established through camera calibration and knowledge of sensor frames relative to the vehicle frame. This transforms pixel coordinates into euclidean rays that can be rotated and translated.

3.2 Pose restricted correspondence search

Narrow-baseline vision systems usually constrain the search for correspondences to small windows around interest points. The underlying assumption is that inter-image point motion is small and that it is sufficient to search in a new frame in a small area centered around the (transformed) position of the interest point in the previous frame. These heuristically restricted searches can fail in the wide-baseline case, where the apparent motion of interest points can be comparable to the size of the image, or where relief is significant. Wide-baseline systems based exclusively on imagery rely heavily on feature descriptors that are sufficiently discriminating to be able to propose matches even when compared to all other features in the image. Several semi-local constraints and consistency checks can be applied [118] [128] if the descriptor provides information on a local reference frame. These approaches tend to fail in scenes with repetitive structure or when the temporal sequence has low overlap.

In the case of a calibrated and instrumented imaging platform, uncertain relative pose information is available to restrict the search for correspondences along regions consistent with the pose prior. Prior pose knowledge relaxes the demands on the complexity of the feature descriptor since the descriptor is no longer required to be unique globally, rather only in a local sense.

Putative matches thus derived can be separated into inliers and outliers based on consistency with a motion model. For uncalibrated systems and a general scene, the fundamental matrix encodes the motion in a form that is convenient for robust estimation. Camera calibration constrains the fundamental matrix into the essential matrix. While more complex to estimate directly, the essential matrix offers advantages over the fundamental matrix since it can be determined even in the case of planar scenes.

3.2.1 Point Transfer Mapping

The point transfer mapping parametrized by pose and calibration parameters, and dependent on scene depth provides a physically meaningful framework to limit correspondence search. To place this in context we briefly review the approach used in [90] [21]. We assume

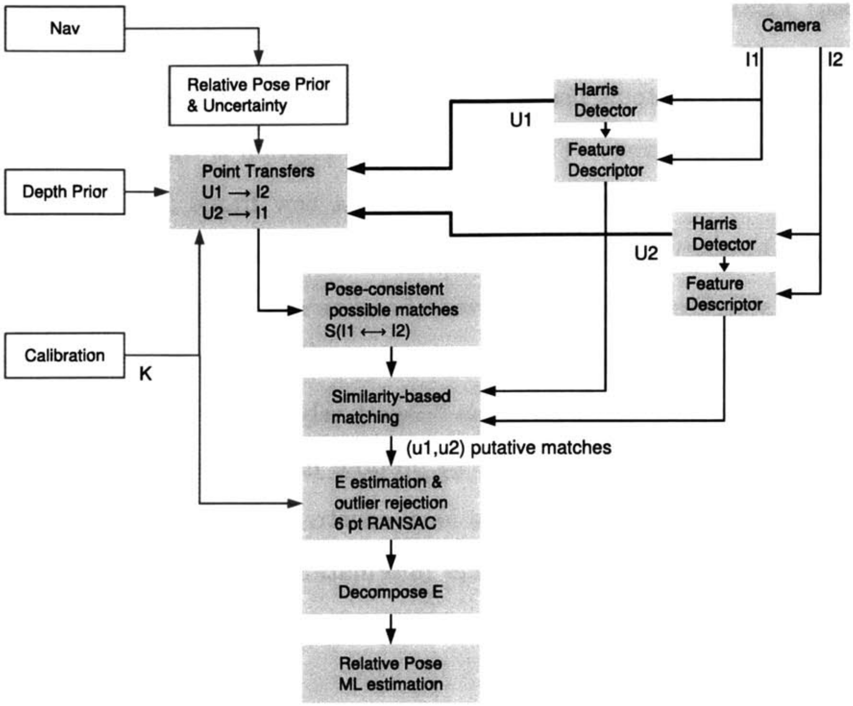

Nav
Relative Pose Prior & Uncertainty
Depth Prior
Point Transfers U1 ⟶ I2 U2 ⟶ I1
Calibration
K
Camera
I1
I2
U1
Harris Detector
Feature Descriptor
U2
Harris Detector
Feature Descriptor
Pose-consistent possible matches S(I1 ⟷ I2)
Similarity-based matching
(u1,u2) putative matches
E estimation & outlier rejection 6 pt RANSAC
Decompose E
Relative Pose ML estimation

Figure 3-1: Overview of our approach to relative pose estimation from instrumented and calibrated platforms. Unshaded blocks represent additional information compared to the uninstrumented/uncalibrated case. Given two images, we detect features using the Harris interest point detector. For each feature we then determine search regions in the other image from sensor based pose and depth information. Putative matches are identified based on similarity and constrained by regions. We then use RANSAC and the proposed 6-point algorithm to robustly estimate the essential matrix which is then decomposed into its proper motion parameters. The pose is then refined by minimizing the reprojection error over all matches considered inliers.

projective camera matrices $\mathrm{P} = \mathrm{K}[\mathrm{I} \mid \mathbf{0}]$ and $\mathrm{P}' = \mathrm{K}[\mathrm{R} \mid \mathbf{t}]$, where K is the matrix of intrinsic camera parameters [39],

$$\mathrm{K} = \begin{bmatrix} \alpha_x & s & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where α_x is the focal length in pixel widths, α_y is the focal length in pixel heights, (u_0, v_0) is the coordinate of the principle point in pixels, and s is the skew in pixel shape. R is the $[3 \times 3]$ orthonormal rotation matrix from camera frame implied in P to the reference frame used by P′ parameterized by XYZ convention Euler angles $\mathbf{\Theta} = [\phi, \theta, \psi]^\top$ (roll, pitch, heading) [34], and $\mathbf{t}$ is the $[3 \times 1]$ translation vector from the frame of P$'$ to P as represented in frame of P'.

Given an interest point with pixel coordinates (u, v) in image I, we define its vector representation $\mathbf{u} = [u, v]^\top$, as well as its normalized homogeneous representation $\underline{\mathbf{u}} = [\mathbf{u}^\top, 1]^\top$. Likewise we define a vector of the imaged scene point as $\mathbf{X} = [X, Y, Z]^\top$ and its normalized homogeneous representation $\underline{\mathbf{X}} = [\mathbf{X}^\top, 1]^\top$ and note that equality in expressions involving homogeneous vectors is implicitly defined up to scale.

The two view point transfer is then given by [90] [39]

$$\underline{\mathbf{u}}' = \frac{\mathrm{KRK}^{-1}\underline{\mathbf{u}} + \mathrm{K}\mathbf{t}/Z}{\mathrm{R}_3^\top \mathrm{K}^{-1}\underline{\mathbf{u}} + t_z/Z} \tag{3.1}$$

where $\mathbf{R}_3^\top$ is the third row of R and t_z is the component of $\mathbf{t}$ along the z axis.

When the depth of the scene point Z is known in the frame of camera P, then (3.1) describes the exact two-view point transfer mapping. However, if we let Z vary, (3.1) traces out the corresponding epipolar line in I'.

3.2.2 Point Transfer Mapping with Uncertainty

The point transfer mapping given in (3.1) can be viewed as a function of a 12 element measurement vector $\mathbf{\Phi}$ [21]. The measurement vector $\mathbf{\Phi}$ is composed from elements of the calibration matrix denoted in vector form as $\mathbf{k} = [\alpha_x, \alpha_y, s, u_0, v_0]^\top$; the six measured pose quantities obtained from the navigation sensors on the underwater vehicle; and the scene depth Z as measured in the frame of P.

$$\mathbf{\Phi} = [\mathbf{k}^\top, \mathbf{\Theta}^\top, \mathbf{t}^\top, Z]^\top \tag{3.2}$$

$\mathbf{\Phi}$ is uncertain and we assume that it is described by a probability distribution. For modeling purposes we represent $\mathbf{\Phi}$ by the first two moments of the distribution $\overline{\mathbf{\Phi}} = E[\mathbf{\Phi}]$ and $\Sigma_{\mathbf{\Phi}} = E[\mathbf{\Phi}\mathbf{\Phi}^\top] - E[\mathbf{\Phi}]E[\mathbf{\Phi}]^\top$. Defining $f(\mathbf{\Phi}; \mathbf{u})$ to be the non-homogeneous point transfer mapping given in (3.1), then to first order we can approximate the mean and covariance of $\mathbf{u}'$ as

$$\overline{\mathbf{u}}' \approx f(\overline{\mathbf{\Phi}}; \mathbf{u}) \tag{3.3}$$

$$\Sigma_{\mathbf{u}'} \approx \mathrm{J}\Sigma_{\mathbf{\Phi}}\mathrm{J}^{\top} \tag{3.4}$$

where J is the $[2 \times 12]$ Jacobian matrix of $f(\mathbf{\Phi}; \mathbf{u})$ with respect to $\mathbf{\Phi}$ evaluated at $\overline{\mathbf{\Phi}}$.

If $\mathbf{\Phi}$ is Gaussian, then to first order the distribution on $\mathbf{u}'$ will also be Gaussian with the given statistics in (3.3) and (3.4). The Gaussian model allows us to generate a bounded search region in I' when trying to constrain possible correspondences for $\mathbf{u}$.

$$(\mathbf{u}' - \overline{\mathbf{u}}')^{\top}\Sigma_{\mathbf{u}'}^{-1}(\mathbf{u}' - \overline{\mathbf{u}}') = k^2 \tag{3.5}$$

defines an ellipse in (u', v') space and k^2 follows a χ_2^2 distribution.

In the case where no knowledge of Z is available, by picking any finite value for Z and letting σ_Z be very large, we recover a search band around the prior pose measured epipolar line in I' whose width corresponds to the uncertainty in the other elements of $\mathbf{\Phi}$ (Figure 3-2). In the case where knowledge of average scene depth exists, such as from an altimeter on an underwater vehicle, and constraints on the minimum and maximum distance to the scene can be imposed, then Z_{avg} and an appropriate σ_Z can be chosen to limit the search to a segment of the epipolar line.

So far we have described how to associate a region to the uncertain transfer of a point from I to I'. We can perform the transfer in the opposite direction by swapping $\underline{\mathbf{u}}$ with $\underline{\mathbf{u}}'$ and replacing R with $\mathrm{R}^{\top}$, $\mathbf{t}$ with $-\mathrm{R}^{\top}\mathbf{t}$ and Z with Z' in (3.1). By intersecting the possible correspondences from I to I' and from I' to I we form a set of possible *bidirectional* correspondences consistent with the relative pose and its uncertainty. We can express this as

$$S_{I \leftrightarrow I'} = S_{I \rightarrow I'} \cap S_{I \leftarrow I'} \tag{3.6}$$

Where S is the set of possible correspondences and the subindex shows in which direction the pose constraint is used. In practice, we choose I as the image with fewer interest points

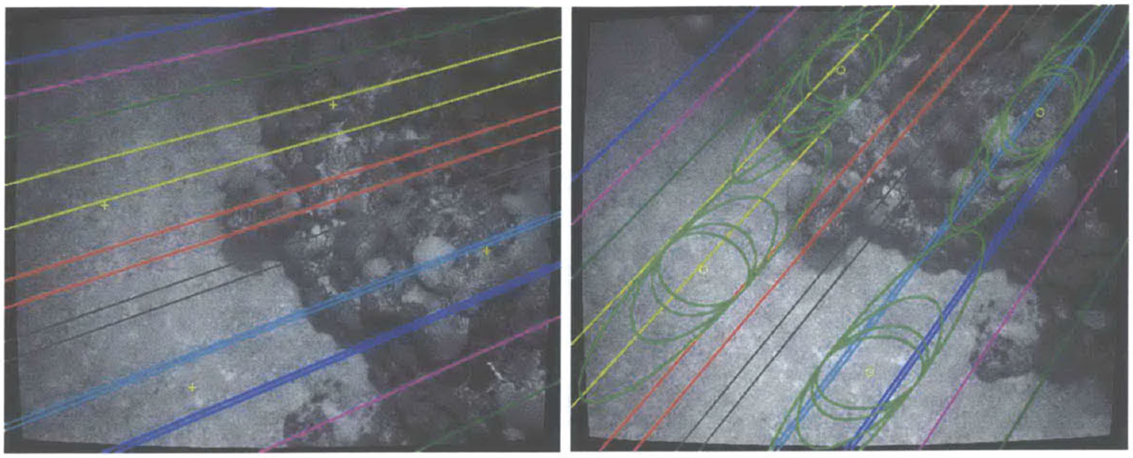


Figure 3-2: Transfer of four feature points from the left image to the right image based on a pose prior and depth information. A sampling of epipolar lines is included as a visual reference. These epipolar lines are based on the pose prior and only approximate the true epipolar lines. The 99% confidence ellipses show that increasing the scene depth uncertainty (by doubling the standard deviation of the scene depth) grows the possible correspondences along the epipolar lines.

and then determine possible correspondences in I' according to Φ and Σ_Φ . Fewer transfers are performed overall if only these possible interest points in I' are transferred back to I. This procedure can be represented as

$$S_{I \leftrightarrow I'} = S_{(I \rightarrow I') \rightarrow I} \tag{3.7}$$

Figure 3-3 illustrates the 99% confidence level pose restricted correspondence search regions for a pair of underwater images. A sampling of interest points and sensor instantiated epipolar lines are shown in the top image; their associated candidate correspondence search regions are shown in the bottom image. The search regions are determined using an altimeter measurement of the average scene depth and setting σ_Z to 0.75 meters. Figure 3-4 shows the pose restricted correspondence matrix for the image pair. Note that the set of possible correspondences has been reduced from a full matrix, to a sparse matrix. The resulting candidate set is 50 times smaller than the set of all possible matches.

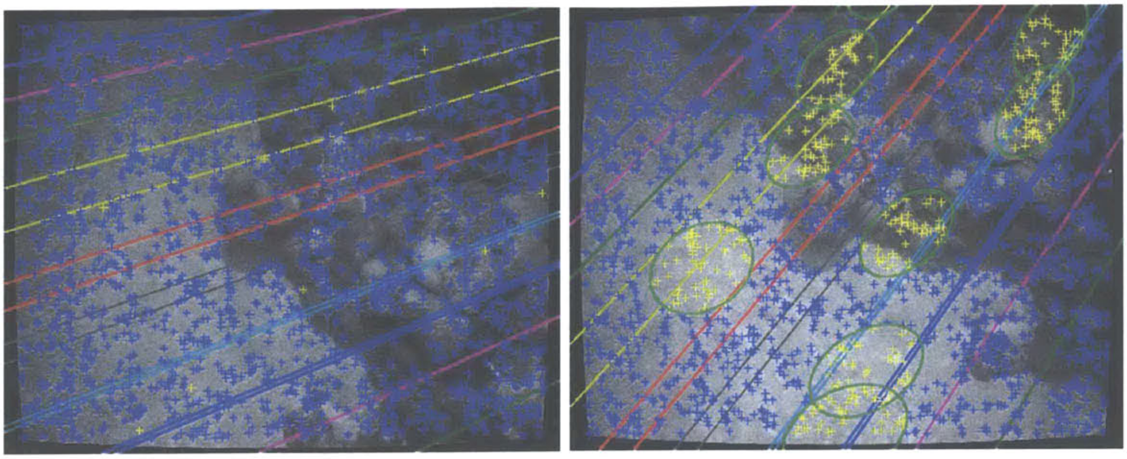


Figure 3-3: Prior pose restricted correspondence search on a pair of underwater coral reef images. (top) Interest points are shown in blue. A sampling of interest points (yellow) are transferred to the right image. (bottom) The 99% confidence regions for the transferred points based on the pose prior and depth standard deviation of 0.75m. The candidate interest points that fall within these regions are highlighted in yellow.

3.3 Essential Matrix estimation

Relative pose from calibrated cameras is a 5 DOF problem (3 DOF for rotation and 2 DOF for direction of motion between cameras), because of loss of scale. Minimal five-point algorithms [45][125][26] tend to be ill-posed, have complex implementations, and can present up to 20 solutions that then have to be tested. In this section we present a method that uses six point matches to determine relative motion using the essential matrix.

The $[3 \times 3]$ essential matrix E encodes the relative motion between two cameras [39]. In terms of the motion parameters it has the following form

$$\mathrm{E} = [\mathbf{t}]_\times \mathrm{R}. \tag{3.8}$$

where $[\mathbf{t}]_\times$ is the skew symmetric matrix based on $\mathbf{t}$ such that $[\mathbf{t}]_\times \mathbf{a} = \mathbf{t} \times \mathbf{a}$. The essential matrix E can be considered as a special case of the fundamental matrix, satisfying the following relationship:

$$\underline{\mathbf{x}}'^\top \mathrm{E} \underline{\mathbf{x}} = 0 \tag{3.9}$$

where $\underline{\mathbf{x}} = [x, y, 1]^\top$ and $\underline{\mathbf{x}}' = [x', y', 1]^\top$ are normalized image point correspondences (i.e.

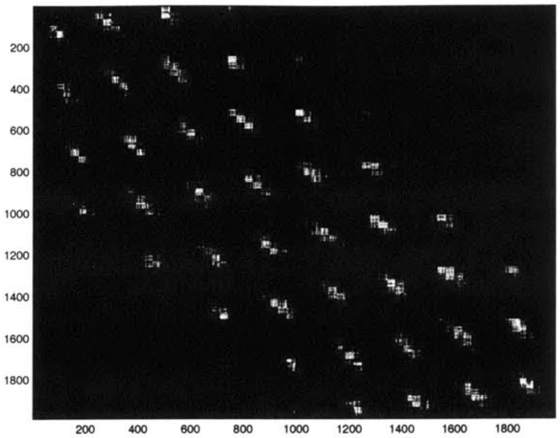


Figure 3-4: Candidate correspondence matrix for the image pair shown in Figure 3-3. Nearly 2000 interest points were selected in each image. Without any prior pose knowledge this matrix would be full (almost 4 million elements). Pose restricted search regions reduce the correspondence matrix to this sparse form (91,016 elements).

$\underline{\mathbf{x}} = \mathrm{K}^{-1}\underline{\mathbf{u}}$ and $\underline{\mathbf{x}}' = \mathrm{K}^{-1}\underline{\mathbf{u}}'$). As a fundamental matrix, E has a null determinant and because of calibration it has two equal singular values. Consider

$$\mathrm{E} = \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \tag{3.10}$$

and define $\mathbf{e} = [e_{11}, e_{12}, e_{13}, e_{21}, e_{22}, e_{23}, e_{31}, e_{32}, e_{33}]^{\top}$ as its vector representation. Then (3.9) can be written as

$$\begin{bmatrix} x'x & x'y & x' & y'x & y'y & y' & x & y & 1 \end{bmatrix} \mathbf{e} = 0 \tag{3.11}$$

With a set of n point matches $(x_i, y_i) \leftrightarrow (x'_i, y'_i)$ we can form a linear system of the form

$$\mathbf{Ae} = \mathbf{0}_{n\times 1} \tag{3.12}$$

where A =

$$\begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \end{bmatrix} \tag{3.13}$$

For $n = 8$ and non-critical motion and point configurations, we have the classic 8-point algorithm [64] which solves for the $\mathbf{e}$ that satisfies (3.12). With $n = 7$ we can find $\mathbf{e}$ as the linear combination of the two generators of the right null space of A and impose the $det(\mathrm{E}) = 0$ constraint. However when the (x_i, y_i) are coplanar the rank of A drops to 6, and the 8 and 7-point algorithms can no longer be used [89]. With $n = 6$ point matches, A will have rank 6 and $\mathbf{e}$ will be a linear combination of the generators of the right null space of A determined by SVD, i.e. $\mathbf{e} = a\mathbf{e}_1 + b\mathbf{e}_2 + c\mathbf{e}_3$ or in matrix form

$$\mathrm{E} = a\mathrm{E}_1 + b\mathrm{E}_2 + c\mathrm{E}_3 \tag{3.14}$$

Homogeneity of the equations implies that E is determined only up to scale, and therefore can be expressed in terms of two parameters

$$\mathrm{E} = \alpha\mathrm{E}_1 + \beta\mathrm{E}_2 + \mathrm{E}_3\,. \tag{3.15}$$

The values of α and β must be determined such that E is an essential matrix. A $[3 \times 3]$ matrix is an essential matrix (one null and two equal singular values) if and only if it satisfies the Demazure constraint [24]

$$\mathrm{E}\mathrm{E}^{\mathsf{T}}\mathrm{E} - \frac{1}{2}trace(\mathrm{E}\mathrm{E}^{\mathsf{T}})\mathrm{E} = 0 \tag{3.16}$$

By replacing E in (3.16) with (3.15) we generate a system of 9 homogeneous polynomial equations of degree 3 in α and β. This can be considered a homogeneous linear system in the terms $\alpha^3, \alpha^2\beta, \alpha^2, \alpha\beta^2, \alpha\beta, \alpha, \beta^3, \beta^2, \beta, 1$. With 9 linear equations and 9 unknowns it is possible to solve uniquely for the vector of unknowns and therefore obtain E. This technique is known as the 6-point linear algorithm [88]. Notice that this approach will satisfy the Demazure constraint only approximately since there is no guarantee that the relationship between different powers of α and β will be preserved.

3.3.1 Planar Scenes and Failure of the Linear 6-Point Algorithm

The linear six-point algorithm fails in cases where all of the 3D points lie in a plane. In such a case system A will still have rank 6, but the system defined by (3.16) will now drop to rank 4 rather than 9, because the linear dependence between 3D points introduces additional relations [89]. This result significantly reduces the applicability of the linear 6-point algorithm in practical situations. It could be used in a model selection framework [121][96], but we seek a simpler approach.

3.3.2 A Six-Point Algorithm Robust to Planar Scenes

The Hofmann-Wellenhof method for six points [88] uses the constraints

$$\mathrm{E}\mathrm{E}^\top\mathrm{E}\mathrm{E}^\top - \frac{1}{2}trace(\mathrm{E}\mathrm{E}^\top)\mathrm{E}\mathrm{E}^\top = 0 \tag{3.17}$$

and $det(\mathrm{E}) = 0$ on the six homogeneous linear equations represented in A. It results in a system of 7 polynomial equations of degree 4 in α and β. Manipulating these equations produces a system of two polynomial equations of degrees 8 and 9 in β. The common solutions of these two polynomials allows one to solve for α and then subsequently E.

The basic idea behind our method is to use only 4 equations from the Demazure constraint and solve the resulting system of polynomials of degree 3. By always using a system of rank 4 we will find a solution even in the presence of planar scenes. We show that by manipulating this system we can generate a polynomial of degree 6 in β, and then solve for α.

Consider four equations from the Demazure constraint in terms of α^3, $\alpha^2\beta$, α^2, $\alpha\beta^2$, $\alpha\beta$, α, β^3, β^2, β, 1. To pick four equations, we perform SVD on the 9 equations of the Demazure constraint and then select the four right singular vectors associated with the largest singular values. Performing Gauss-Seidel elimination on the four equations we have

$$\begin{array}{cccccccccc}
\alpha^3 & \alpha^2\beta & \alpha^2 & \alpha\beta^2 & \alpha\beta & \alpha & \beta^3 & \beta^2 & \beta & 1 \\
\hline
1 & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & 1 & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & 1 & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array} \tag{3.18}$$

Here a blank represents zero and '$\cdot$' represents some value not eliminated by Gauss-Seidel. This system can be represented as

$$\alpha^3\beta_a^0 + \alpha\beta_b^1 + \beta_c^3 = 0 \tag{3.19}$$

$$\alpha^2\beta_d^1 + \alpha\beta_e^1 + \beta_f^3 = 0 \tag{3.20}$$

$$\alpha^2\beta_g^0 + \alpha\beta_h^1 + \beta_i^3 = 0 \tag{3.21}$$

$$\alpha\beta_j^2 + \beta_k^3 = 0 \tag{3.22}$$

where the β_p^n represents an n^{th} degree polynomial in β and the subscript p is an identifier for one of 11 distinct polynomials. By multiplying the second equation (3.20) by β_g^0 and the third equation (3.21) by β_d^1, then subtracting we obtain

$$\alpha(\beta_e^1\beta_g^0 - \beta_h^1\beta_d^1) + \beta_f^3\beta_g^0 - \beta_i^3\beta_d^1 = 0\,. \tag{3.23}$$

Notice that the above expression no longer depends on α^2. Defining $\beta_s^2 = \beta_e^1\beta_g^0 - \beta_h^1\beta_d^1$ and $\beta_t^4 = \beta_f^3\beta_g^0 - \beta_i^3\beta_d^1$, together with the fourth equation (3.22), we have

$$\alpha\beta_s^2 + \beta_t^4 = 0 \tag{3.24}$$

$$\alpha\beta_j^2 + \beta_k^3 = 0 \tag{3.25}$$

Cross multiplying by the polynomials of degree two β_j^2 and β_s^2, and subtracting we obtain a single polynomial equation of degree six

$$\beta_t^4\beta_j^2 - \beta_k^3\beta_s^2 = 0 \tag{3.26}$$

We solve this polynomial and use the real roots to solve for α using (3.22). For each real pair (α, β) we calculate the corresponding essential matrix according to (3.15). The proposed six-point algorithm will produce up to six possibly valid essential matrices.

Using synthetic data sets (generated for both planar and non-planar scenes) and random relative motion, we have determined that one of the essential matrices produced by this six-point algorithm always corresponds to the true camera motion for perfect (noise free) measurements. We have also observed that for perfect measurements of points in a planar configuration, the proposed six-point algorithm always produces two essential matrices, one which corresponds to the true essential matrix, and one which corresponds to the (incorrect) output of the linear six-point algorithm.

3.4 Robust Essential Matrix Estimation

The following two statements must hold for the proposed six-point algorithm to be useful in the context of estimating the essential matrix from a large set of putative correspondences. First, we must be able to select the correct solution from up to six essential matrices. Second, the quality of the solution must degrade gracefully in the presence of measurement noise.

we select a solution with a RANSAC approach, testing the solutions against the entire correspondence set and selecting the one with the most inliers. We determine inliers based on the reprojection error using implicit triangulation [57] which is more efficient than the solution based on a sixth-degree polynomial [39].

To test how the performance of this algorithm degrades in the presence of noise, we performed 1000 trials with randomly generated scenes and motions. For each trial the essential matrices computed by the six-point algorithm were decomposed into their respective rotation and translation representation. The essential matrix with rotation and translation that was closest (minimum error) to the true motion was selected. In order to summarize results in one quantity, we define a pose error measure as the sum of (1) the angle of rotation between the true rotation matrix R and the estimated $\hat{\mathrm{R}}$ using the axis-angle representation [87], and (2) the angle between the translation direction vectors. These trials were then repeated with different levels of noise added to the pixel coordinates.

Figure 3-5 shows the minimum, median, and maximum error for increasing pixel noise variance for a scene with points in a general configuration. The top figure shows results of the linear 6-point algorithm while the bottom figure shows results from the proposed 6-point algorithm. Notice that for perfect measurements (i.e. zero noise) both algorithms produce the correct essential matrix. Figure 3-6 plots the same curves for a test where for each trial the 3D points were in a planar configuration of random orientation. Notice that the linear 6-point algorithm fails even for perfect measurements.

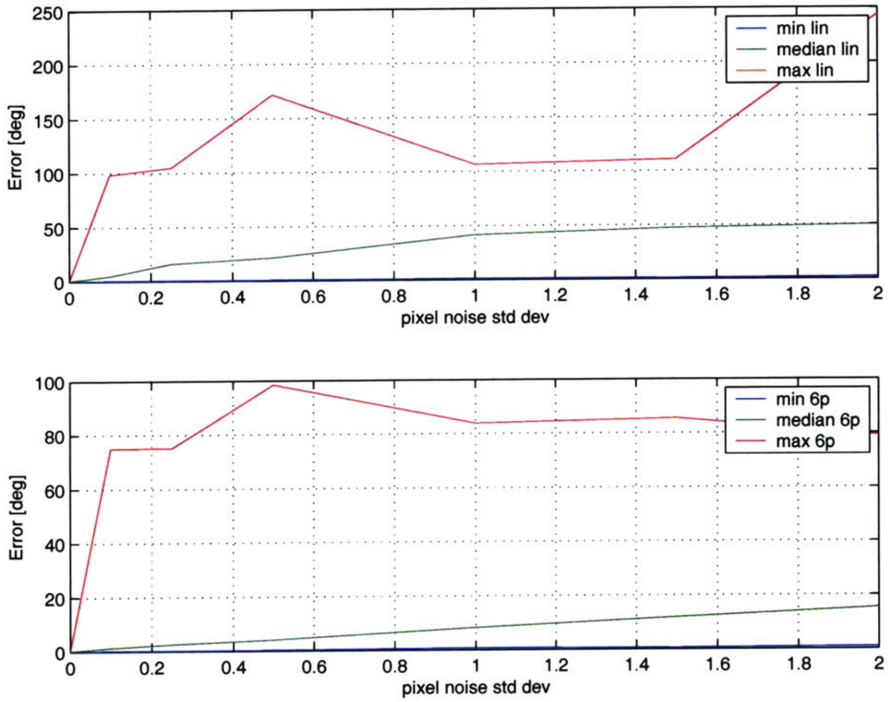


Figure 3-5: Noise test with general scenes. The minimum, median, and maximum pose error (defined as the sum of the rotation error and the angular error of the baseline direction vector) over 1000 trials, plotted against noise variance. (top) Linear 6-point algorithm, (bottom) proposed 6-point algorithm.

Even though the proposed 6-point algorithm degrades in the presence of noise, Figures 3-5 and 3-6 show that a large number of estimates will be close to the true motion. This suggests that the algorithm can be used effectively in a RANSAC context where it is reasonable to expect that there will be point combinations yielding an essential matrix close to the true one and will explain a large fraction of the correctly matched points.

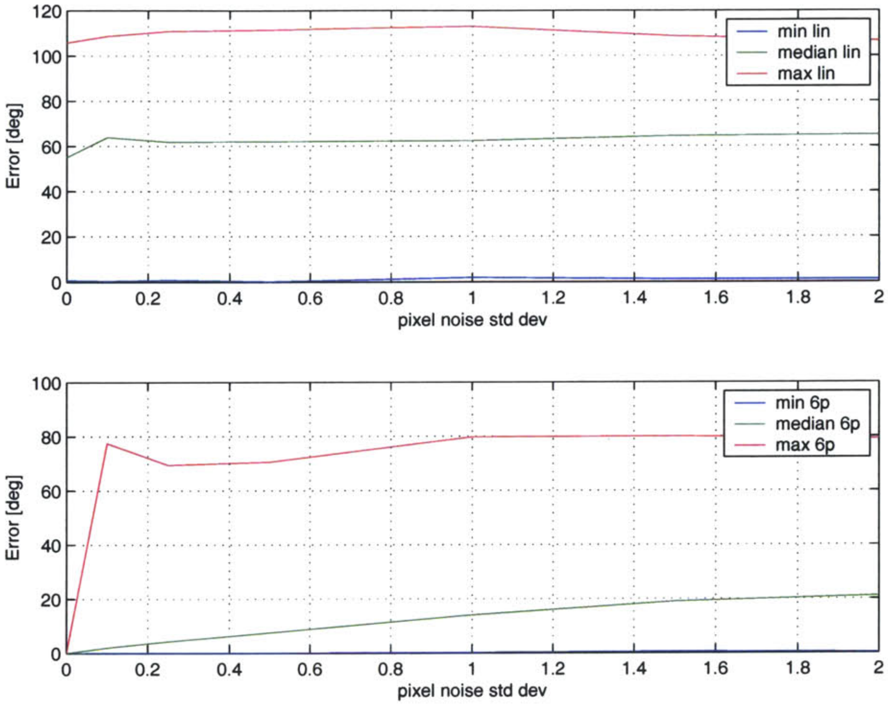


Figure 3-6: Noise test with planar scenes. The minimum, median, and maximum pose error (defined as the sum of the rotation error and the angular error of the baseline direction vector) over 1000 trials, plotted against noise variance. (top) Linear 6-point algorithm, (bottom) proposed 6-point algorithm. The failure of the linear 6-point algorithm for planar scenes can be seen in the high errors even in the absence of noise.

3.4.1 Two view critical configurations

Planar or nearly planar scenes are frequently encountered in surveys of the ocean floor. For the uncalibrated case there is a three degree of freedom ambiguity in the parametrization of the solution that generates a continuum of fundamental matrices consistent with the data. In the case of a calibrated camera, two views of an unknown plane will have at most two valid essential matrices [65]. The ambiguity can be resolved by requiring all points to be in front of both cameras except in the case where all points are closer to one camera than the other. This situation can happen when the vehicle motion has a significant component toward or away from the bottom.

Planar scenes are a particular case where scene points and the camera centers fall on a ruled quadric [66] [55]. In the general case of ruled quadrics there will be up to a three-fold ambiguity in motion and structure for the uncalibrated case. For the calibrated case the number of interpretations is two. Each interpretation will place the scene points and camera centers on distinct ruled quadrics. A dense set of points (hundreds) from a natural scene is unlikely to fall on a ruled quadric, but in cases of low overlap (tens of points) this could happen. In section 3.5 we use the motion prior from navigation instruments to disambiguate image-based solutions.

3.4.2 Reprojection Error

Given a set of n_{in} measured correspondences $S_{in} = \{\mathbf{u}_i \leftrightarrow \mathbf{u}'_i\}$, under the assumption of isotropic Gaussian noise corrupting the interest point locations, it can be shown [39] that the MLE for the fundamental matrix $\mathrm{F} = \mathrm{K}^{-\top}\mathrm{E}\mathrm{K}$ minimizes the sum of squared reprojection errors:

$$D(F, \hat{\mathbf{u}}_i, \hat{\mathbf{u}}'_i) = \sum_i d(\mathbf{u}_i, \hat{\mathbf{u}}'_i)^2 + d(\mathbf{u}'_i, \hat{\mathbf{u}}_i)^2 \tag{3.27}$$

where $d(\cdot, \cdot)$ represents the Euclidean distance and $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{u}}'_i$ are the estimated ideal correspondences (*i.e.*, before corruption with Gaussian noise) that exactly satisfy $\underline{\hat{\mathbf{u}}}'_i \mathrm{F} \underline{\hat{\mathbf{u}}}_i$.

The reprojection errors are used to rank the quality of the essential matrices proposed in the RANSAC loop. The number of inliers for a proposed essential matrix is determined by the number of correspondences with reprojection errors below a threshold t. This threshold is set based on the expected noise in feature locations and with some testing on actual images. Calculating the reprojection error requires triangulating the ideal feature points with algorithms such as Hartley and Sturm's optimal triangulation method [40] which requires solving sixth degree polynomials. Torr and Zisserman show [123] that the optimally corrected correspondences proposed by Kanatani [56] are equivalent to iterating the Sampson approximation [39] and yield a close approximation to the MLE estimate obtained by the optimal triangulation method. The ideal correspondences are calculated as [56]

$$\underline{\hat{\mathbf{u}}}_i = \underline{\mathbf{u}} - \frac{\underline{\mathbf{u}}_i'^\top \mathrm{F} \underline{\mathbf{u}}_i}{\underline{\mathbf{u}}_i'^\top \mathrm{F} \Sigma_0 \mathrm{F}^\top \underline{\mathbf{u}}_i' + \underline{\mathbf{u}}_i^\top \mathrm{F}^\top \Sigma_0 \mathrm{F} \underline{\mathbf{u}}_i} \Sigma_0 \mathrm{F}^\top \underline{\mathbf{u}}_i' \tag{3.28}$$

$$\underline{\hat{\mathbf{u}}}_i' = \underline{\mathbf{u}}_i' - \frac{\underline{\mathbf{u}}_i'^\top \mathrm{F} \underline{\mathbf{u}}_i}{\underline{\mathbf{u}}_i'^\top \mathrm{F} \Sigma_0 \mathrm{F}^\top \underline{\mathbf{u}}_i' + \underline{\mathbf{u}}_i^\top \mathrm{F}^\top \Sigma_0 \mathrm{F} \underline{\mathbf{u}}_i} \Sigma_0 \mathrm{F} \underline{\mathbf{u}}_i \tag{3.29}$$

$$\tag{3.30}$$

where $\Sigma_0 = diag(1, 1, 0)$ is the assumed covariance for homogeneous coordinates. We approximate the reprojections from triangulation by letting $\underline{\mathbf{u}} \leftarrow \underline{\hat{\mathbf{u}}}$ and $\underline{\mathbf{u}}' \leftarrow \underline{\hat{\mathbf{u}}}'$ and iterating until $\underline{\hat{\mathbf{u}}}' \mathrm{F} \underline{\hat{\mathbf{u}}} = 0$ is sufficiently satisfied. Usually one or two iterations are enough.

3.4.3 From the essential matrix to motion estimates

The essential matrix that best explains the data according to RANSAC is decomposed into a rotation and translation according to the following result from [39].

Assume that the first camera $\mathrm{P} = [\mathrm{I}|\mathbf{0}]$ and the second camera is $\mathrm{P}' = [\mathrm{R}|\mathbf{t}]$. We seek to determine P' or equivalently R and $\mathbf{t}$ given $\mathrm{E} = [\mathbf{t}]_\times \mathrm{R}$. Let

$$\mathrm{W} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.31}$$

and assume that the SVD decomposition of E is, up to scale, $\mathrm{E} \sim \mathrm{U} diag(1, 1, 0) \mathrm{V}^\top$, then the translation is given, up to scale, by $\mathbf{t} \sim \mathrm{U}[0, 0, 1]^\top = \mathbf{U}_3$ and the rotation matrix R is $\mathrm{R}_a = \mathrm{UWV}^\top$ or $\mathrm{R}_b = \mathrm{UW}^\top \mathrm{V}^\top$. Under the assumption of unit baseline magnitude, there is a four-fold ambiguity in P':

$$\mathrm{P}' = [\mathrm{R}_a|\mathbf{U}_3] \, or \, [\mathrm{R}_a| - \mathbf{U}_3] \, or \, [\mathrm{R}_b|\mathbf{U}_3] \, or \, [\mathrm{R}_b| - \mathbf{U}_3] \tag{3.32}$$

One of these choices corresponds to the true relative pose. The others correspond to a reversal in baseline direction, a rotation of 180° around the line connecting both cameras ('twisted pair'), and the twisted pair with a reversed baseline. To determine which is the correct solution we check that triangulated points are in front of both cameras.

3.4.4 Outlier Rejection (RANSAC)

To eliminate outliers (correspondences inconsistent with the motion model) an essential matrix between the two images is estimated using RANdom SAmple Consensus (RANSAC) [28]. The basic steps for outlier rejection based on RANSAC are augmented to include checking for physically realizable point configurations. The added robustness comes at the expense of additional computation, though this is incurred only when a proposed essential matrix seems superior to the current 'best' estimate. To be physically realizable, a configuration of points and relative pose must:

- place all points in front of both cameras (cheirality constraint) [39],
- the scene points lie only a few meters in front of the camera. This constraint can be invoked due to the strong attenuation of light underwater. As described in §1.2.1 the attenuation lengths underwater for the visible spectrum are in the range of 5-25 m, and
- the points must not lie between both cameras (the baseline cannot go through the surface implied by the scene points). The ocean floor is a 'solid surface' and both cameras must be on the same side of it.

Enforcing these constraints resolves many cases of ambiguities but does not resolve all ambiguous pairs. It is important to bear in mind that during the RANSAC stage we are mainly interested in determining matches that are consistent with an essential matrix. If the inliers support two or three distinct interpretations this is not a problem at the RANSAC stage. It only becomes an issue when determining and refining the final motion estimate.

The steps for the robust estimation of the essential matrix are:

- Start with the set S of potential correspondences (based on similarity), the set of inliers $S_{in} = \emptyset$ and the number of inliers $n_{in} = 0$.
- Repeat for N trials :
  - Randomly select a set p of six matches from the potential correspondences.

- Calculate E(p) the essential matrix implied by the subset. There can be one to six essential matrices. For each essential matrix:
    - Calculate the reprojection error d for all putative matches through the triangulation procedure described in §3.4.2.
    - Determine $S_{in}(p)$, the set of inliers according to E(p), as the set of matches with $d < t$ pixels. The number of inliers according to E(p) is $n_{in}(p)$.
    - If $n_{in}(p) > n_{in}$ apply the cheirality, light attenuation and 'solid surface' constraints:
        - Explicitly triangulate $S_{in}(p)$.
        - Reduce $S_{in}(p)$ and $n_{in}(p)$ to the correspondences that are in front of both cameras, are only a few meters away, and that are not between cameras.
        - If $n_{in}(p) > n_{in}$ then $n_{in} \leftarrow n_{in}(p)$ and $S_{in} \leftarrow S_{in}(p)$

The RANSAC algorithm produces an E that best explains most of the data (in the sense that the reprojection errors are less than the threshold t). Under the assumption that 2D features are localized with a standard deviation of σ, the distance squared between the measured and the estimated 2D feature location follows a χ_2^2 distribution. The cumulative chi-squared distribution $F_2(t^2) = \int_0^{t^2} \chi_2^2(x)\,dx$ includes 99% of the inliers for $t^2 = 9.21\sigma^2$ or $t = 3.03\sigma$. In practice we assume $\sigma = 1$ and our results are not overly sensitive to t. The number of iterations, N, is calculated adaptively based on the current estimate of the fraction of outliers [39]. Figure 3-7 shows the resulting image-based points considered inliers by RANSAC. The epipolar geometry in the figure is a refinement by maximum a posteriori estimation from the RANSAC inliers (Section §3.5). Figure 3-8 illustrates the triangulated correspondences and the cameras in the frame of the first camera.

3.5 Final motion estimate

The previous section recognizes that the output of the RANSAC stage is a set of inliers associated with one of possibly several interpretations of motion. The six point algorithm

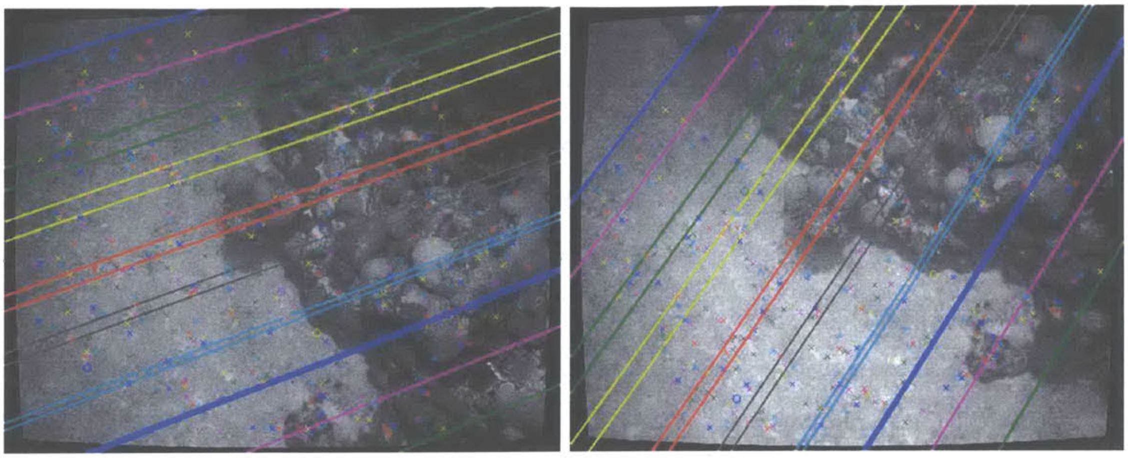


Figure 3-7: Epipolar geometry and correspondences. The given image pair illustrates the MAP refined image-based epipolar geometry. RANSAC determined 398 consistent inliers designated 'x', from the putative set of 405 matches. The rejected outliers are designated 'o'.

can be used with more than six points and in fact we use it with all inliers to generate possible essential matrices.

To disambiguate the motion we must rely on additional information. Three possible approaches are:

- Keep track of multiple hypothesis and decide on a particular motion based on consistency of motion and structure over several frames.
- Select the relative pose encoded in the essential matrix that is closest to the relative pose prior from navigation sensors.
- If there are scene points in common between three images, use resection to determine the camera pose relative to the existing structure.

We choose to use a combination of the second and third approaches since they do not delay the decision and they tend to be simpler. If there are common scene points between three or more views we use resection to generate an initial guess to relative pose. This will be discussed in more detail in the next chapter in the context of building sequential submaps. If there is not enough overlap for resection and we have multiple interpretations for the essential matrix, we choose the relative pose closest to the navigation prior. More specifically, the image-based relative pose with the smallest Mahalanobis distance $||\mathbf{p}_i - \mathbf{p}_{nav}||_{\Sigma_{nav}}$,

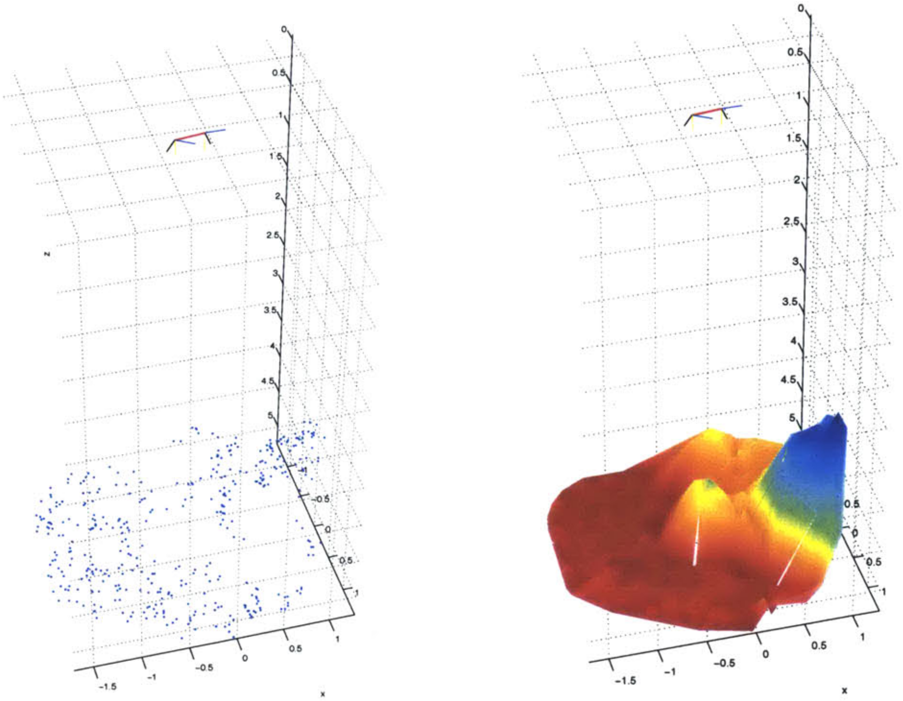


Figure 3-8: Triangulated inliers for the pair in figure 3-7. Coordinates in meters, in the reference frame of the first camera.(left) 3D feature locations. (right) Interpolated surface, colorcoded by depth from the first camera. The camera frames are shown as a blue,black and yellow frame (x,y,z) connected by the baseline (red).

based on Σ_{nav} the covariance of the prior is selected as the initial guess.

$$||\mathbf{p}_i - \mathbf{p}_{nav}||_{\Sigma_{nav}} = \sqrt{\left(\mathbf{p}_i - \mathbf{p}_{nav}\right)^\top \Sigma_{nav}^{-1} \left(\mathbf{p}_i - \mathbf{p}_{nav}\right)} \tag{3.33}$$

where $\mathbf{p}_i = [\mathbf{t}_i^\top, \mathbf{\Theta}(\mathbf{R_i})^\top]^\top$ are the translation and orientation parameters (as Euler angles) for the i^{th} Essential matrix, and $\mathbf{p}_{nav}$ is the similarly defined relative pose from the navigation sensors. Since relative pose is recovered only up to scale from images, the baseline magnitude is normalized to unit length and the covariance is constrained to be zero in the direction of motion. The baseline of the image-based solution is then scaled according to the projection of the prior baseline:

$$\mathbf{t} = \frac{\mathbf{t}_i^\top \mathbf{t}_{nav}}{||\mathbf{t}_i||} \mathbf{t}_i \tag{3.34}$$

3.5.1 Bundle Adjustment

The final relative pose estimate is based on a bundle adjustment of pose and 2D feature locations. From Bayes rule we have

$$p(\mathbf{x}|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \tag{3.35}$$

which in terms of parameter estimation can be interpreted as posterior distribution $p(\mathbf{x}|\mathbf{z})$ of a vector of parameters (associated to a model) $\mathbf{x}$ given observations $\mathbf{z}$ is proportional to the likelihood of the data given the parameters $p(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{x})$. The maximum a posteriori (MAP) estimate $\mathbf{x}^*$ maximizes the total posterior probability of the model given the observations and prior knowledge. If the prior is assumed to be uniform, then we have a maximum likelihood estimate (MLE) that selects the model for which the probability of the observed data is highest. From the point of view of estimation the distinction between MLE and MAP is nonexistent since the prior knowledge can be considered an additional observation. We choose to refer to MLE estimation when using only image-based measurements and MAP estimation when including navigation sensor measurements, though in practice the navigation information is included as additional observations.

We assume conditionally independent measurements. The MLE estimate is then

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \prod_i p(\mathbf{z}_i|\mathbf{x}) \tag{3.36}$$

For image-based measurements $\mathbf{z} = \mathbf{u}$ given the relative pose and structure $\mathbf{x} = [\mathbf{p}^\top, \mathbf{X}^\top]^\top$ the measurements can be considered to have Gaussian distributions centered around the true reprojections

$$p(\mathbf{u}|\mathbf{x}) \propto e^{[\mathbf{u}-\phi(\mathbf{x})]^\top [\mathbf{u}-\phi(\mathbf{x})]} \tag{3.37}$$

Taking the negative loglikelihood we express the MLE problem as a minimization of the cost function

$$[\mathbf{u} - \phi(\mathbf{x})]^\top [\mathbf{u} - \phi(\mathbf{x})] \tag{3.38}$$

Since the measurements are assumed to be independent the measurement covariance is diagonal and the cost function can be expressed as

$$\sum_i ||{}^c\mathbf{u}_i - \phi(\mathbf{p}_c, \mathbf{X}_i)||^2 \tag{3.39}$$

where ${}^c\mathbf{u}_i$ is the measurement on camera c for feature i, and $\mathbf{p}_c$ is the relative pose estimate from imagery and $\mathbf{X}_i$ the estimate of the position of the i^{th} 3D feature point.

For MAP estimation the pose sensors provide a relative pose prior. The initial guess close to the nav-based pose together with the extra cost term that penalizes large deviations from the nav-prior provide a robust two-view relative pose estimate. The cost function being minimized then takes the form

$$\sum_i ||{}^c\mathbf{u}_i - \phi(\mathbf{p}_c, \mathbf{X}_i)||^2 + ||\mathbf{p}_c - \mathbf{p}_{nav}||_{\Sigma_{nav}} \tag{3.40}$$

with the additional term accounting for the relative pose prior, which have the form of a Mahalanobis distance similar to (3.33) with $\mathbf{p}_c$ the relative pose vector estimate from imagery.

3.5.2 Robust estimation

The minimization of squared residuals is optimal in the maximum likelihood sense for zero mean Gaussian noise. While this is considered a good approximation for the reprojection error of true correspondences, there are situations in which the noise does not satisfy the Gaussian assumption. Mismatches can lead to incorrect correspondences with large reprojection errors. In some cases specular reflections or shadows can degrade localization of a feature to the point that the measurement model does not describe the observed errors. These issues are particularly relevant in the multi-view case where data association errors lead to errors that are not obvious in the two view case.

A Gaussian noise model has a distribution with small tails, reflecting that large errors are unlikely. But in practice large errors occur more often than the Gaussian distribution suggests. When this is ignored (and noise is assumed Gaussian) the minimization of squared residuals is strongly affected by outliers.

Least squares minimizes

$$E_{LS}(\mathbf{x}) = \sum_i (r_i(\mathbf{x}))^2 \tag{3.41}$$

where $r_i(\mathbf{x}) = \frac{z_i - h_i(\mathbf{x})}{\sigma_i}$ is the weighted residual for the i^{th} measurement.

M-estimators [133] reduce the sensitivity to outliers by replacing the r_i^2 with a $\rho(r_i)$ that grows more slowly for large r_i while remaining symmetric, positive definite and having a minimum at zero.

$$E_M(\mathbf{x}) = \sum_i \rho(r_i(\mathbf{x})) \tag{3.42}$$

Several choices of $\rho(r)$ have been proposed. The Cauchy M-estimator [126] weighs the residuals in a manner that assumes a Cauchy distribution rather than a Gaussian, which allows for a larger proportion of large errors

$$\rho_C(r) = \frac{c^2}{2} \log(1 + (r/c)^2) \tag{3.43}$$

We use this estimator in all bundle adjustment calculations throughout this thesis. The soft outlier threshold $c = 2.3849$ achieves 95% asymptotic efficiency on the standard normal distribution [133].

Chapter 4

Submap Generation

4.1 Overview

Generating a 3D reconstruction from an extended sequence of images, where each image views only a small fraction of the whole scene, requires additional considerations compared to the two view case. As we advance through the sequence generating structure and motion estimates that are locally consistent, the solution will drift at larger scales. Estimates will be strongly correlated locally but weakly correlated at greater distances [134] because image-based constraints are fundamentally local. If the camera trajectory revisits parts of the scene it is possible to establish additional constraints on the camera poses as long as the data association problem is addressed (recognizing part of the scene as having been visited before). This requires a reliable way of matching images or sections of reconstruction that are not temporally adjacent as well as a mechanism to infer proximity of views that are not temporally adjacent (to avoid exhaustively checking all possible image pairs).

Purely sequential methods are simple and temporally consistent but have limited ability to correct for drift since the covariance between camera poses is not stored permanently [8] [52] [134]. Other incremental methods explicitly keep representations that allow for error distribution over the complete trajectory but only at the expense of complexity that grows quadratically with the number of features and views [71]. Incremental methods are usually implemented as recursive filters (*e.g.*,some form of EKF) in real-time applications or as approximate initializations for small scale bundle adjustment (assuming small drift).

Our objective is to perform a bundle adjustment over a large sequence of images with trajectories that potentially close multiple loops. While the temporal sequence offers an ordering of the imagery, the processing should not be constrained to a temporal order only. While batch solutions based on factorization [120] [92] [115] recover all camera poses and structure simultaneously they tend to have restrictions on the imaging model and typically expect all features to be viewed in all images, which is clearly not applicable to underwater surveys. Also, factorization methods do not address the data association problem of recognizing trajectories that revisit parts of the scene. In addition, a sequential method offers the opportunity to recognize mismatches and some degree of robustness to outliers (which negatively impact factorization methods).

The reconstruction process along a temporal sequence can be viewed as occurring at multiple scales, each scale possessing unique advantages. While the temporal sequence is formed by processing individual image pairs this scale is not the best to determine additional spatial relationships. Subsequences contain information on 3D structure where structure and motion remain significantly correlated. At this larger scale it is easier to recognize that we are revisiting an area since it considers multiple images, with multiple views of interest points and estimates of their location. This is the approach taken to submap matching in the next chapter.

Several approaches to SFM and Simultaneous Localization and Mapping (SLAM) are inspired by local to global or hierarchical methods. Fitzgibbon and Zisserman [29] presented a technique based on triplets of images as the subsequence that allow for the distribution of errors associated with loop closing. They do not discuss, however, how to recognize loop closures. The Atlas framework addresses loop closures for autonomous mapping [13]. Nister [84] refined the approach by adaptively selecting the images in the triplet to improve reconstruction.

4.1.1 Assumptions

Our assumptions in generating submaps include

- A navigation-based pose prior is available. The prior is useful in providing scale and regularization in MAP bundle adjustment of pairs and triplets of images as well as

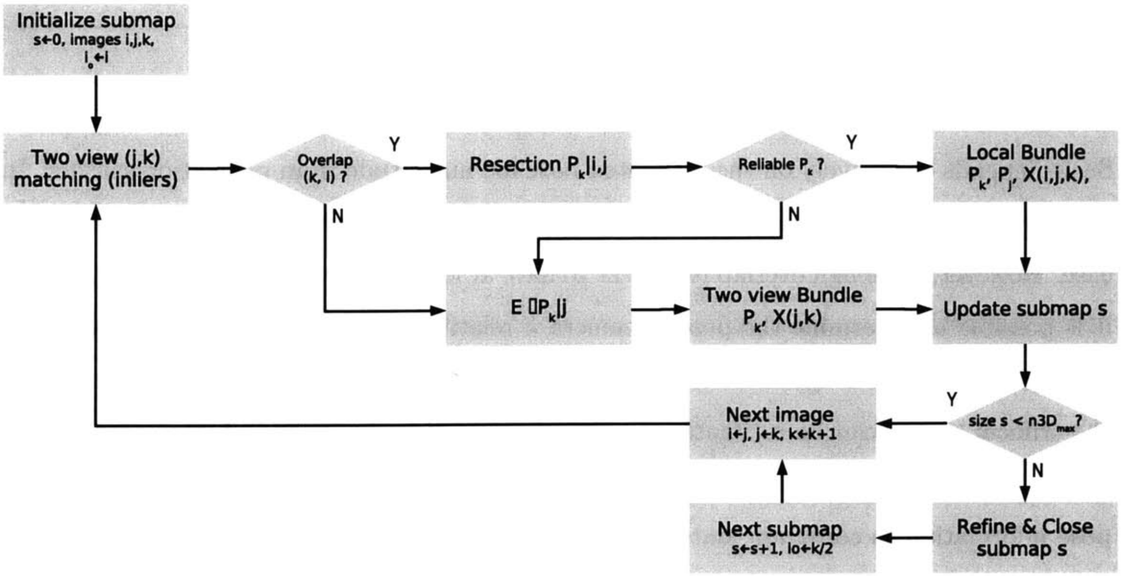


Figure 4-1: Overview of approach to submap generation. The image sequence is processed incrementally. Two view matching proposes putative matches between images j and k. Resection of the pose of k is attempted if some of the correspondences in j have already been matched to the previous image i (there is structure defined by i and j). If the resection is reliable (based on at least 15 points) the triplet i, j, k is bundle adjusted. If there are enough points for resection or there aren't enough inliers, the relative pose is estimated by the two view relative pose procedure of the previous chapter. The new camera pose and structure are then incorporated into the submap. New views are added onto the same submap until the number of 3D features exceeds $n3D_{max}$, (we use 1500-2000 features). At this point the map is bundle adjusted and closed. A new map is initialized with 50% overlap (using the second half of the image sequence).

the complete submap.

- The sequence is mostly unbroken. If temporally adjacent images can not be related we can close the submap and start a new one. The transformation between submaps is determined by the navigation prior.

We propose to use a local to global approach that uses the pose sensors to constrain the camera locations. This chapter concentrates on the submap generation, which corresponds to a local reconstruction based on growing sub-sequences from robust two view matching and resection.

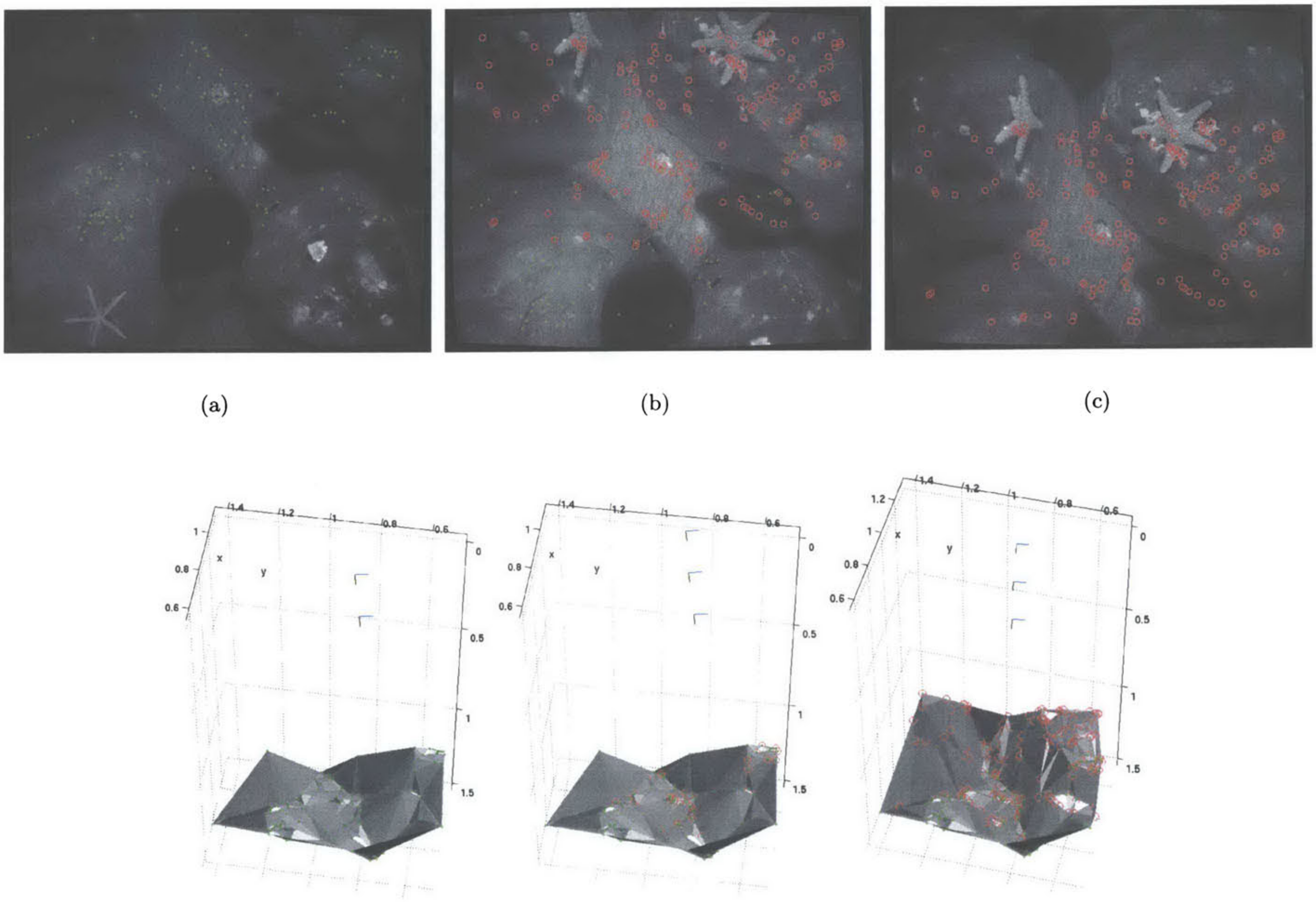
4.2 Resection

A sequence of images can be processed pair-wise using the two view motion estimation. Scale depends exclusively on the estimated baseline magnitude from navigation sensors. In addition, navigation estimates are crucial in resolving ambiguous solutions in the two view case. However, if enough overlap is present to have at least three views of the same structure it is possible to determine the pose of camera k relative to the structure from views i and j. In this case scale is inherited from the triangulated structure. In the uncalibrated case the trifocal tensor can be estimated robustly from proposed correspondences across three images [39]. Though generally effective this approach does not lend itself well to using prior pose information to constrain matches and tends to be more sensitive to critical surfaces.

By establishing putative correspondences between images j and k using the procedure described in §3.2.1 we can then establish putative correspondences between image features in k and the 3D structure associated with the image features of j. Minimal sets of these correspondences are used in a RANSAC loop to determine the pose of k by resection. A proposed resection is evaluated in the RANSAC loop by reprojecting the putatively corresponding structure onto the resected camera. Structure points with a reprojection error below a threshold t are considered inliers. Figure 4-2 illustrates the process by which a submap grows by relying on common features and existing features.

We use a modified version of Fischler and Bolles method for resection [28]. The basic approach consists of selecting a triplet of structure points and their corresponding projections on the camera to be resected. Since the camera is calibrated, the images of the structure points correspond to euclidean rays going through the center of projection, the actual 3D point and its image. Their method first determines the length of the rays from the camera projection center to the 3D points (legs of the tetrahedron). Figure 4-3 illustrates the geometry of the problem.

This can be expressed using the cosine rule in a system of three quadratic equations relating the distances between 3D points R_{ab}, R_{ac}, R_{bc}, the angles between rays (or between 3D points, as measured from the camera), $\theta_{ab}, \theta_{ac}, \theta_{bc}$ and the length of the rays (the unknowns) a, b, c:



(a) (b) (c)

(d) (e) (f)

Figure 4-2: Illustration of growth of a submap based on resection. Images (a) and (b) have corresponding features marked by green dots. The structure and motion implied by those correspondences is illustrated in (d) with units in meters. Images (b) and (c) have correspondences marked by red circles. The features viewed by the three images are marked by both a green dot and a concentric red circle. (e) These features are used in resection to initialize the pose of the third camera. (f) Then the additional correspondences between (b) and (c) are triangulated and the poses refined.

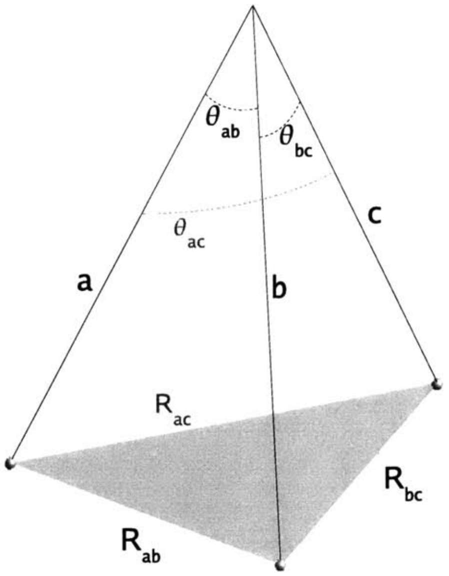


Figure 4-3: Resection geometry. Tetrahedron formed by the camera projection center (top) and three feature points (represented by small spheres at the base of the pyramid). From an image and known correspondences it is possible to measure the angles $\theta_{ab}, \theta_{ac}, \theta_{bc}$ between features. The feature locations are known so distances R_{ab}, R_{ac}, R_{bc} are known. We seek to determine the lengths a, b, c of the rays to the features.

$$R_{ab}^2 = a^2 + b^2 - 2ab\cos(\theta_{ab}) \tag{4.1}$$

$$R_{ac}^2 = a^2 + c^2 - 2ac\cos(\theta_{ac}) \tag{4.2}$$

$$R_{bc}^2 = b^2 + c^2 - 2bc\cos(\theta_{bc}) \tag{4.3}$$

It can be shown that this system has at most four possible solutions. Each solution is included as a possible model in the RANSAC loop (the assumption, which works well in practice, is that other data points can disambiguate the motion). In [28] these solutions are found by reducing the system into a quartic polynomial in one unknown corresponding to the ratio of the two legs of the tetrahedron $x = b/a$. This equation can be solved directly which then allows solving for the actual legs. The resection isn't complete at this stage since determining the length of the rays from resection is equivalent of expressing the position of the 3D points in the reference frame defined by the camera. To resect the camera, *i.e.* to place the camera in the reference frame of the structure, we register the structure in the

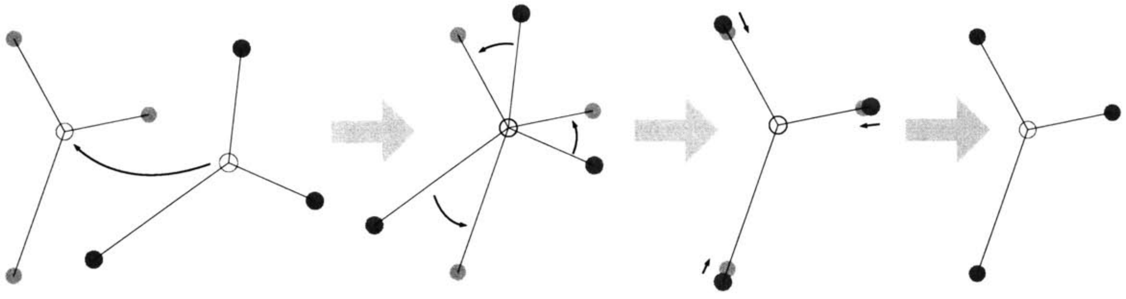


Figure 4-4: Stages of registration source points (dark) to target points (light gray) using Horn's algorithm. From left to right: translation of points to common origin (centroids), alignment of corresponding rays (from centroid to points), adjustment of scale, to yield a final configuration. Adapted from [1].

camera reference frame ${}^k\mathbf{X}$ to the structure in the original frame ${}^0\mathbf{X}$ using Horn's absolute orientation algorithm [43], described briefly in the next section.

4.2.1 Absolute orientation

The goal is to find the similarity transformation (translation, rotation and scale) that aligns the source 3D points ${}^s\mathbf{X}_i$ in the reference frame of the camera onto the ${}^t\mathbf{X}_i$ corresponding points on the target camera t. The process can be understood as a sequence of transformations illustrated for a 2D case in figure 4-4. The sequence of steps are:

1. Translate to a common origin (defined by the centroids).

$$ {}^s\bar{\mathbf{X}}_i = \frac{1}{N}\sum_k^N {}^s\mathbf{X}_k \qquad {}^t\bar{\mathbf{X}}_i = \frac{1}{N}\sum_k^N {}^t\mathbf{X}_k \tag{4.4} $$

$$ {}^s\tilde{\mathbf{X}}_i = {}^s\mathbf{X}_i - {}^s\bar{\mathbf{X}}_i \qquad {}^t\tilde{\mathbf{X}}_i = {}^t\mathbf{X}_i - {}^t\bar{\mathbf{X}}_i \tag{4.5} $$

2. Rotate ${}^s\tilde{\mathbf{X}}_i$ so that the rays from the origin to the source points align with the corresponding ray of the target point ${}^t\tilde{\mathbf{X}}_i$. The rotation is determined by using SVD of the cross covariance matrix of the rays after translation to the common origin [44] [129].

3. Compute scale so that the overall magnitude of rays is the same

$$c = \sqrt{\frac{\sum_{i=1}^{N} |{}^{t}\mathbf{X}_i|^2}{\sum_{i=1}^{N} |{}^{s}\mathbf{X}_i|^2}} \tag{4.6}$$

4. Translate the rotated and scaled source points to the frame of the target.

The resulting transformation sequence that maps the source points onto the target points is thus

$${}^{t}\mathbf{X} = c \cdot {}^{t}_{s}\mathrm{R} \cdot \left({}^{s}\mathbf{X} - {}^{s}\bar{\mathbf{X}}\right) + {}^{t}\bar{\mathbf{X}} = c \cdot {}^{t}_{s}\mathrm{R} \cdot {}^{s}\mathbf{X} + {}^{t}\bar{\mathbf{X}} - c \cdot {}^{t}_{s}\mathrm{R} \left({}^{s}\bar{\mathbf{X}}\right) \tag{4.7}$$

By defining $\mathbf{t} = {}^{t}\bar{\mathbf{X}} - c \cdot {}^{t}_{s}\mathrm{R} \left({}^{s}\bar{\mathbf{X}}\right)$ and $\mathrm{R} = {}^{t}_{s}\mathrm{R}$ the transformation can be expressed as:

$${}^{t}\mathbf{X} = c \cdot \mathrm{R} \cdot {}^{s}\mathbf{X} + \mathbf{t}. \tag{4.8}$$

During resection the rays from the camera are scaled to be consistent with the scale implied by the structure. Therefore, for perfect data $c = 1$. Given noise in measurements and uncertainty in the structure we expect $c \approx 1$.

4.2.2 Robust Resection

The resection approach described in §4.2 is at the core of a RANSAC loop applied to the the proposed correspondences between 2D features in the view to be resected and already existing 3D features. We have one constraint we can easily enforce at this stage: the scale, c, implied by the pose must be nearly unity. If this is not the case it is not necessary to check reprojection errors. The steps for the robust estimation of the pose of the view relative to the structure and the correct correspondences are:

- Start with the set S of potential correspondences between 2D features and 3D structure, the set of inliers $S_{in} = \emptyset$ and the number of inliers $n_{in} = 0$.
- Repeat for N trials :
    - Randomly select a 3 point subset p from the potential correspondences.

- Calculate $\mathrm{P}(p) = [c\mathrm{R}|\mathbf{t}]$ the pose of the camera implied by the tetrahedron formed by subset (§4.2). There can be up to four solutions. For each solution, if $(0.9 < c < 1.1)$:
    - Calculate the reprojection errors d for all putative matches by projecting the structure onto the proposed camera.
    - Determine $S_{in}(p)$, the set of inliers according to $\mathrm{P}(p)$, as the set of matches with $d < t$ pixels. The number of inliers according to $\mathrm{P}(p)$ is $n_{in}(p)$.
    - If $n_{in}(p) > n_{in}$ then $n_{in} \leftarrow n_{in}(p)$ and $S_{in} \leftarrow S_{in}(p)$

The RANSAC algorithm produces a pose that best explains most of the data in the sense that the reprojections errors are less than the threshold t. Under the assumption that 2D features are localized with a standard deviation of σ, the distance squared between the measured and the estimated 2D feature location follows a χ_2^2 distribution. The cumulative chi-squared distribution $F_2(t^2) = \int_0^{t^2} \chi_2^2(x)\, dx$ includes 99% of the inliers for $t^2 = 9.21\sigma^2$ or $t = 3.03\sigma$. In practice we assume $\sigma = 1$ and have observed that the results are not overly sensitive to t. The number of iterations N is calculated adaptively based on the current estimate of the fraction of outliers [39], typically being tens or hundreds of iterations. We use a simple count of points in common between three views (typically 15 points) to determine if resection is not possible or deemed unreliable (i.e. not enough common points, or not enough inliers exist between the three views). In either case we switch to the two-view relative pose estimation to incorporate the latest view onto the submap. Figure 4-1 illustrates this decision process.

4.3 Local Bundle Adjustment

The resection stage produces the approximate pose of the camera that is most consistent with the proposed correspondences between image points and 3D structure. To refine the pose we turn to Zhang and Shan's local bundle adjustment method [134] for inspiration. This is a variant of sequential methods that is shown to approximate the optimal global bundle adjusted solution for sequences of images with short feature tracks while significantly reducing computational costs. The approach considers the bundle adjustment problem of the latest three views while reducing the free parameters to the latest camera pose and the feature points it views. It takes advantage of points seen in the three views as well as those in the last two views. Though efficient, this technique does not handle uncertainty in a consistent fashion. By considering the first two cameras of the triplet fixed, the uncertainty of the estimates (third camera and structure) are expressed relative to a frame fixed implicitly by the relative pose of the first two cameras (including scale). In the context of maximum a posteriori estimation we have prior information of the relative pose between the first and second camera as well as between the second and third cameras. We choose to fix the origin on the frame of the first camera and leave the second and third cameras to be adjusted. In essence we solve the MAP estimate of the trifocal tensor as a way to produce an estimate of the latest pose and the uncertainty in pose and structure.

Given three views $0, 1, 2$ and the measured (noisy) correspondences between the views $\{{}^0\mathbf{u}_i \leftrightarrow {}^1\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, and the correspondences between pairs of views $\{{}^1\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, $\{{}^0\mathbf{u}_i \leftrightarrow {}^2\mathbf{u}_i\}$, $\{{}^0\mathbf{u}_i \leftrightarrow {}^1\mathbf{u}_i\}$, under the assumption of isotropic Gaussian noise corrupting the interest point locations, it can be shown that the MLE for the the poses and structure minimizes the sum of squared reprojection errors:

$$\sum_{c=0}^{2}\sum_{i} d({}^c\mathbf{u}_i, {}^c\hat{\mathbf{u}}_i)^2 + \sum_{c=1,2}\sum_{j} d({}^c\mathbf{u}_j, {}^c\hat{\mathbf{u}}_j)^2 + \sum_{c=0,2}\sum_{k} d({}^c\mathbf{u}_k, {}^c\hat{\mathbf{u}}_k)^2 + \sum_{c=0,1}\sum_{l} d({}^c\mathbf{u}_l, {}^c\hat{\mathbf{u}}_l)^2 \tag{4.9}$$

where $d(\cdot,\cdot)$ represents the Euclidean distance, ${}^c\hat{\mathbf{u}}_m$ are the estimated ideal correspondences (*i.e.*, before corruption with Gaussian noise) for camera c, and m the index into the correspondence set. The role of the structure is implicit in (4.9). More explicitly, we have that the projection of a 3D point $\mathbf{X}_i$ onto a camera c with pose $\mathbf{p}_c$ is ${}^c\hat{\mathbf{u}}_i$:

$${}^c\hat{\mathbf{u}}_i = \phi(\mathbf{p}_c, \mathbf{X}_i) \tag{4.10}$$

Using the camera projection (4.10) we expand (3.27)

$$\min_{\mathbf{p}_1,\mathbf{p}_2,\{\mathbf{X}_i\},\{\mathbf{X}_j\},\{\mathbf{X}_k\},\{\mathbf{X}_l\}} \sum_{c=0}^{2} \sum_{i=1}^{N_{012}} ||{}^{c}\mathbf{u}_i - \phi(\mathbf{p}_c, \mathbf{X}_i)||^2 + \sum_{c=1,2} \sum_{j=1}^{N_{12}} ||{}^{c}\mathbf{u}_j - \phi(\mathbf{p}_c, \mathbf{X}_j)||^2$$
$$+ \sum_{c=0,2} \sum_{k=1}^{N_{02}} ||{}^{c}\mathbf{u}_k - \phi(\mathbf{p}_c, \mathbf{X}_k)||^2 + \sum_{c=0,1} \sum_{l=1}^{N_{01}} ||{}^{c}\mathbf{u}_l - \phi(\mathbf{p}_c, \mathbf{X}_l)||^2 \quad (4.11)$$

The MAP estimate adds cost terms based on relative pose prior (from pose sensors) similar to the ones used in the relative pose MAP estimation, which biases the solution to the scale implied by the navigation sensors.

$$||\mathbf{e}_{01}||_{\Sigma_{nav}} + ||\mathbf{e}_{12}||_{\Sigma_{nav}} \quad (4.12)$$

where using the composition notation from Smith, Self and Cheeseman [114] the discrepancy between vision and navigation-based relative pose is given by

$$\mathbf{e}_{ij} = \ominus\hat{\mathbf{x}}_{ij} \oplus \mathbf{x}_{ij}^{nav} = \ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i \oplus \mathbf{x}_{ij}^{nav} \quad (4.13)$$

and the weighted error is

$$||\mathbf{e}_{ij}||_{\Sigma_{nav}} = e_{ij}^{\top}\Sigma_{ij}^{-1}e_{ij} \quad (4.14)$$

where Σ_{ij} corresponds to the estimated covariance of e_{ij} propagated from the covariance of $\mathbf{x}_{ij}^{nav}$.

4.4 Submap size

We have proposed using reconstructions of subsequences as the basic unit from which to derive the network of images and feature points for a final global adjustment. An important issue in this approach is setting the size of submaps. There are multiple implications and trade-offs that merit consideration. Processing the temporal sequence can be assumed to have a fixed cost per image (two view or resection). We advocate performing a bundle adjustment over all views and feature points in a submap at the time of its closure. This guarantees that self consistent maps should improve matching. We use the local bundle adjusted sequence as an initial guess for a proper bundle adjustment. As described in the next chapter, once the image sequence has been processed into submaps these are matched to find any additional constraints on the network of submaps (and therefore images). The size (or number) of submaps affects the complexity of multiple bundle adjustments, the reliability of matching submaps, and the complexity of the resulting network of submaps. We discuss these points and suggest that it suffices to close submaps based on the number of features they contain. Our current implementation can perform bundle adjustment and submap matching in a few seconds for submaps with fewer than 2000 3D features; there is a rapid increase in runtime for larger submaps. Thus we choose to create submaps with at least three images and limit the number of 3D features in each submap to be 1500-2000.

4.4.1 Bundle adjustment complexity

Each step in a sparse bundle adjustment of N features and M views has complexity $\mathcal{O}((N+M)M^2)$, linear in N and cubic in M associated with the inversion of the sparse normal equations [71]. If we break down the problem into S submaps with no overlap and perform bundle adjustment on each submap individually each bundle adjustment is of complexity $\mathcal{O}((1/S)(N+M)(M/S)^2)$ assuming that the features and views are evenly distributed in each submap. The complexity for the total sequence (the bundle adjustment of S submaps) is $\mathcal{O}(\frac{(N+M)M^2}{S^2})$. Therefore, S smaller bundle adjustments reduces the overall complexity in proportion to S^2. In the presence of overlap between submaps, the complexity grows linearly with the overlap fraction v. We can show this by defining the actual number of submaps as $S_v = S/(1-v)$, the complexity of processing one submap does not change but the overall complexity is $\mathcal{O}(\frac{1}{1-v}\frac{(N+M)M^2}{S^2})$. This result suggests that, if a sequence is to be split into submaps and each submap bundle adjusted, then there are significant computational savings to be had by using smaller maps.

4.4.2 Uncertainty in Structure and Motion

An incremental reconstruction can drift relative to the 'true' structure because the imaging process relates only features that are spatially close to each other (local correlation). The longer the sequence used in the submap, the greater the possible deviation from the 'true' geometry. If submaps are to be registered as sets of 3D points related by a similarity transformation it is necessary to consider the effect of drift on the reconstruction. Actual drift can be quantified if ground truth is available. In general this is not the case, and we choose to use the estimate of covariance in 3D feature positions as an indication of possible drift.

Geometrical quantities (such as 3D structure) derived from multiple camera measurements are expressed in a reference frame (including scale) that can be freely chosen. This *gauge freedom*, or coordinate frame ambiguity, is inherent to the imaging process. Image projections do not depend on the chosen gauge while reconstructed results in different gauges are equivalent modulo the gauge [70] since they produce the same projections regardless of the reference frame. However, the choice of reference frame will affect the apparent covariance of structure and motion [78][126]. We note that normally the reference frame is defined in the reconstructed space of structure and motion (for example, coincident with the frame of the first camera). Since all reconstructed quantities are uncertain the frame is also uncertain yet the elements used to define the frame appear perfectly known. For example, if the frame is fixed to the first camera, the uncertainty of the first camera relative to this frame will be zero (since the two frames are coincident) while other reconstructed quantities may appear to have increased uncertainty.

Our local bundle adjustment procedure fixes part of the gauge (scale) implicitly through the relative pose prior provided by navigation sensors. The reference frame origin and orientation is coincident with the first camera.

The covariance of poses are calculated from the sparse bundle adjustment by the block inversion of the approximate Information matrix $(J^{\top}J)$. The covariance of pose is expressed relative to the frame fixed to the first camera with scale implied by the navigation-based relative pose priors (zero gauge freedoms). This is convenient as camera pose uncertainty is expressed relative to the first camera, illustrating the trend towards higher uncertainties with number of images as illustrated in Figure 4-6. Since the succeeding submap is initialized using one of the cameras of the current submap, this representation also provides a direct estimate of the relative transformation and uncertainty between the current submap and its neighbor.

For registration purposes the uncertainty of reconstructed 3D points should reflect the quality of the triangulation rather than an arbitrary choice of reference frame. Points that are triangulated more precisely should be weighed more when registering a set of points. We choose to express the uncertainty of 3D points with six gauge freedoms (orientation and translation). This is achieved by simply eliminating the rows in the Jacobian corresponding to the equations that fix the origin to the first camera before calculating the covariance of the poses and structure by using the pseudo inverse (keeping the six smallest singular values as zeros) [78].

4.4.3 Submap Matching and Network Complexity

To propose putative matches based on similarity between submaps i and j takes time $\mathcal{O}(N_i N_j)$ where N_i and N_j are the number of features in each submap. Since $N_i = \mathcal{O}(N_j)$ we realize that registering submaps by similarity is $\mathcal{O}(N_i^2) = \mathcal{O}((N/S)^2)$. But this has to be considered in the context of the number of matches that have to be performed. Matching all submaps to all submaps is $\mathcal{O}(S^2)$ which would imply that the lower costs of matching smaller maps are offset by the need to match more maps. But for a sparse network where most nodes have edges to a few adjacent nodes, as in a survey with a moving vehicle, we can expect that $\mathcal{O}(S)$ edges exist and that a reasonable matching technique will also perform $\mathcal{O}(S)$ matches. The overall complexity of matching for the sparse network case is $\mathcal{O}(N^2/S)$ which also suggests using more (smaller) submaps will save effort at the submap matching stage. In terms of reliability of matching, matching smaller maps reduces the chances of false positives by requiring less descriminating ability from similarity-based descriptors.

An indirect way of studying this issue is to generate submaps and register them according to the procedure described in the next chapter. Given feature points in correspondence the registration quality can be recalculated using a starting from a small subset and growing along the submap to include all correspondences. If submaps represent a true rigid body reconstruction of the environment, the average registration error of considered points should be roughly constant regardless of the number of points used. If the submaps tend to distort at larger scales the registration quality should degrade as more of the submaps are brought into alignment. The results of figure 4-7 suggest that submaps remain rigid for

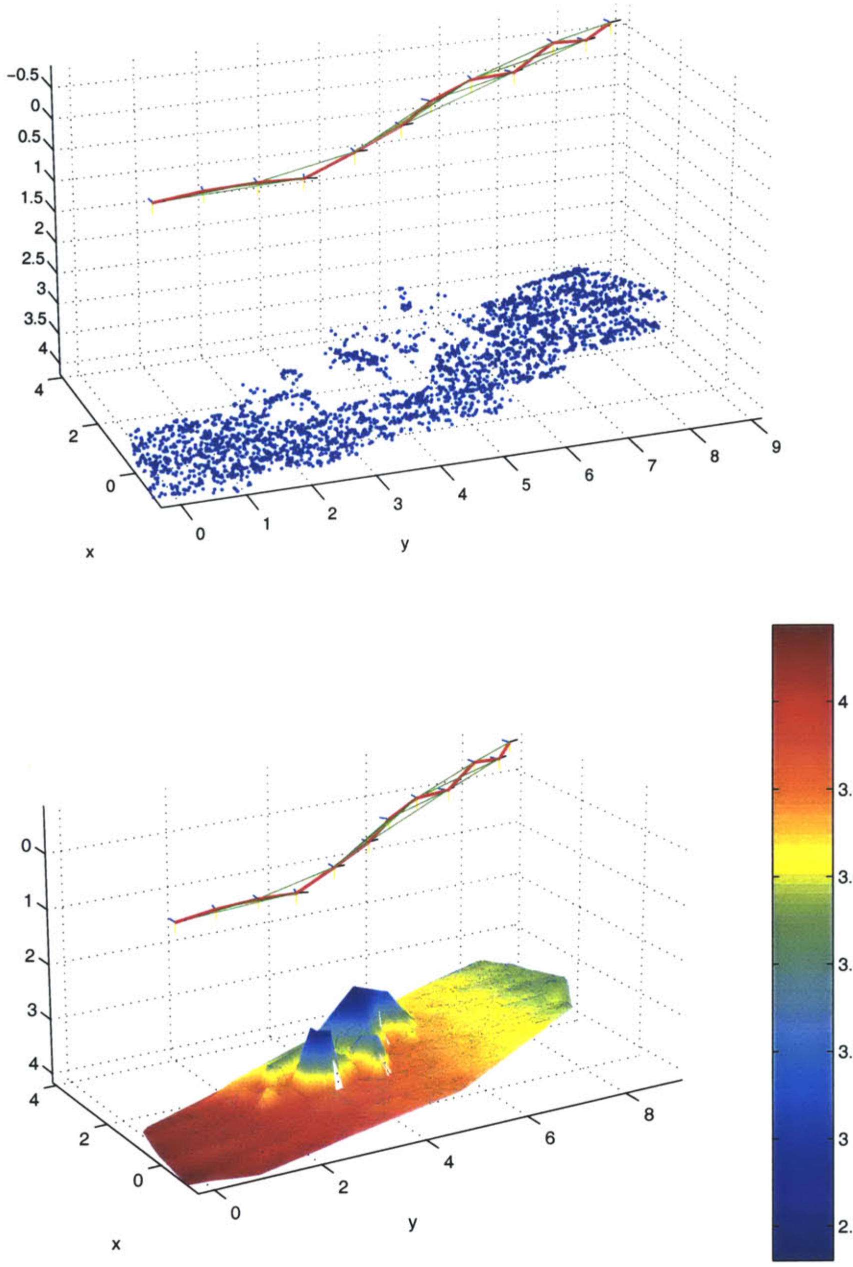
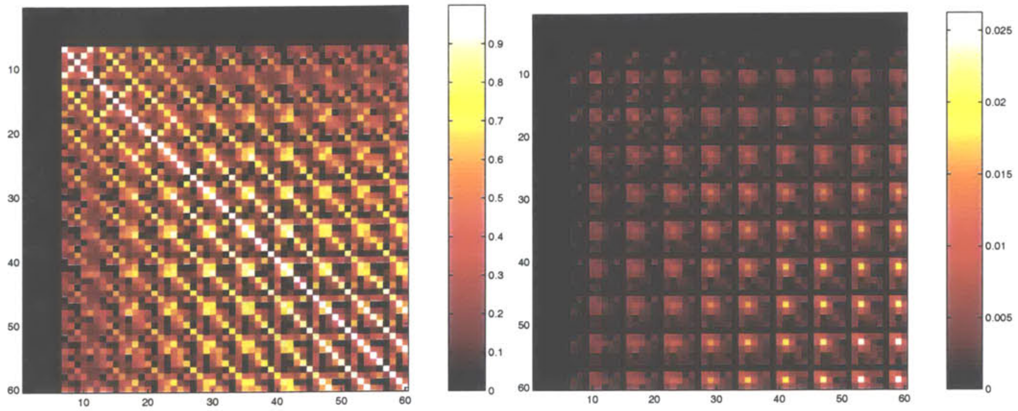


Figure 4-5: Two views of the MAP bundle adjustment of an example of an incremental reconstruction consisting of 12 images and close to 3000 points. Cameras are represented as reference frames (X,Y,Z axis as blue,black,yellow). The temporal sequence is from left to right. Temporally adjacent frames are connected by a red line. Spatially adjacent frames (determined through resection) are linked in green. (Top) The dots represent the estimated position of 3D points in the reference frame defined by the first camera. (Bottom) For ease of interpretation, a surface has been fit through the points using a Delaunay triangulation. The surface is color-coded according to the Z coordinate.



(a)

(b)

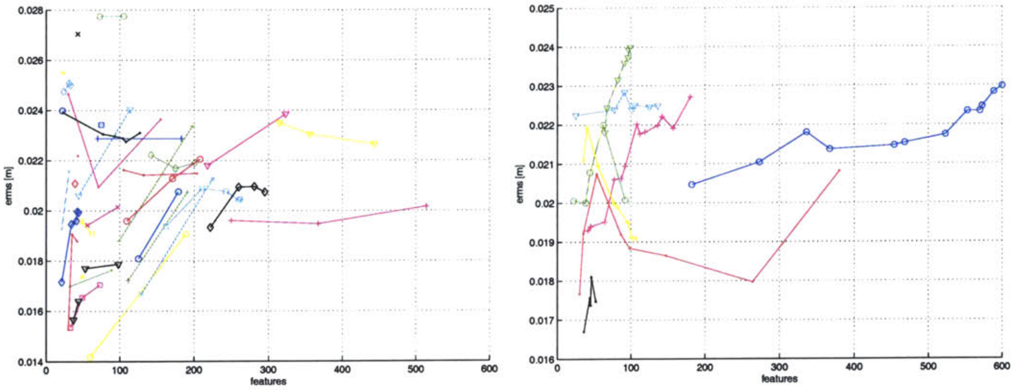
Figure 4-6: (a) Absolute value of the correlation coefficient (normalized covariance) for a large submap (3000 features). Every six rows (or columns) correspond to the pose parameters of a camera $(\phi, \theta, \psi, x, y, z)$. The first camera is fixed which produces zero covariance. (b) The element by element square root of the absolute covariance matrix. The weak coupling between the first images and the last images is apparent. From (b) it is clear that the highest uncertainty is associated with xy.

practical sizes. For typical data and sensors of this application it appears that the increase in complexity is far more significant than drift in determining the size of submaps. In addition, more submaps offer more degrees of freedom or 'hinges' on which to distribute error when closing a loop. This should provide a better initialization for the final bundle adjustment.

4.5 Submap Closing

Once a submap contains enough 3D features it is closed and a new submap is started. The structure associated with the most recent half of the cameras in the map being closed is used to start the new submap. There is a trade-off between the number of submaps and improving the chances of matching across tracklines. In practice, an overlap of around 30% to 50% provides a good balance between the number of maps and improved matching.

We perform a final bundle adjustment using all poses and prior pose information on the submap before closing it. A sparse bundle adjustment adjustment routine [39] [126] is used to minimize the cost function



(a)

(b)

Figure 4-7: Registration error (RMS) between multiple submap pairs as a function of number of features used. (a) For a sequence of images the maximum submap size was set to 1200 features. Each track is associated with a submap pair and illustrates the evolution of the RMS registration error as the set of features considered increases from those implied by the first pair of images in the submap to all corresponding features. (b) The evolution for the same image sequence with maximum submap size of 3000 features. There are fewer and longer tracks since there are fewer submaps (and each submap is larger). If submaps correspond to rigid bodies then these curves should be approximately flat. If submaps are close to rigid bodies for small scales and 'drift' for larger scales the general trend should be for an increase in RMS error as more features are considered.

$$\sum_{c} \sum_{i} ||{}^{c}\mathbf{u}_i - \boldsymbol{\phi}(\mathbf{p}_c, \mathbf{X}_i)||^2 + ||\mathbf{e}_{c,c+1}||_{\Sigma_{nav}} \tag{4.15}$$

where $\mathbf{p}_c$ is the pose estimate from imagery for the c^{th} camera, $\mathbf{e}_{c,c+1}$ is the residual vector between the relative pose estimate from navigation sensors and imagery (4.13), and $\mathbf{X}_i$ the estimate of the position of the i^{th} 3D feature point.

This is the same procedure used on the triplets (after resection) but considers all views. The initial guess is provided by the incremental submap.

The relative pose between the new submap and the previous submap corresponds to the pose (in the reference frame of the submap being closed) of the camera that becomes the origin of the new submap.

Chapter 5

Global representation

5.1 Overview

The submap generation stage of the previous chapter yields a sequence of overlapping submaps and estimates of the relative transformations between adjacent submaps. In order to produce a final, consistent bundle adjusted representation of the structure it is necessary to place submaps in a common *global* frame. It is also important to recognize instances of the same 3D points in multiple submaps (due to parallel tracklines or loop closures) so that they are reconstructed only once in the final representation. This chapter discusses the refinement of spatial relationships between submaps before attempting to produce a global reconstruction.

The problem of transitioning between local and global representations is related to the one confronted by local to global mapping and localization approaches [14],[3] in which a robot explores and maps without an explicit global map. The global, self consistent map is established only in post-processing. This thesis assumes that an underwater vehicle performs a preprogrammed survey relying on uncertain navigation. While sufficient to process data temporally (as in the previous chapter) the overall uncertainty in navigation motivates using redundancy at local levels (*i.e.*, overlap) to form a globally consistent set of poses.

The spatial relationships we know (between temporally adjacent submaps) and the ones we seek (between spatially adjacent but temporally non-adjacent submaps) can be ab-

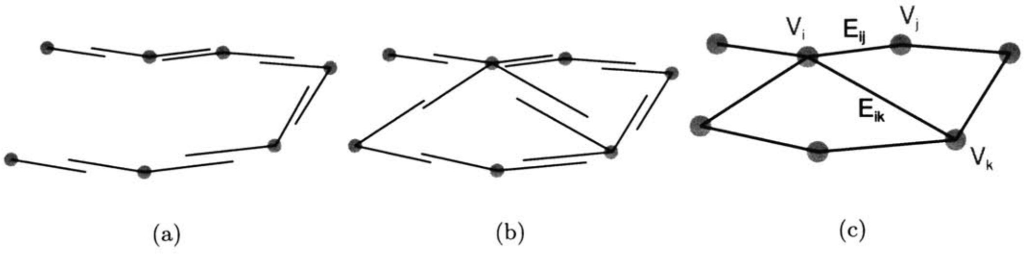


Figure 5-1: Placing nodes (Gray circles) in a globally consistent frame. From relative transformations (black links) in a temporal sequence (a), to proposing and verifying new additional links (b) to a network with nodes consistent with the relative transformations (c).

stracted into a graph, where each submap reference frame corresponds to a node and each edge corresponds to a coordinate frame transformation between submaps. More precisely, the submaps and transformations can be represented as a graph $G = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{V_i\}_1^N$ the set of vertices and $\mathcal{E} = \{E_{ij}\}_1^M$ represents the set of edges. A directed link between nodes V_i and V_j is denoted by E_{ij}. Figure 5-1 illustrates these concepts.

The set of nodes that link to node V_j is given by $S(V_j) \equiv \{V_i \in \mathcal{V} | E_{ij} \in \mathcal{E}\}$. Each node V_i has an uncertain frame or pose x_i associated to it and each edge E_{ij} has an uncertain relative transformation x_{ij} associated to it. Such topological representations are common in the SLAM community [61] [104].

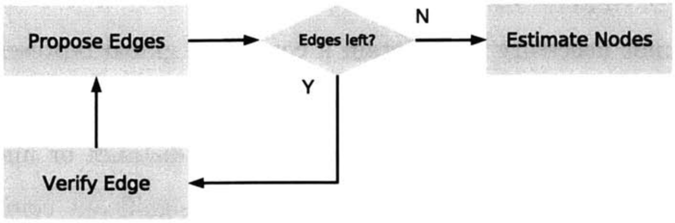
In terms of the graph, generating a global representation from submaps can be broken into three distinct processes:

- Edge proposal (§5.3). Discover potentially new edges starting with only the temporal links $\mathcal{E}_{initial} = \{E_{12}, E_{23}, \ldots, E_{N-1,N}\}$.

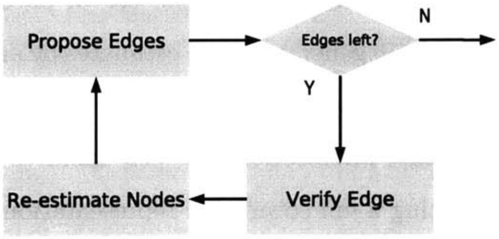
- Edge validation (§5.2). Check the proposed edges against the known data or constraints (for example, map-matching).

- Node estimation (§5.4). Generate globally consistent estimates of the frames associated with $\mathcal{V}$.

We assume that edge validation is computationally intensive, and that it is therefore desirable to propose only edges which are likely to be valid. Otherwise, we could simply propose that all nodes are connected to all nodes and verify which of all possible $\mathcal{O}(N^2)$



(a)



(b)

Figure 5-2: Consistent estimates of nodes (the submap frames in a global reference frame) depend on establishing additional links between nodes. (a) These can be proposed and verified entirely in 'relative space' based on the composition of links before calculating the node frames. (b) Alternatively, the node estimates can be refined as new links are proposed and verified, and used to propose new edges.

edges exist.

The proposal of probable edges requires some knowledge of the node positions such as their means and covariances. This suggests a 'global space' approach that interleaves edge proposal (and validation) with node estimation. In essence, the node estimates are generated from the temporal sequence and updated as new links are discovered (figure 5-2 b).

Node estimation is computationally intensive as nodes must be placed in a global frame while satisfying all the spatial relationships implied by the edges. Addition of a new edge can dramatically alter the placement of some nodes (for example, in the case of loop closures) and can be subject to convergence to local minima.

Further reflection leads to the realization that edge proposal is essentially a local problem – by definition, submaps overlap only if the acquiring cameras were close by. Composition of transformations implied by edges allows us to position an individual submap relative to any other submaps. Approaches that operate in 'relative space', such as Atlas[12], exploit the fact that it is not necessary to maintain a consistent global representation to propose new edges. That is, it may often be sufficient in many applications to use a 'reasonable' path between nodes, such as the shortest path under a distance or uncertainty measure, instead of attempting to fuse information. This offers significant computational savings compared to global solutions that attempt to update the node estimates.

There are some suboptimal (in the sense of the use of uncertainty) approaches that straddle this definition. Sharp [109] generates node estimates by enforcing cycle consistency. This requires keeping track of all cycles and distributing error in a manner that does not use uncertainty optimally. Covariance intersection (CI) [54] operates in global space but nodes are updated without constructing a full covariance representation. CI is often criticized for producing very conservative estimates of uncertainty. This, on its own, is not a significant problem when using nodes for initializing a bundle adjustment. However, we have observed that high uncertainty nodes gain little from precise relative measurements. This is typical of networks where there is uncertainty in the overall orientation and position of the network and suggests that CI alone cannot solve our problem.

This chapter discusses the use of the transformations between submaps to place submaps in a global frame, as well as a procedure to propose and verify additional spatial relationships between submaps. Regardless of how links are proposed, either by a local (such as shortest path) or global approach (covariance intersection) links are verified in the same fashion.

5.1.1 Assumptions

The underlying assumption in this approach is that uncertainty in the frames and transformation (nodes and edges respectively) can be characterized adequately to allow proposing additional edges (submaps with common features). This implies that if overlap does exist between submaps the uncertainty in their relative position should suggest the overlap.

5.2 Edge Validation

Establishing additional edges between submaps implies determining the transformation that maps common 3D points in the reference of one submap to the other. Using submaps to match 'across track' (or spatially adjacent but temporally discontinuous) views offers significant advantages over two view matching:

- Matching images with knowledge of structure offers a stronger constraint than epipolar geometry (no loss of scale).

- Matching sequences of images typically increases the number of matching features in low overlap situations (*e.g.* parallel tracklines). Individual image pairs with low overlap might fail to match or match unreliably (too few feature points).

- Since each 3D point in a submap is imaged at least twice, there is redundancy in the appearance of 3D points that can be exploited when matching features across submaps.

We assume that submaps are internally consistent, given that each is bundle adjusted before being closed. The scale of each submap is derived from the multiple baseline measurements from the navigation system. Because the vehicle acquired the imagery using the same set of instruments we expect that all submaps will have approximately the same scale.

Registering two sets of 3D points with unknown correspondences is traditionally performed with Iterative Closest Point (ICP) techniques [11] [132]. In its strictest sense, ICP is only a refinement of the transformation between two sets of 3D points that are already relatively well aligned and in which all points in one set have a match in the other. ICP variants [101] extend the domain of applicability but remain unreliable when the initial guess is poor, when there is low overlap (*i.e.* a small fraction of common points) between the sets of 3D points or when there is low variability in the 3D structure. In practice, ICP is most successful with dense data sets (typically laser scanned) while the submap matching problem we confront involves relatively sparse sets of points.

One way of improving the robustness and range of applicability of ICP-type algorithms is to associate descriptors with the 3D features. The additional information provided by the descriptors makes the correspondence proposal less dependent on the initial alignment of the data sets. Some descriptors are based on the local geometry around a feature such as curvature and oriented points [37] or spinmaps [53]. Other sensing modalities can also provide descriptors, such as image patches or color [35].

While the sparse set of 3D points contained in the submaps do not consistently offer discriminating structure, the very fact that they exist as 3D points implies that their appearance in multiple views is characteristic enough to effectively establish correspondences (and be reconstructed by the SFM algorithm). We extend the feature description and similarity based matching between images to matching submaps by relying on the appearance of 3D points to propose corresponding features between submaps. The average of the descriptors of the 2D neighborhoods in all views (*i.e.*, observations) is used as the appearance of the 3D point. The underlying assumption is that a similarity measure which was effective to match 3D points along track will also be effective when matching across submaps. This requires descriptors that are robust to lighting changes, or scenes in which lighting does not change significantly between submaps.

5.2.1 3D Feature Descriptors

For similarity-based matching purposes, we propose to describe a 3D feature by the average of all acquired 2D views of the neighborhood around the feature. We assume that for each view the neighborhood is represented in a canonical frame as described in Chapter 2 (*i.e.* an affine invariant region mapped onto a circle with orientation known to a few degrees from navigation).

Given an image patch formed by averaging N image patches $\bar{f}(x,y) = \frac{1}{N}\sum_k^N f_k(x,y)$ (all in the canonical frame), with the moments for each patch f_k

$$A(k)_{nm} = \frac{n+1}{\pi} \int\int_{x^2+y^2\leq 1} f_k(x,y) V^*_{nm}(x,y)\, dx\, dy \tag{5.1}$$

The moments for the average patch are given by

$$\begin{aligned}\bar{A}_{nm} &= \frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1}\bar{f}(x,y)V^*_{nm}(x,y)\,dx\,dy & (5.2)\\ &= \frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1}\left(\frac{1}{N}\sum_k^N f_k(x,y)\right)V^*_{nm}(x,y)\,dx\,dy & (5.3)\\ &= \frac{1}{N}\sum_k^N\left(\frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1} f_k(x,y)V^*_{nm}(x,y)\,dx\,dy\right) & (5.4)\\ &= \frac{1}{N}\sum_k^N A(k)_{nm}. & (5.5)\end{aligned}$$

Due to superposition and linearity the moment of an average image patch corresponds to the average of the moments. Thus for a 3D feature $\mathbf{X}$ viewed by N cameras, with an extracted 2D region f_k from the k^{th} camera, and associated feature descriptor $\mathsf{s}(f_k)$ (§2.5.2) we construct a descriptor for the 3D feature as the average of all 2D descriptors:

$$\mathsf{s}(\mathbf{X}_i) = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathsf{s}(f_{ik}) \qquad (5.6)$$

5.2.2 Similarity measure

Putative 3D feature correspondences between different submaps are proposed based upon similarity of descriptors. The measure of §2.5.2 (which approximates the cross correlation between patches in the invariant frame) is used to propose matches. No pose prior is used in this case given that the relative transformation between temporally distant submaps can be very uncertain due to the drift inherent in the dead-reckoned navigation and in the sequential structure from motion.

Since submap sizes are limited to less than 2000 feature points, matching all 3D feature points against all other feature points presents a similar computational cost to that associated with matching two images.

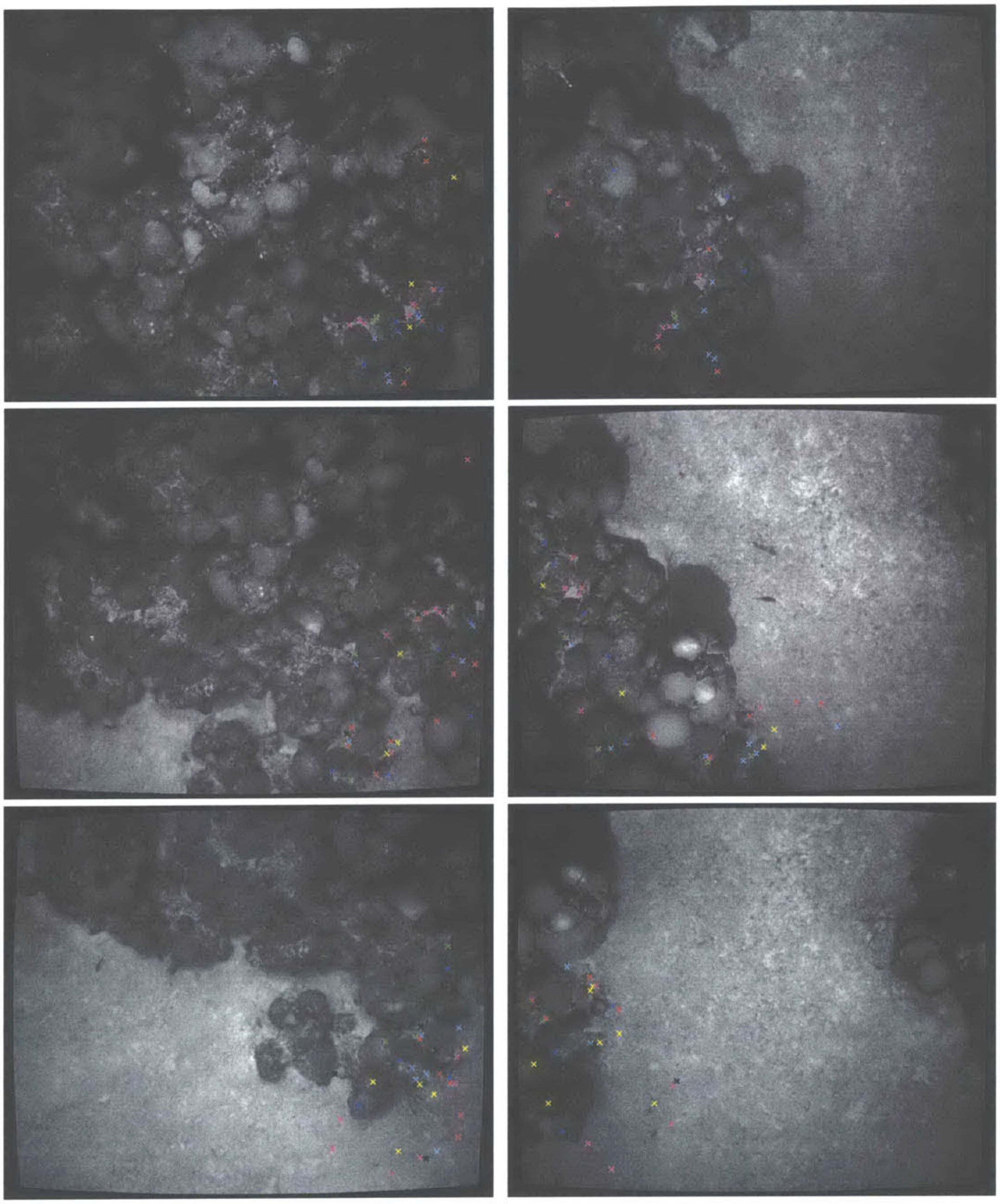


Figure 5-3: Improved matching through use of submaps. Each column contains images from a different submap (the images on the left are from the first trackline of the survey, the images on the right are from the second trackline and rotated 180° to facilitate comparison). It is difficult to reliably find corresponding features between any pair of images across the columns. But when columns are considered as a whole (*i.e.*, as submaps) it is easier to find common features and to reliably estimate the transformation between submaps. Corresponding features found through submap matching are shown as 'x'. 3D features are color-coded consistently across all images.

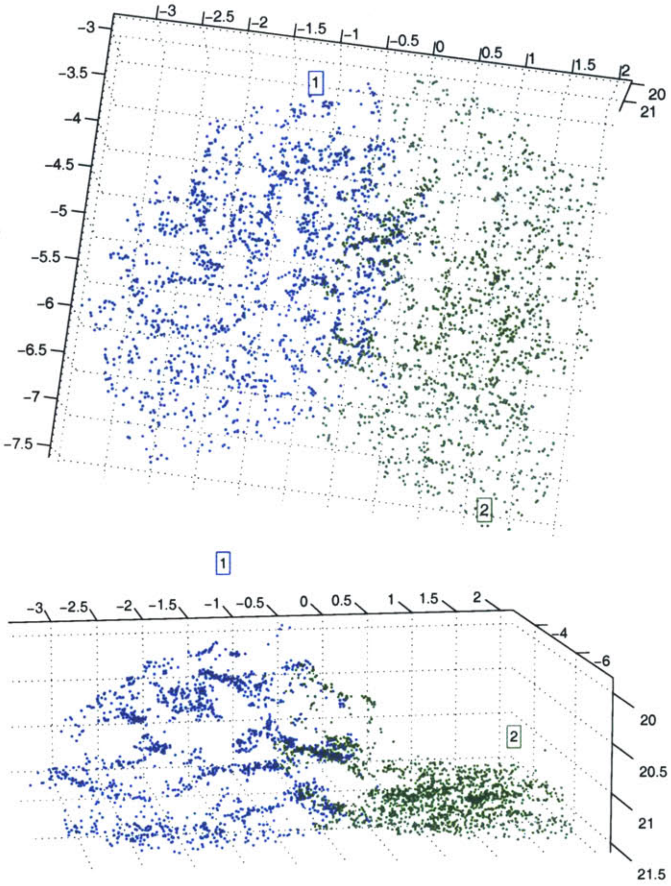


Figure 5-4: Views of the registered submaps that contain the images in figure 5-3. The blue dots correspond to the 3D features of the submap on the first trackline of the survey (*i.e.* the images on the left column of figure 5-3). The green dots correspond to features in a submap on the second trackline of the survey (*i.e.* images in the right column in figure 5-3).

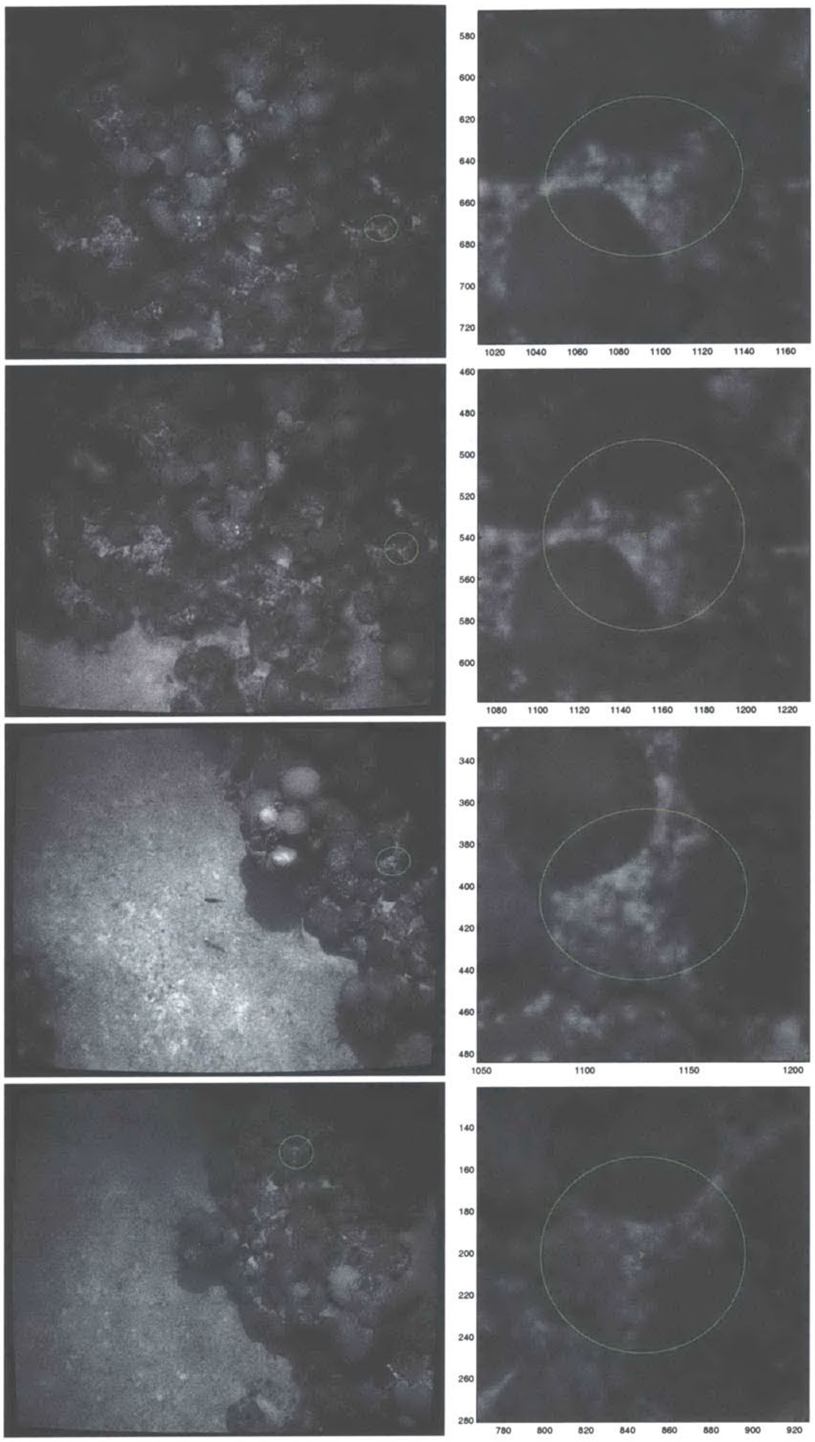


Figure 5-5: Multiple views of a 3D feature:(left column) the image and the feature neighborhood extracted as described in §2.5.2 and (right column) a detail of around the feature point. The top two rows correspond to images that belong to a submap on the first trackline of the survey, the bottom two rows are from a submap from the second trackline.

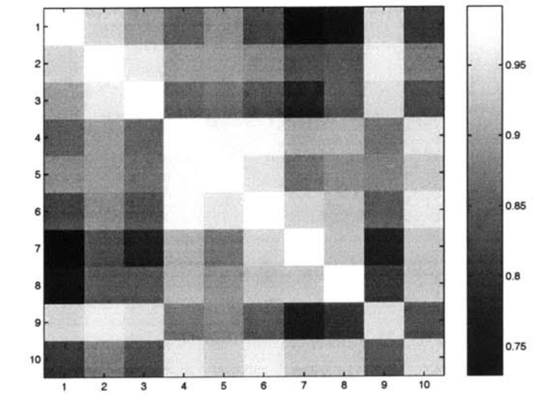


Figure 5-6: Similarity scores between descriptors corresponding to different views of the feature of figure 5-5. The first three entries correspond to views in the first trackline submap, the next five entries are views of the same feature in the second trackline submap. Similarity is highest along the diagonal (corresponding to self-similarity, as expected it is close to 1 since the similarity score based on moments only approximates the correlation between the image patches). Similarity is higher between views that belong to the same submap (usually above 0.9) than across submaps (less than 0.9). There is also more variability in the scores when matching across submaps. The ninth entry is the average descriptor of the feature for the first trackline submap and the tenth entry corresponds to the average descriptor of the feature for the second submap. As expected, the average descriptor is similar to the descriptors in its own submap. The similarity between the average descriptors (9,10) and (10,9) (across submaps) is within the range of the individual matches.

5.2.3 3D to 3D matching

Given putative correspondences between 3D points from two submaps, we seek to register the two sets of 3D points. The goal is to find the similarity transformation (translation, rotation and scale) that aligns the 3D points ${}^{s}\mathbf{X}_i$ from source submap s onto ${}^{t}\mathbf{X}_i$, the corresponding points on the target submap t.

5.2.4 Robust outlier rejection

To support robust outlier rejection we utilize RANSAC based on a minimal set of three points (with Horn's algorithm [44]). This determines the inliers in the putative correspondence set and an initial approximation of the transformation. A second pass is then used with a limited search range based upon the estimate from the first pass, and typically produces more proposals and correct matches. The RANSAC loop is modified to include prior knowledge regarding the transformation scale between submaps. As the scale of the submaps is derived from the same instruments, registered submaps should have a similarity transformation with a scale close to unity. This helps speed up the RANSAC loop by allowing us to only evaluate the support of transformations with scale c such that $0.9 \leq c \leq 1.1$. If the scale is out of this range the set of potential correspondences is assumed to have at least one outlier.

5.2.5 Uncertainty in transformation

For simplicity we ignore the estimated covariance of 3D points (from the submap bundle adjustment) in the RANSAC loop. In this case the solution from Horn's algorithm is equivalent to an unweighted least squares. We then use this solution as an initial guess for a refinement based on the uncertainties of all corresponding structure points, which corresponds to minimizing the sum of Mahalanobis distances

$$\mathbf{d}_k = {}^{t}\mathbf{X}_k - {}^{t}_{s}\hat{\mathrm{T}} \cdot {}^{s}\mathbf{X}_k \tag{5.7}$$

$${}^{t}_{s}\mathrm{T}^{*} = arg \min_{{}^{t}_{s}\mathrm{T}} \sum_{k} \mathbf{d}_k{}^{\top} \Sigma_k^{-1} \mathbf{d}_k \tag{5.8}$$

with the covariance of the error approximated by the first order propagation of the covariance of the points being registered

$$\Sigma_k \approx \frac{\partial \mathbf{d}_k}{\partial {}^{t}\mathbf{X}_k} \Sigma_{{}^{t}\mathbf{X}_k} \frac{\partial \mathbf{d}_k}{\partial {}^{t}\mathbf{X}_k}^{\top} + \frac{\partial \mathbf{d}_k}{\partial {}^{s}\mathbf{X}_k} \Sigma_{{}^{s}\mathbf{X}_k} \frac{\partial \mathbf{d}_k}{\partial {}^{s}\mathbf{X}_k}^{\top} . \tag{5.9}$$

We assume that the estimates of structure points between submaps are uncorrelated, which is a valid assumption for submaps that do not share any cameras (*e.g.* across-track submaps).

The covariance of the transformation parameters can be estimated to first order from the Jacobian of the cost function being minimized in (5.8) evaluated at the optimum.

5.3 Edge proposal

Starting from a temporal sequence we wish to establish additional links between overlapping submaps (which will lead to establishing additional links between overlapping imagery). This can be viewed as a refinement of a graph where each node is a submap reference frame and each edge (or link) a relative transformation. Since submaps can be linked only to spatially neighboring submaps, the graph is expected to be sparse. This would require verifying only $\mathcal{O}(N)$ links if the node positions were well known. Yet as links are added we expect the spatial relationships between nodes to change, possibly requiring additional link checks. Verifying edges (map matching) is expensive computationally, so our approach must concentrate effort on likely links by considering uncertainty in the node positions and by updating node position estimates as links are added.

Possible approaches to estimating links (*i.e.* transformations between nodes) include

- Estimating relative transformations from current global estimates $\hat{\mathbf{x}}_{ik} = \ominus\hat{\mathbf{x}}_{wi} \oplus \hat{\mathbf{x}}_{wk}$.

- Estimating relative transformations from composition of relative transformations $\hat{\mathbf{x}}_{ik} = \mathbf{x}_{ij} \oplus \mathbf{x}_{jk}$.

These are related to the approach used to estimate the current network topology. If estimates of the node poses are maintained in a common, global reference frame then additional links can be inferred by measuring distances and uncertainties between nodes. Though the proposal process is simple and is less demanding as more of the topology becomes known (fewer possible links to consider), maintaining nodes in a common frame requires enforcing consistency among the cycles that may form as additional edges are included in the graph. It should be noted that while consistency is important before attempting a bundle adjustment (§5.4) it is not essential when attempting to establish edges in a sparse graph.

The alternative approach is to remain in relative frame space and use composition of relative transformations to express the relative transformation between nodes that do not have a direct link. Because there may be multiple paths between nodes, an approximate solution is to use a shortest path algorithm such as Dijkstra's. The Atlas framework advocates this approach for map matching [14]. In this case a consistent representation is not constructed but the proposal process is more complex since it must place nodes relative to a base node by composition along the shortest path. As more edges become available more paths must implicitly be considered.

5.3.1 Estimation of unmeasured links through shortest path

An uncertainty measure can be used as the path length in Djikstra's algorithm to choose among all paths when generating an estimate of the transformations between submaps. Starting from a base node a shortest path spanning tree is grown incrementally by choosing the edge that connects the shortest path tree so far to a candidate node (*i.e.*, a node that has an edge to the current tree) such that the composed uncertainty to the new node is the smallest of all possible candidate nodes.

To estimate relative transformations we compose multiple measured transformations and propagate their uncertainties. This is performed recursively along the possible paths:

$$ {}^{k}_{i}\hat{\mathrm{T}} = {}^{k}_{j}\mathrm{T} \cdot {}^{j}_{i}\hat{\mathrm{T}} \tag{5.10} $$

$$ \Sigma_{i,k} \approx \mathrm{J}_{i,j} \Sigma_{i,j} \mathrm{J}_{i,j}^{\mathsf{T}} + \mathrm{J}_{j,k} \Sigma_{j,k} \mathrm{J}_{j,k}^{\mathsf{T}} \tag{5.11} $$

where $\hat{\mathbf{x}}$ represents an estimate of $\mathbf{x}$.

The determinant of the covariance is an attractive uncertainty measure since it can be related to the volume of the uncertainty ellipsoid and is invariant under rigid body transformations [13]. We apply this approach to the six DOF problem of submap transformations. Since the determinant magnitude is small we prescale by a constant factor before calculating the determinants.

Dijkstra's algorithm is greedy but can be shown to provide the shortest path to all nodes when the path length is additive [108]. By using composition to the next node this assumption is violated and the greedy behavior may not yield the shortest path to all nodes. Assume two paths A and B offer different estimates to a node. If A composes to the smallest determinant, then Dijkstra's algorithm will choose that path. But this does not guarantee that nodes connected to this one will compose to a smaller determinant than B. Abusing notation, $det(A) < det(B) \nRightarrow det(A \oplus C) < det(B \oplus C)$.

5.3.2 Proposal strategies

After estimating relative transformations between a pair of submaps it is necessary to determine which submaps are likely to overlap. This depends on several factors such as camera field of view, altitude, terrain and camera trajectory in each submap. A simple approach is to calculate the distance and uncertainty between the centroids of the structure of each submap according to the relative transformation and its uncertainty.

A maximum distance for overlap can be estimated based on the camera field of view and the altitude of the cameras. We use as maximum distance the horizontal dimension of the footprint. That is, for overlap calculations we model the submap as a disc with diameter equal to the width of the footprint. This is a simple and conservative measure since submaps tend to be longer than their width. A more detailed model could keep track of the corners or even the convex hull of the submap footprint but this simple model performs satisfactorily.

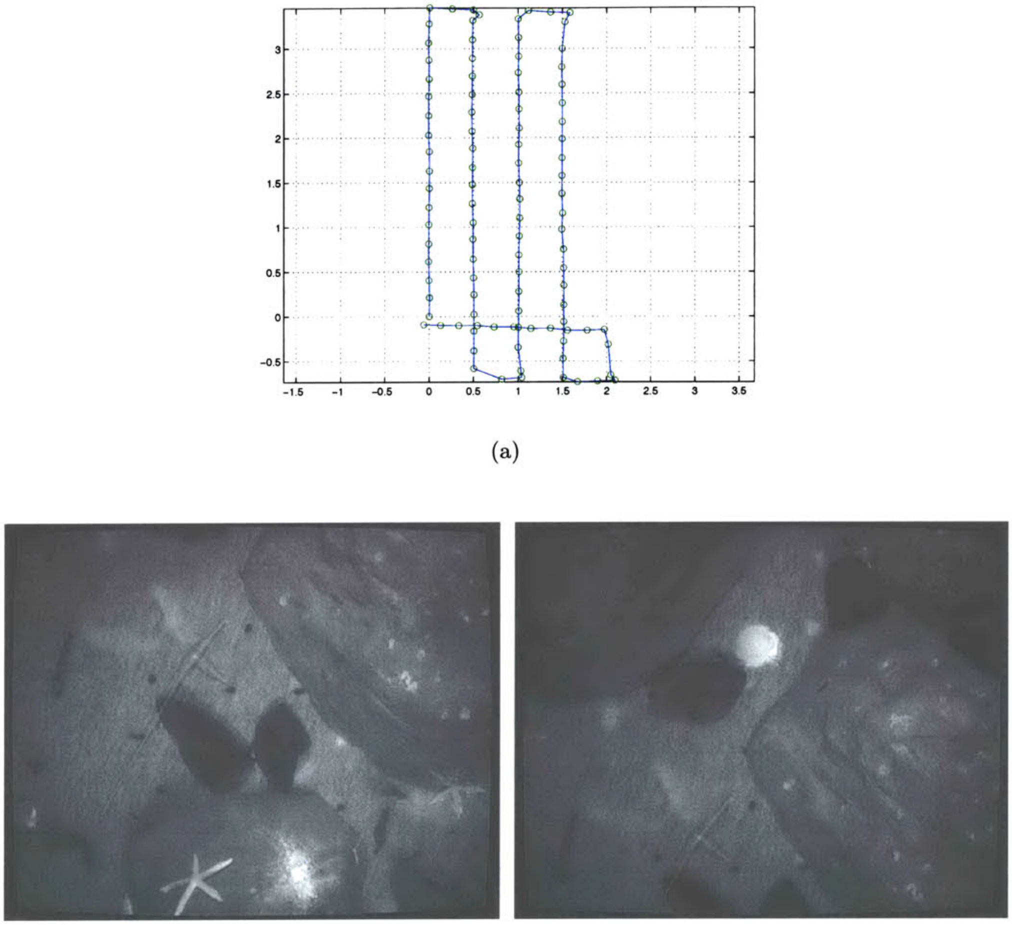
The proposal stage calculates a 99% confidence interval for the distance between submaps. If the maximum distance for overlap is within the interval (or greater) then overlap is considered possible. The most likely link is the one that has the highest proportion of the confidence interval within the maximum distance for overlap.

By proposing the most likely link within range the graph tends to 'zipper up' nodes, closing loops last. Alternatively, the least likely link within range of overlap could be verified first. Because it propose transformations with large uncertainty it relies heavily on being able to match maps without useful priors. For the same reason there is a lower probability that the nodes actually do overlap. This results in a low ratio of verified to proposed links.

The proposal and verification steps are repeated until the maximum number of allowable links is achieved, which is user-defined. A good choice is eight times the number of submaps which allows on average maps to connect to the previous and next map in the temporal sequence and up to six other nearby maps.

5.4 Node Estimation: Global poses from relative poses

Our objective is to place nodes in a global frame such that they are consistent with all the relative measurements (frame transformations). This can be formulated directly as an



(a)

(b)

(c)

Figure 5-7: Four tracklines taken from the JHU tank data set used to illustrate the link proposal and verification stage.(a) EKF track of the vehicle with circles marking the vehicle location when acquiring images. Units are in meters. (b) and (c) Sample images of a bottom with both flat and 'rocky' sections.

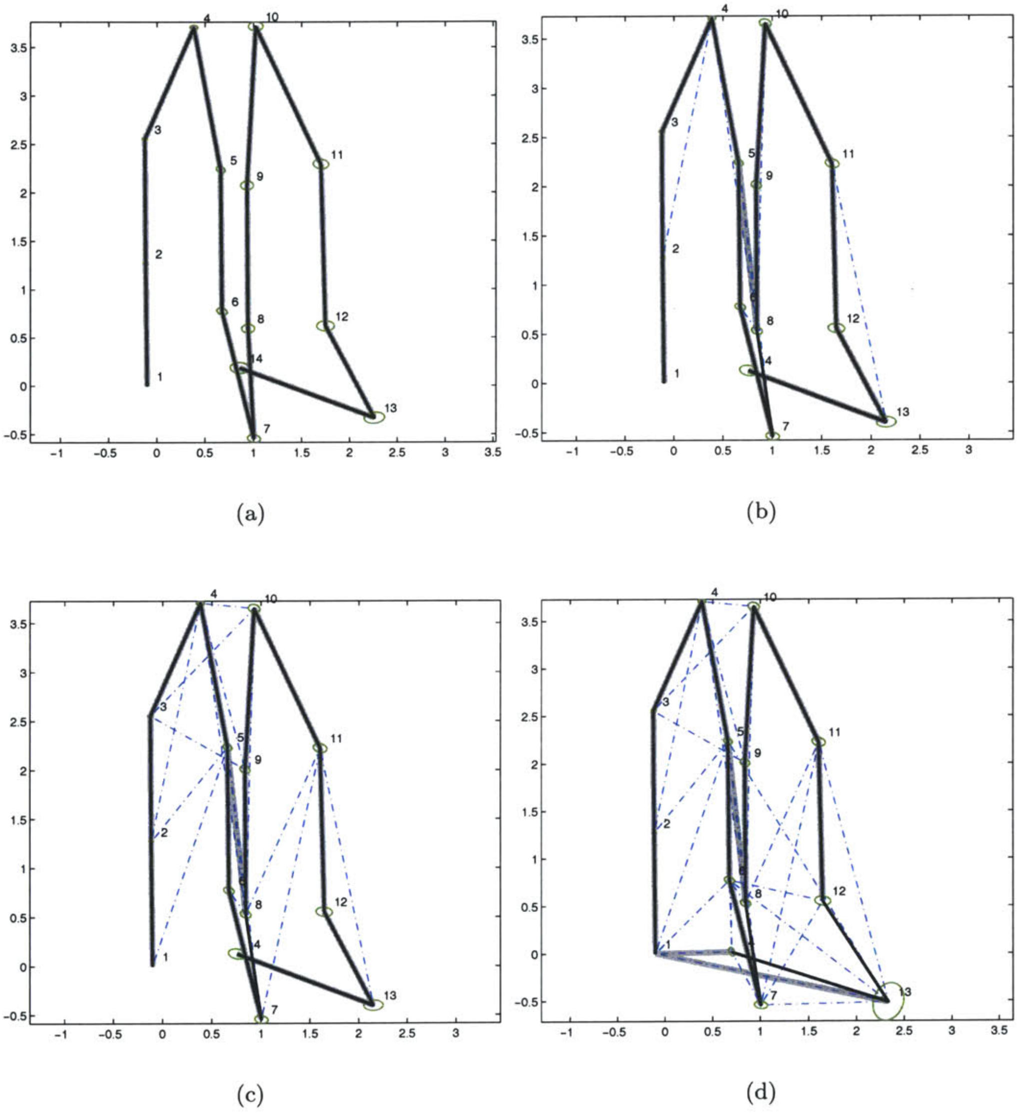
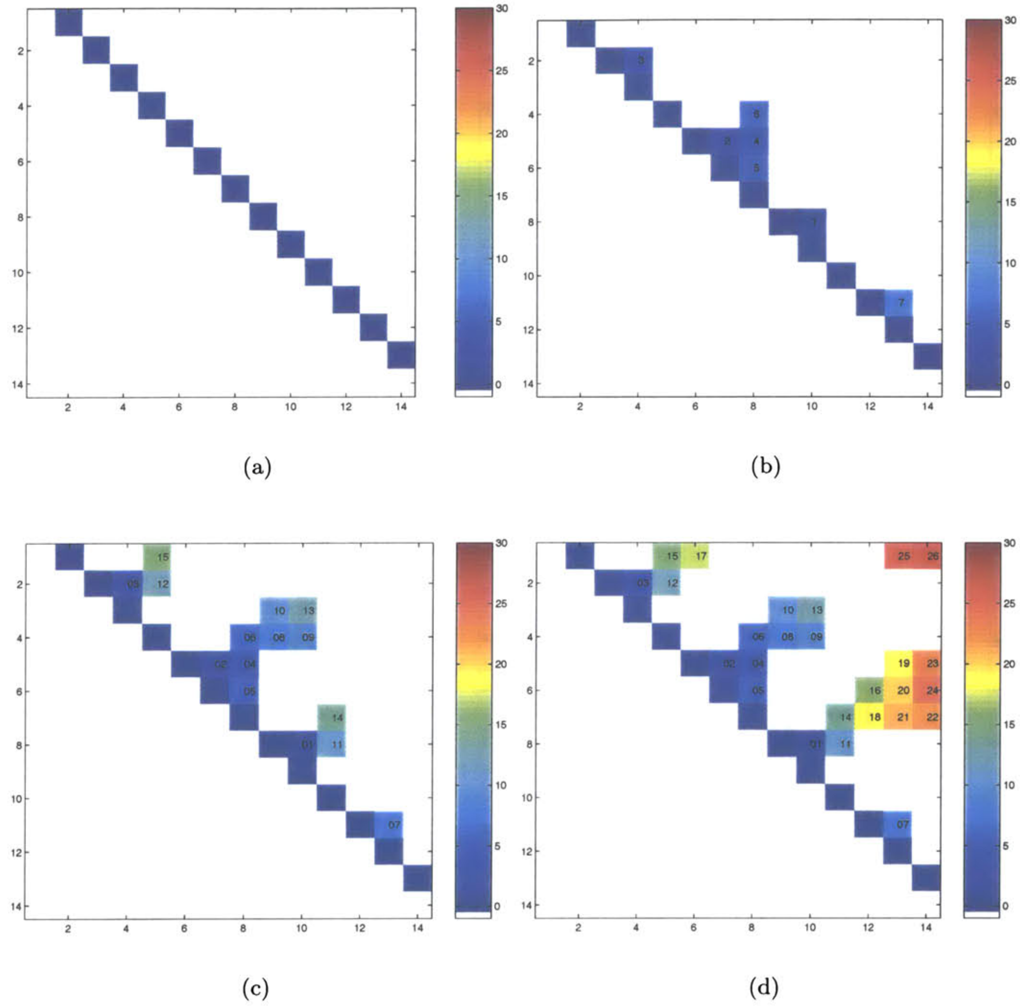


Figure 5-8: Start (a), intermediate (b), (c) and final (d) stages of the link proposal and verification for four tracklines of the JHU data set (Figure 5-7). The temporal sequence was processed into 14 submaps (labeled at the origin of each submap). The layout of the nodes (submaps reference frames) by composing transformations according to the shortest path algorithm. Black is the temporal sequence, gray the shortest path and dashed additional links (not used in the shortest path). The 99% confidence ellipses for the node xy positions are shown in green. Units in meters.



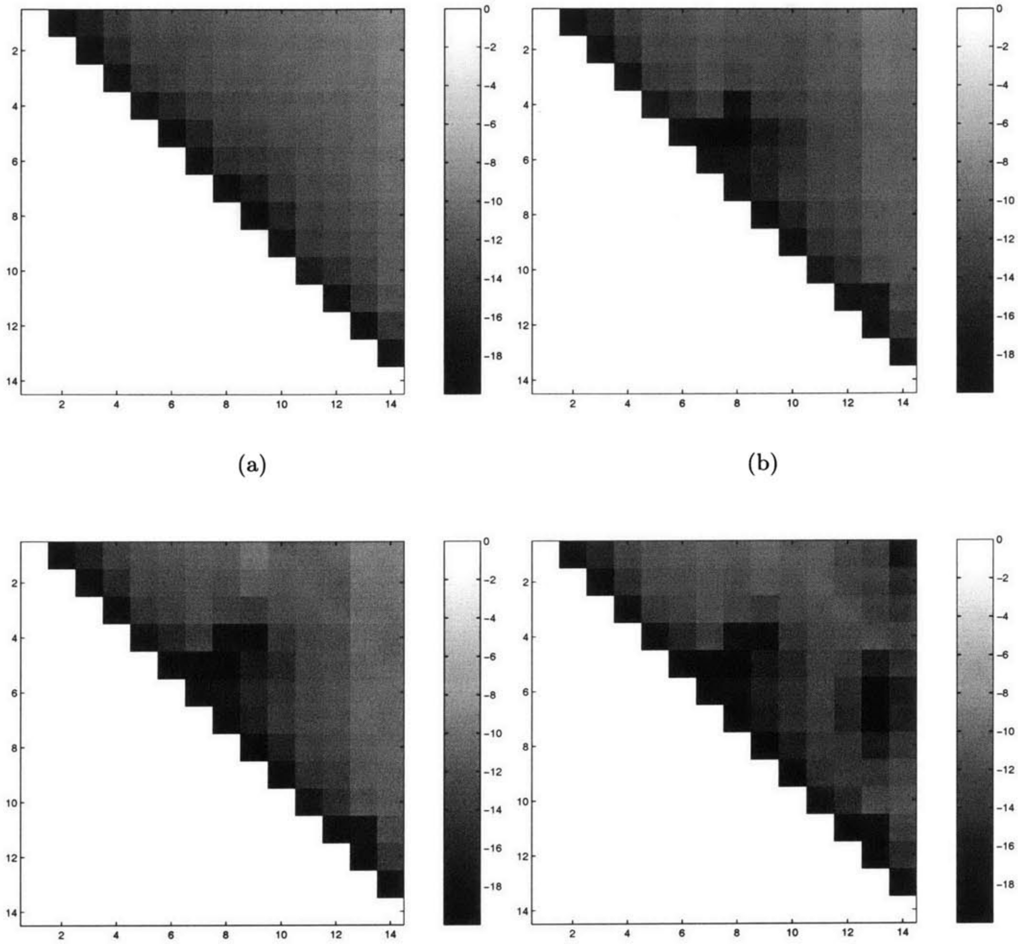
(a)

(b)

(c)

(d)

Figure 5-9: History of verified links color-coded and numbered according to the order of addition. (a) Start with only links from the temporal sequence. (b) and (c) are intermediate steps and (d) is the final adjacency matrix. Links are 'grown' by connecting closely related submaps first. The stages correspond to those in Figure 5-8.



(a)

(b)

(c)

(d)

Figure 5-10: (a) Start, (b),(c) intermediate and (d) final relative transformation uncertainty shown as the log of the determinant of the uncertainty as calculated by composition along shortest path for the stages shown in figures 5-8 and 5-9.

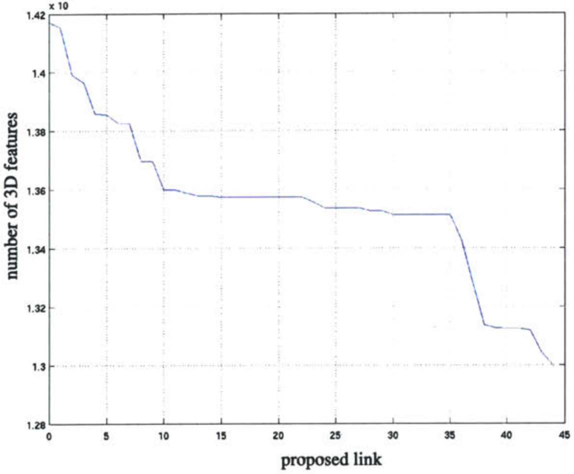


Figure 5-11: Evolution of the number of 3D features that are considered unique. As maps are matched 3D features that correspond across submaps are fused into one unique 3D feature.

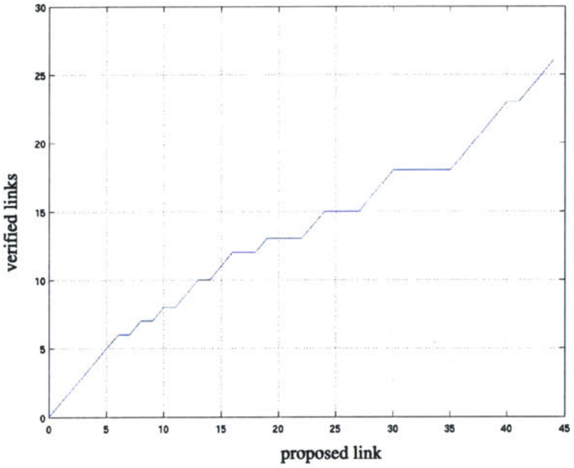


Figure 5-12: The evolution of verified links plotted against proposed links.

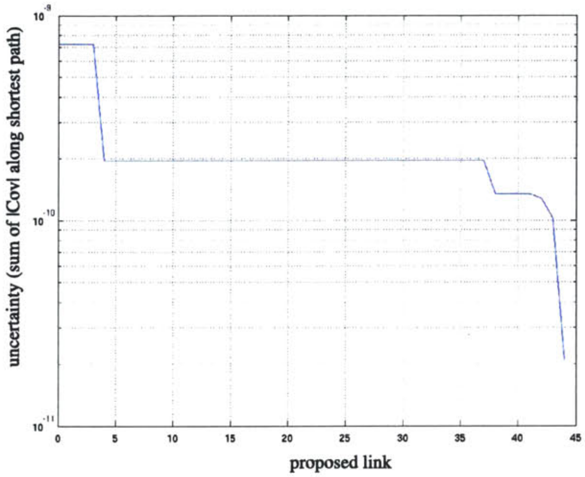


Figure 5-13: The evolution of the uncertainty in the relative transformations. The sum of determinants of the covariances plotted against link proposal. Estimated based on the shortest path compositions.

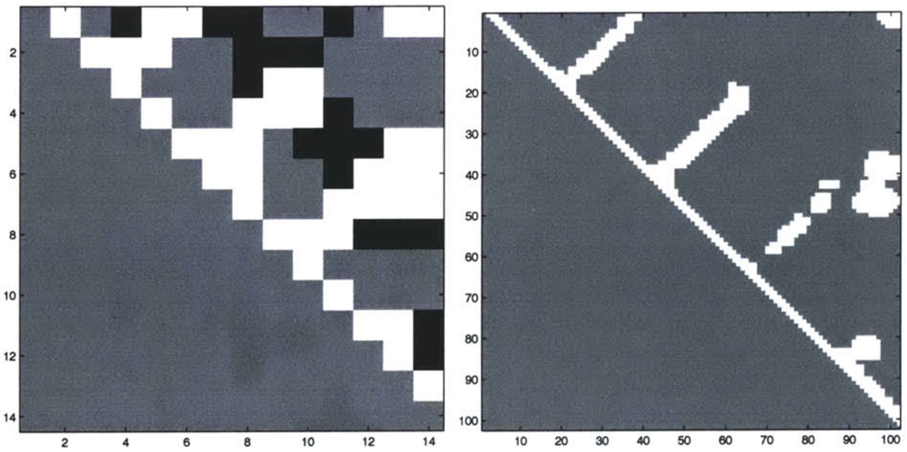


Figure 5-14: The adjacency matrix for (a) submaps and (b) images. Each 3D feature that is matched between submaps links all images that view that 3D feature. Verified links appear as white, proposed but not matched links in (a) are shown in black.

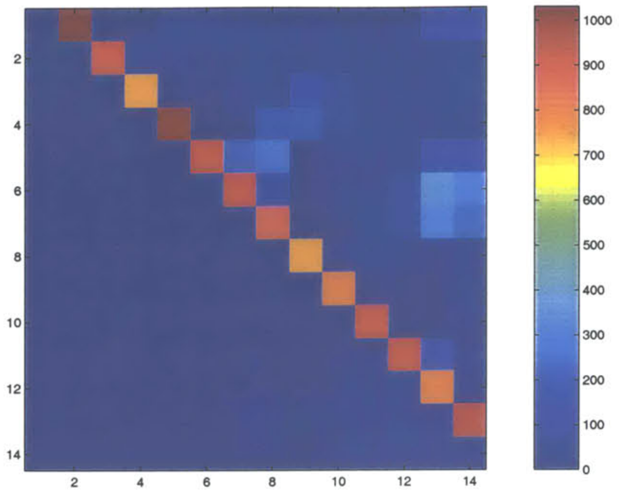


Figure 5-15: The number of matching features between submaps.

optimization problem to yield either a batch nonlinear least squares or a recursive nonlinear least squares solution [67] [19]. These approaches suffer from requiring to maintain the cross-covariances between submap frames. Sharp et al [109] have proposed a cycle consistency approach that operates in relative space but produces consistent global estimates without having to estimate or store cross-covariances. The graph can be seen as a distributed network and consistent, conservative global estimates can be generated through fusion by covariance intersection [104].

5.4.1 Nonlinear weighted least squares

We seek to determine the global poses that best explain all the relative pose measurements and respect the a priori distribution coming from navigation. By defining a cost function associated with these discrepancies we can then optimize an initial guess.

We define e_{ij} as the disparity pose vector between the composition of the estimates of global transformations ${}^{w}_{i}\hat{\mathrm{T}}$, ${}^{w}_{j}\hat{\mathrm{T}}$ and the measured relative transformation ${}^{j}_{i}\mathrm{T}$. Throughout this discussion we use $\hat{\mathbf{x}}$ to represent an estimate of $\mathbf{x}$. In Smith, Self & Cheeseman's [114] (SSC) notation, the relative pose vector implied by the estimates of pose is obtained from a tail-to-tail operation:

$$\hat{\mathbf{x}}_{ij} = \ominus\hat{\mathbf{x}}_{wi} \oplus \hat{\mathbf{x}}_{wj} \tag{5.12}$$

where the transformation/pose parameters are related to the homogeneous transformation as $\mathbf{x}_{ik} = \rho({}^{i}_{k}\mathrm{T})$. The disparity between the relative pose measurement $\mathbf{x}_{ij}$ (the MAP estimate from imagery and navigation) and the relative pose $\hat{\mathbf{x}}_{ij}$ from the tail-to-tail composition of estimates $\hat{\mathbf{x}}_j$ and $\hat{\mathbf{x}}_i$ is the error measure we seek to minimize

$$\mathbf{e}_{ij} = \ominus\hat{\mathbf{x}}_{ij} \oplus \mathbf{x}_{ij} = \ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i \oplus \mathbf{x}_{ij} \tag{5.13}$$

$\mathbf{e}_{ij}$ can be thought of as the residual transformation in a short cycle formed by the tail-to-tail estimate of the transformation $\ominus\hat{\mathbf{x}}_j \oplus \hat{\mathbf{x}}_i$ and the measured transformation by map matching (or from the temporal sequence) $\mathbf{x}_{ij}$. Ideally the residual transformation should be the identity (corresponding to no rotation and no translation). We use the rotation vector representation (where the direction of the vector specifies the axis of rotation and the magnitude of the vector corresponds to the angle of rotation) for the orientation parameters of the residual transformation [87].

$$\mathbf{e}_{ij} = \rho\left({}^{j}_{i}\mathrm{T}^{-1} \cdot {}^{j}_{w}\hat{\mathrm{T}} \cdot {}^{w}_{i}\hat{\mathrm{T}}\right) \tag{5.14}$$

We also define the disparity between the global pose according to navigation and our estimate of global pose

$$\mathbf{e}_{i} = \rho\left({}^{w}_{i}\mathrm{T}^{-1}{}^{w}_{i}\hat{\mathrm{T}}\right) \tag{5.15}$$

or directly in SSC notation:

$$\mathbf{e}_{i} = \ominus\hat{\mathbf{x}}_{wi} \oplus \mathbf{x}_{wi} \tag{5.16}$$

In a similar fashion to [67] we seek a set of global transformations $\mathcal{T}^*$ of all N submaps $\mathcal{T} = \{\mathbf{x}_{w1} \cdots \mathbf{x}_{wN}\}$ that minimizes this error over all links. We formulate this as a weighted non-linear least squares optimization:

$$\mathcal{T}^* = arg\min_{\mathcal{T}} \sum_{ij} e_{ij}^\top \Sigma_{ij}^{-1} e_{ij} + \sum_i e_i^\top \Sigma_i^{-1} e_i \tag{5.17}$$

where Σ_{ij} corresponds to the estimated covariance of e_{ij} propagated from the covariance of $\mathbf{x}_{ji}$ and Σ_i corresponds to the estimated covariance of e_i propagated from the covariance of $\mathbf{x}_{wi}$.

An alternative to minimizing the discrepancy between the composition of global poses is to directly minimize the 3D distances between corresponding points of submaps, though computationally more intensive because the number of equations is proportional to the number of corresponding points instead of to the number of measured edges. However, this reduces the sensitivity to poorly triangulated networks [1] where the error in the frame transformations might appear small at the expense of large errors in the structure. The error measure becomes

$$\mathbf{d}_{ijk} = {}^{w}_{i}\hat{\mathrm{T}} \cdot {}^{i}\mathbf{X}_k - {}^{w}_{j}\hat{\mathrm{T}} \cdot {}^{j}\mathbf{X}_k \tag{5.18}$$

$$\mathcal{T}^* = arg\min_{\mathcal{T}} \sum_{ij} \sum_{k} \mathbf{d}_{ijk}{}^\top \Sigma_{ijk}^{-1} \mathbf{d}_{ijk} + \sum_i e_i^\top \Sigma_i^{-1} e_i. \tag{5.19}$$

In cases where the frame-based refinement is unsatisfactory (*i.e.*, the reprojection errors for the implied camera poses are large or have strong biases) we switch to this cost function.

5.4.2 Fusion through covariance intersection

The problem of generating global pose estimates from multiple relative transformation measurements can be posed as a data fusion problem where the estimates of the node poses are refined by fusing the current estimate with the composition of a node linked to it and the relative pose between them. Covariance intersection (CI) [54] is a conservative scheme that allows fusing two estimates when the cross covariance between them is unknown. This approximation is attractive because it breaks down the large nonlinear optimization problem into independent estimates for each node. Schlegel and Kämpke [104] apply CI to node estimation in a SLAM context. The basic update rule is

$$\mathbf{x}_{wj}^{t+1} = CI\left(\mathbf{x}_{wi}^{t} \oplus \mathbf{x}_{ij}, \mathbf{x}_{wj}^{t}\right) \tag{5.20}$$

where $\mathbf{x}_{wi}^{t} \oplus \mathbf{x}_{ij}$ expresses the estimate of $\mathbf{x}_{wj}$ according to $\mathbf{x}_{wi}$ and the edge connecting them.

In general, for an estimate $\mathbf{x}_a$ the information matrix $\mathrm{H}_{\mathbf{x}_a}$ is the inverse of the covariance, *i.e.* $\mathrm{H}_{\mathbf{x}_a} = \mathrm{P}_{\mathbf{x}_a}^{-1}$. CI weighs the estimates according to their information content in a convex combination such that the uncertainty ellipsoid of the fused estimate contains the ellipsoids from all possible cross-covariances.

$$\mathrm{H}_{\mathbf{x}_c} = \omega \mathrm{H}_{\mathbf{x}_a} + (1 - \omega)\mathrm{H}_{\mathbf{x}_b} \tag{5.21}$$

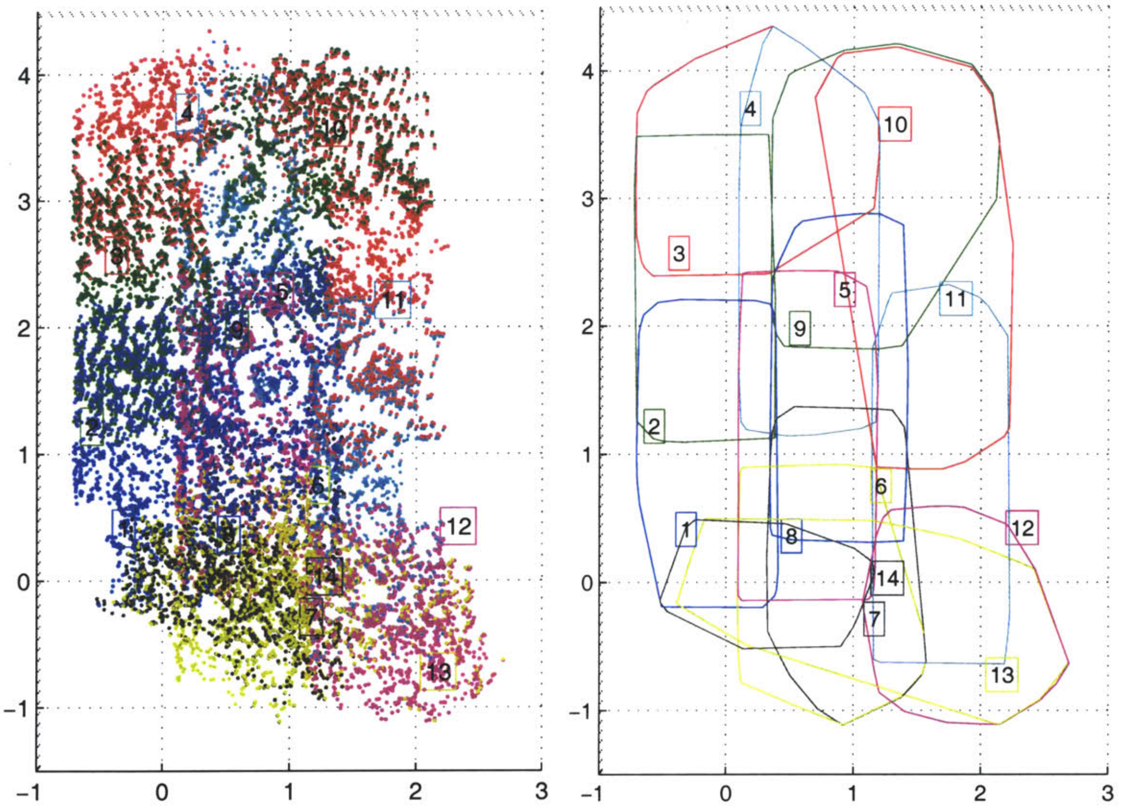
$$\mathbf{x}_c = \mathrm{H}_{\mathbf{x}_c}^{-1}(\omega \mathrm{H}_{\mathbf{x}_a}\mathbf{x}_a + (1 - \omega)\mathrm{H}_{\mathbf{x}_b}\mathbf{x}_b) \tag{5.22}$$

In a network setting the update equation is used over all links and over all nodes until the estimates and uncertainties converge. Unfortunately fusing estimates based on their uncertainty in such a conservative fashion can result in estimates that are not strongly constrained by measurements.

The shortest path algorithm allows for a fast and simple exploration of the topology of the network. As an initial guess to optimize global transformations, the shortest path estimate ignores cycles that in some cases might lead the global optimization to a local minimum. Preliminary tests using CI as a refinement of the shortest path initial guess suggest that the initial residuals tend to be smaller when using the CI refinement. This probably stems from CI using all link constraints.

5.4.3 Camera poses from submaps

Once submaps are placed in a global frame it is then possible to place the cameras that form the submaps in the same global frame. These cameras in the global frame are used as the initial guess for the bundle adjustment of the complete data set.



(a)

(b)

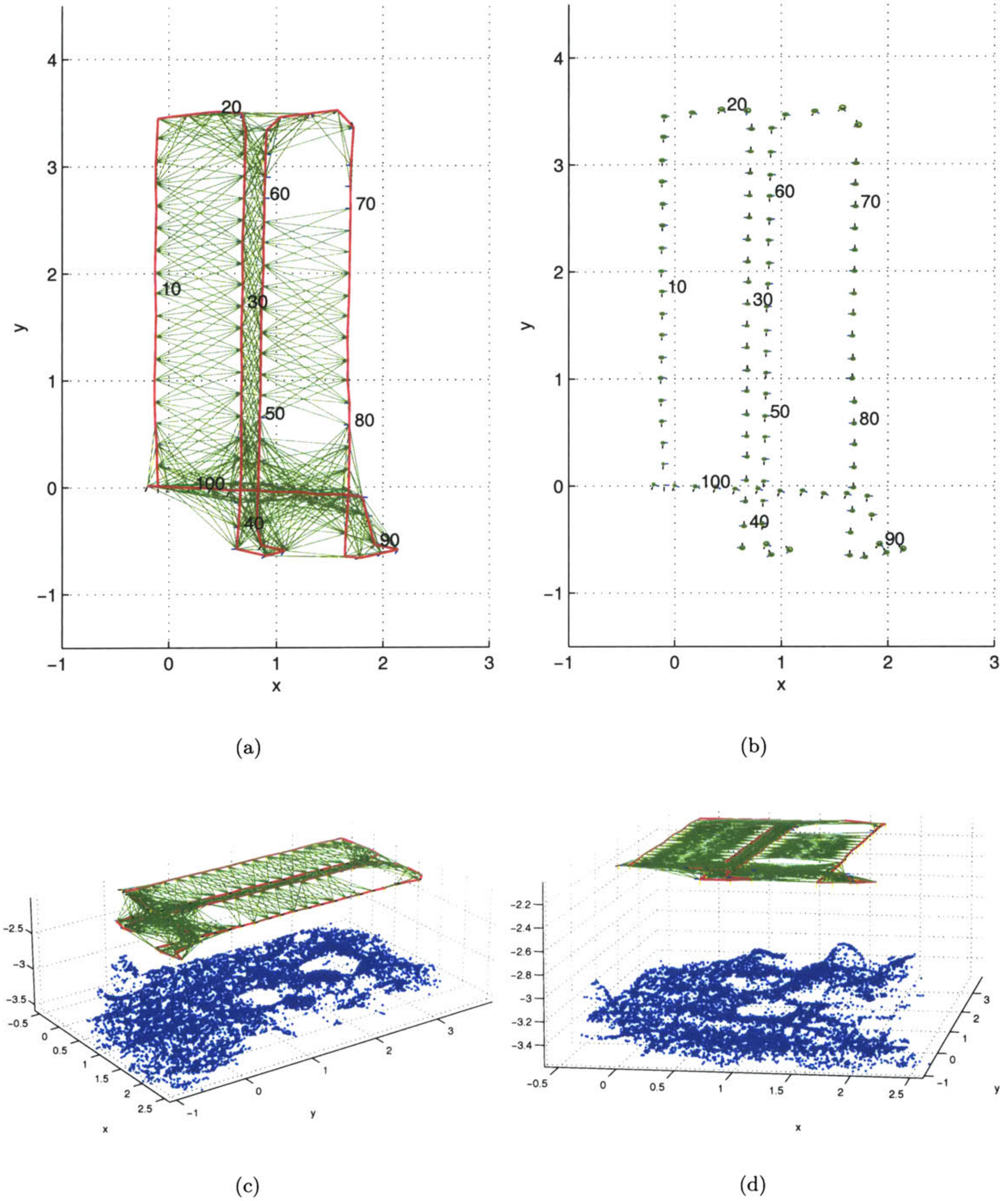
Figure 5-16: Plan view (xy) of the placement of submaps for the four JHU tracklines in a global frame. (a) features color-coded by submap. (b) The convex hull of the submaps shows high overlap in the temporal sequence and varying degrees of overlap across track.

By construction the pose of each camera in a submap is in the frame of the first camera. The transformation from the node to the global frame can be composed with the transformation of the camera pose to the node origin.

Since temporally adjacent submaps share cameras there is more than one way of mapping the cameras that are common between submaps. We use the geometric mean [86] of the pose estimates according to each submap (in the global frame) to obtain an initial guess of the camera positions.

5.5 Bundle Adjustment

Once camera poses are in the global frame the same sparse bundle adjustment routine used to close the submaps is used on the entire data set. We obtain the maximum a posteriori estimate by including cost terms associated with the navigation measurements, as described in §4.5.



(a)

(b)

(c)

(d)

Figure 5-17: Results from bundle adjustment of the four tracklines from the JHU tank. (a) Recovered cameras (xy plane). The trajectory is highlighted by red links while additional spatial links appear as green. Every tenth camera is identified with its place in the temporal sequence. Units in meters. (b) The 99% confidence ellipses for the xy position according to the bundle adjustment, assuming that the xy coordinates of the frame are fixed at the first camera. (c) and (d) two views of the reconstructed terrain with the camera poses and links.

Chapter 6

Validation and Results

6.1 Approach

The structure from motion problem is solved by gathering data with redundancy (multiple views), identifying the redundancy (corresponding features in multiple views) and enforcing consistency between model and data (by minimizing the reprojection error of estimated 3D points on estimated views). To a limited degree the quality of the reconstruction can be assessed by the behavior of the residuals [72] with qualitative assessments such as the distribution and magnitude of reprojection errors reflecting the presence of systematic effects. The presence of outliers can sometimes be noted in the observation residuals although this is not always the case since the measurement might not detect certain errors (for example, a two view mismatch will not have a large error if the incorrect match is along the correct epipolar line). The *precision* of the reconstruction is derived from the covariance of the recovered parameters (pose and structure). But these checks are based on self-consistency between model and measurements which, by definition, is what we try to optimize in the SFM solution. Thus they do not provide insight into the quality of the model relative to the scene.

The underlying goal is to produce an *accurate* reconstruction, so that the recovered parameters closely describe the state of the world. This cannot be determined by examining just the solution since a very consistent (precise) solution might still not be accurate (for example the scale might be off). For this we must use some form of ground truth. Underwater this can be particularly difficult since alternative measurement forms based on sound are limited in resolution or range. This chapter presents validation of the thesis framework through a small scale experiment with position and structure ground truth.

Results from a coral reef survey demonstrate the applicability to real world data and provide an opportunity to discuss some sensor bias and offset corrections based on self consistency.

6.1.1 Assumptions

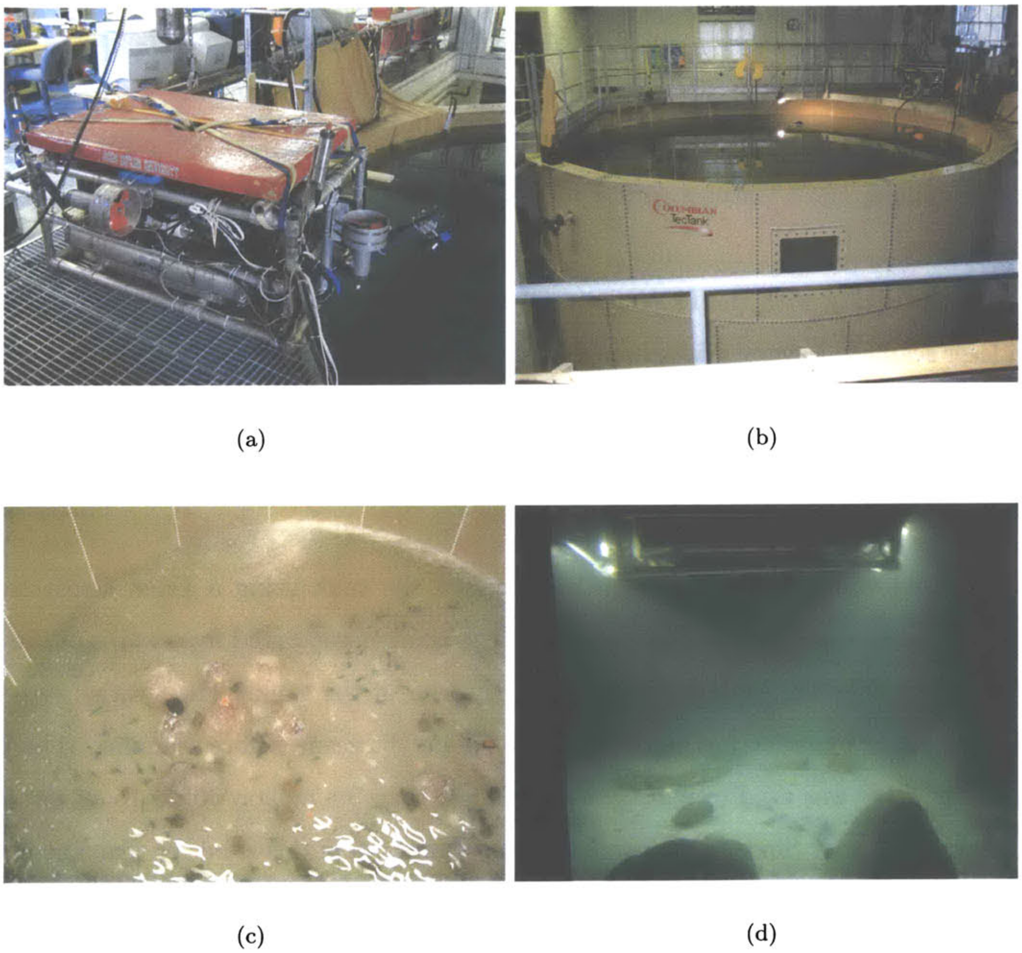
- Realistic sensor frame calibrations. Measurements of sensor to vehicle frame transformation usually have small errors that affect long term navigation estimates.
- Self consistency in the estimated pose and structure can be used to identify and correct for biases in sensor readings and errors in the transformation from sensor to vehicle frame.

6.2 Validation: Test Tank

6.2.1 Context

Limited access to the ocean floor makes it particularly difficult to directly validate results from a survey. To generate a data set with ground truth we used the SeaBED camera system on the Johns Hopkins University (JHU) Remotely Operated Vehicle (ROV) at the JHU test tank (Figure 6-1). The JHU ROV carries a similar navigation suite to the SeaBED AUV. In addition the tank is instrumented with Sharps, a high frequency (300 kHz) acoustic long baseline positioning system which provides an independent, absolute position estimate with sub-centimeter accuracy based on triangulating travel time to transponders.

A single light sources attached to a vehicle can cast shadows that alter the appearance of the scene depending on the heading of the vehicle. This reduces similarity and impacts our ability to match images and submaps that are not adjacent in the temporal sequence. Significant improvements may be realized by using two light sources on the vehicle as shown in figure 6-1(d). For ROVs which have access to essentially unlimited power, such a lighting configuration could be operationally viable. For AUVs this might require a trade-off in range



(a)

(b)

(c)

(d)

Figure 6-1: (a) The JHU ROV on deck. Light booms are visible fore and aft. (b) The JHU test tank, with the ROV visible on the right. (c) A down-looking view into the tank as it was being filled. The carpet and rocks are visible.(d) The swimming ROV as seen through a view port. The dual light configuration reduced cast shadows.

and mission duration, although the advent of LED based lighting systems may enable such multiple light-source configurations.

Results from the SFM reconstruction are shown in Figure 6-2. The evolution of the submap links and the number of common features between submaps and the plane view of the submap layout are shown in Figure 6-3. The reprojection errors for all measurements and their distribution are presented in Figure 6-4. Some outliers are apparent in the reconstruction, though their effect is reduced by the Cauchy M-estimator. Figure 6-5 shows feature tracks of the third submap color-coded according to reprojection error magnitude. For that submap most features persist between two to five views in a submap that consists of 10 images.

After surveying with the ROV the tank was drained and the bottom scanned with an area laser scanner. Several million range and bearing measurements were registered to form a 3D point cloud of the tank bottom with millimeter accuracy.

6.2.2 Structure ground truth

A team from Cullinan Engineering Co. scanned the tank using a Leica Geosystems - HDS2500 (serial number P24) laser scanner. The scanner generated five swaths of the tank bottom from different vantage points along the rim of the tank. The swaths were then registered aided by reflecting markers positioned on the scene before scanning. The Leica Cyclone package was used to register a set of more than 3.8 million points to an estimated accuracy of 1.2 mm. Since ranges were on the order of 5-6 m this is better than the usually quoted ±4 mm accuracy at 50 m. The surface area was approximately 41m^2 resulting, on average, in 9 range measurements for each cm^2 of the bottom.

We initially aligned SFM reconstruction with the laser data by selecting easily recognizable landmarks (Figure 6-6) and then refined through ICP. We note that the carpet was slightly buoyant underwater and was kept on the bottom by multiple lead weights and that after the tank was drained the carpet settled under its own weight. We attempted two registration strategies to overcome the non-rigid transformation between surfaces: using only points belonging to rocks to register (segmenting by height under the assumption that the rocks in the scene did not move), and performing ICP based on the points with

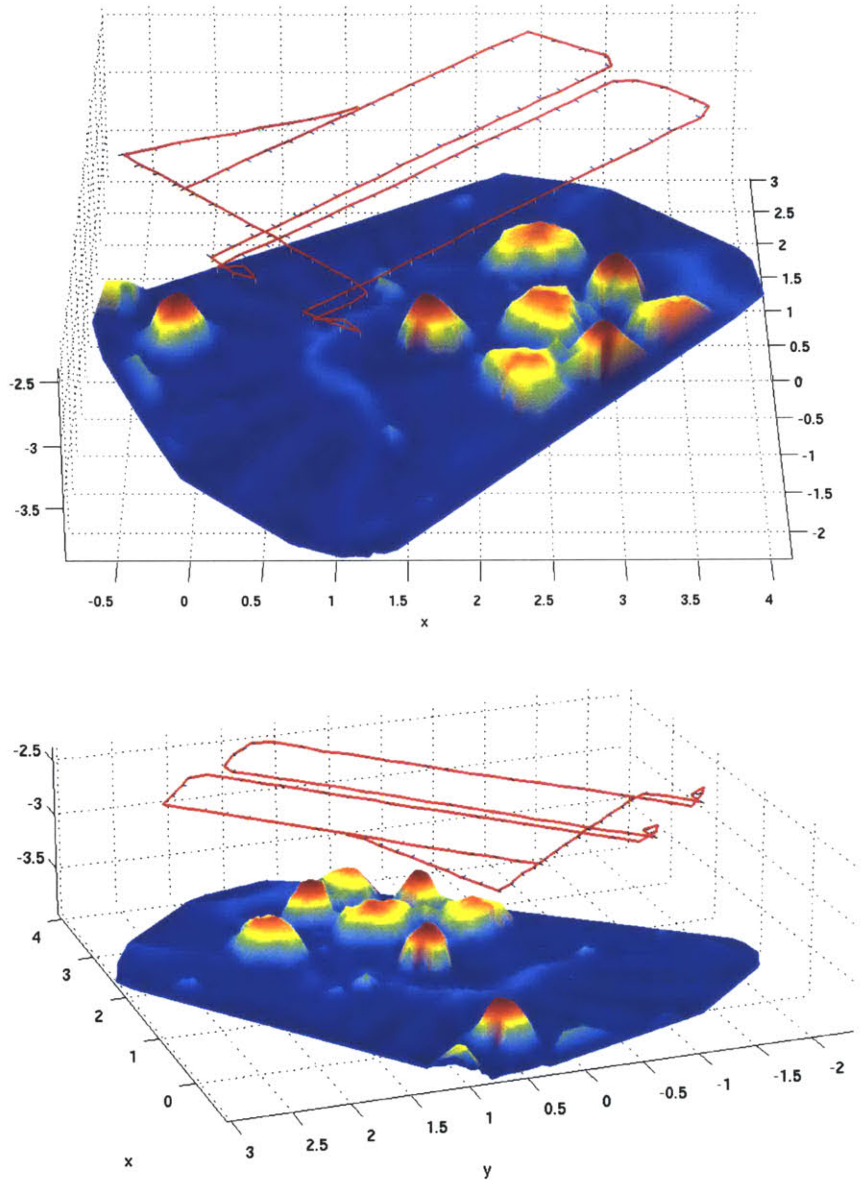
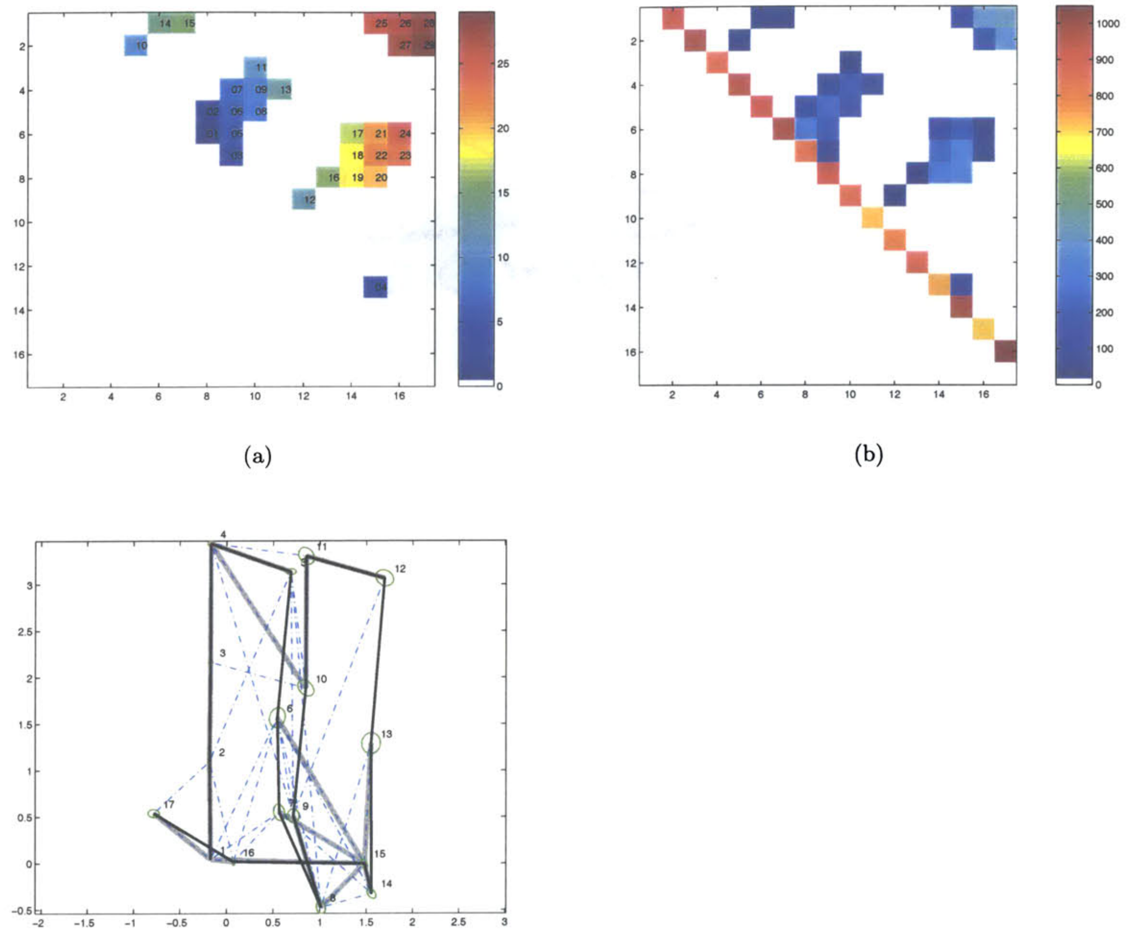


Figure 6-2: Two views of the reconstruction of poses and structure for the JHU tank. The camera poses are connected by a red line. A Delaunay triangulation interpolates a surface between 3D feature points. The structure is color-coded according to height. Units are in meters.



(a)

(b)

(c)

Figure 6-3: (a) Order in which links across track were added to the graph. The 'zipper' effect in parallel tracklines is apparent as links close in time are established before more distant ones. (b) The number of matching features between submaps. The closing of the loop can be seen in the relatively high number of common features between the first and last submaps. (c) The plan view of the submap origins according to the shortest path algorithm: the temporal sequence (fine black), the additional links (dot-dashed blue) and the shortest uncertainty path from the origin node (wide gray).

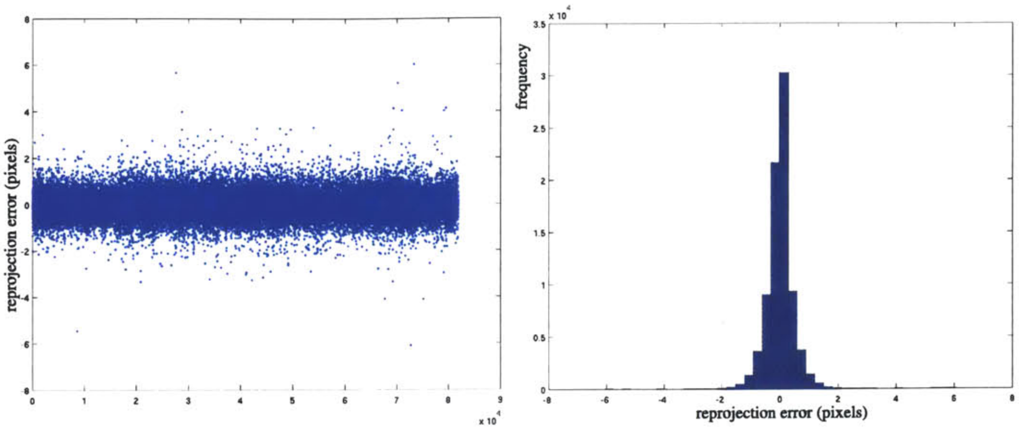


Figure 6-4: (Left) The reprojection errors (both x and y coordinates) for all reconstructed features. Some outliers are present though their effect is reduced by using an m estimator in the bundle adjustment.(Right) A histogram of the same errors. For visualization purposes 95% of the features with lowest associated reprojection errors are displayed in the reconstructions of Figure 6-2.

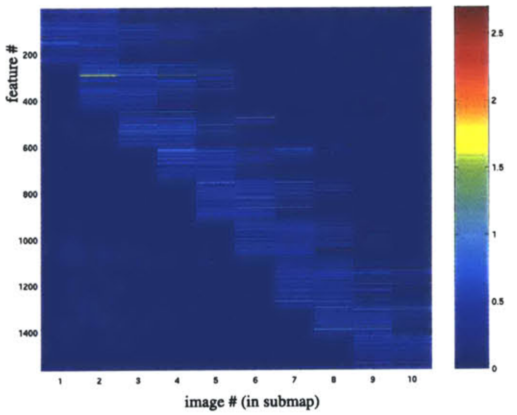


Figure 6-5: The feature tracks and the norm of the reprojection error for the third submap.

registration errors below the median error (under the assumption that at least half the points remained fixed). Results were very similar for both strategies and we present the median-based approach since it highlights regions where the carpet moved.

Figures 6-7 and 6-8 indicate that the registration errors are of the order of centimeters with a 2% change in scale. Though the tank is a relatively small scale reconstruction problem, these results suggest that the approach is capable of delivering reasonable estimates of scene structure.

By using points below the median error to calculate the similarity transformation to register the SFM and laser data we effectively segment the data into two halves, one of which was allowed to deform while the other was not. It is interesting to note from Figure 6-9 that most of the outliers correspond to the broad carpet waves.

6.3 Results: Bermuda survey

6.3.1 Context

In August 2002 the SeaBED vehicle performed several transects on the Bermuda shelf as well as some shallow water engineering trials. This section presents results from a shallow water (20 m approx) area survey programmed with several parallel tracklines for a total path length of approximately 200 m and intending to cover 200 m^2. Due to very strong swell and compass bias the actual path deviated significantly from the assumed path. This data set illustrates the capabilities to infer links in the graph of submaps to yield a consistent reconstruction.

6.3.2 Single Loop

A section of 62 images where the camera trajectory approximately folds back on itself shows matching and reconstruction along the temporal sequence and across track. Figure 6-10 presents Delaunay triangulated surfaces trough the reconstructed points and the camera trajectory. Plan views of the camera trajectory, the links (common 3D features) between views and the uncertainty in the xy position of the cameras are shown in figure 6-11.

Figure 6-12 shows features points and the convex hull of the submaps. Spatial overlap

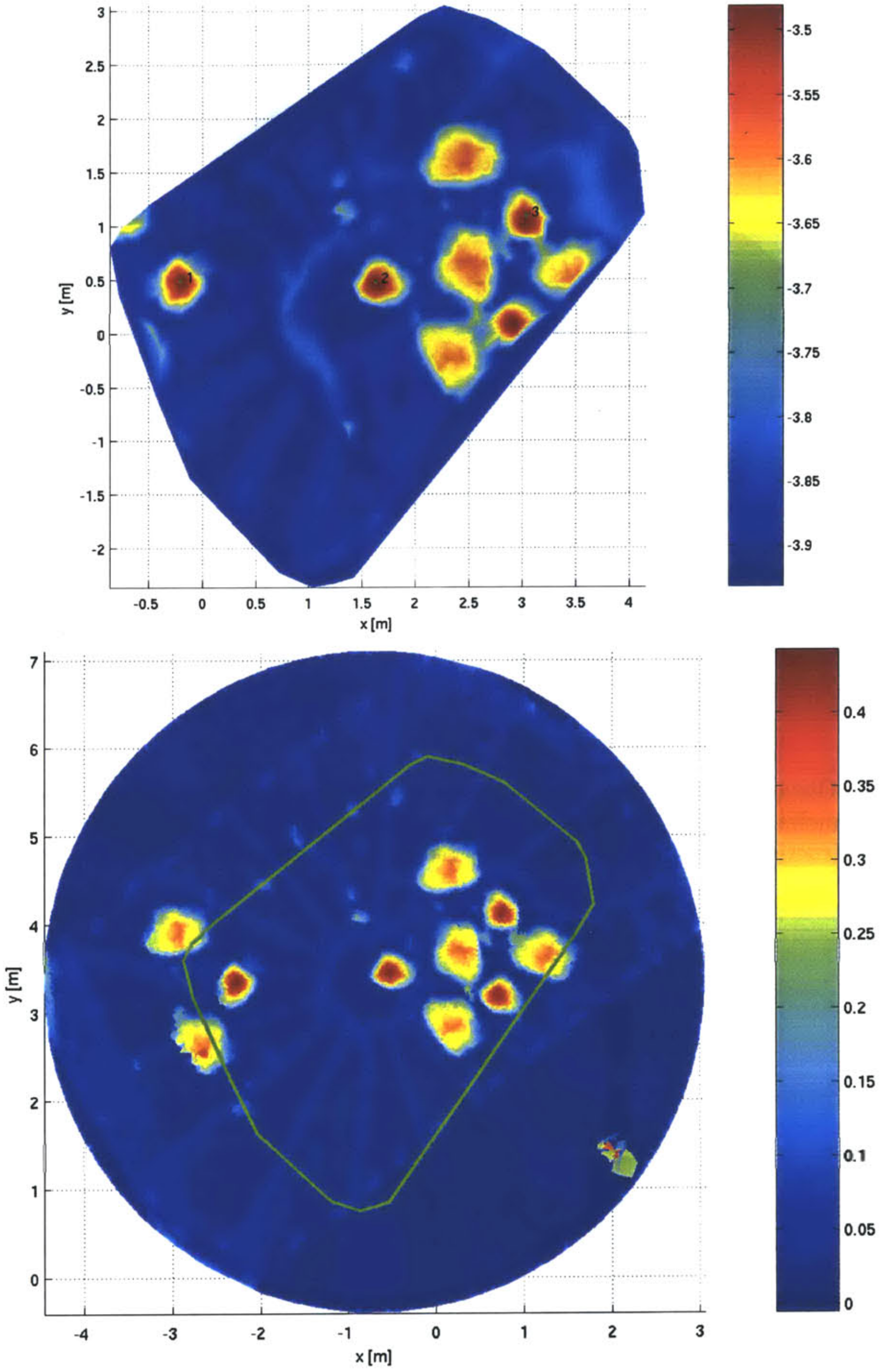


Figure 6-6: (Top) Height map from the SFM reconstruction. Surface based on a Delaunay triangulation. The labeled points were manually selected for the initial alignment with the laser scan. (Bottom) Height map from the laser scan. The outline of the manually registered SFM reconstruction is shown in green.

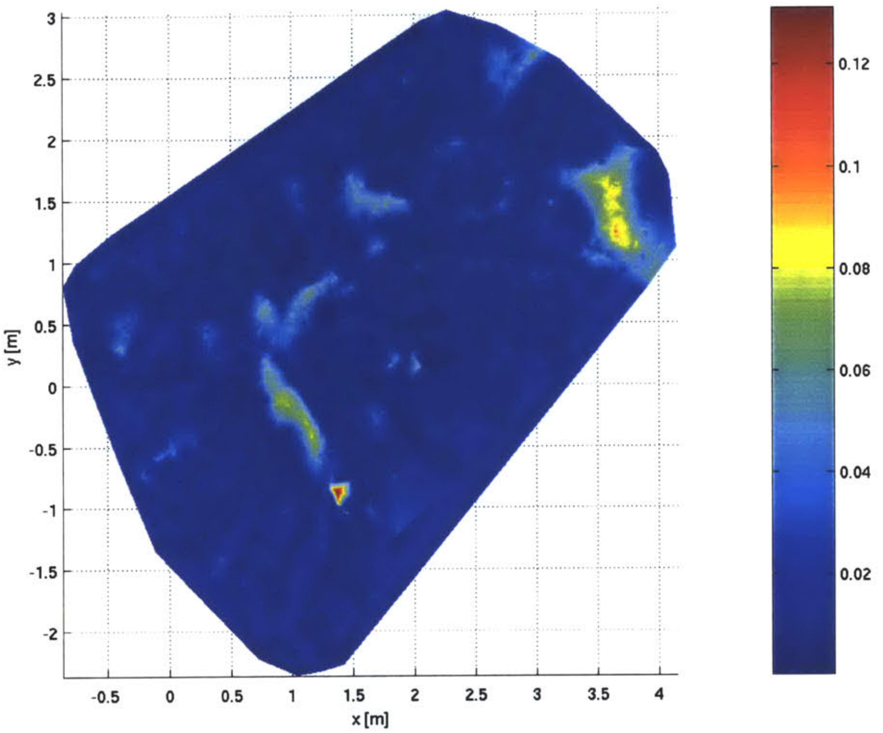


Figure 6-7: Distance map from SFM 3D points to the laser scan after ICP registration. Areas of large discrepancies tend to correspond to the carpet being buoyant for the visual survey. An outlier in the reconstruction produced the large error visible at approximate x=1.4 m,y=0.8 m.

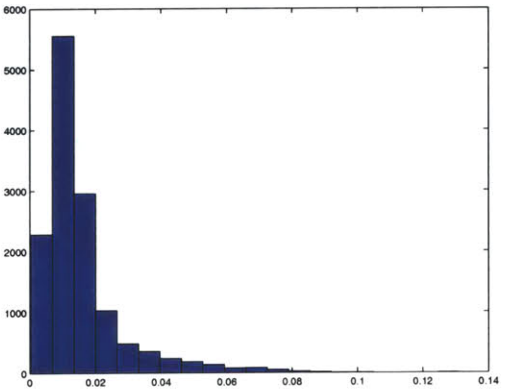


Figure 6-8: The distribution of minimum distances to the laser scan from each recovered 3D point. Because of the moving carpet only the points below the median error were used to calculate the registration transformation. The similarity based registration results in an RMS distance of 3.6 cm. Scale is recovered to within 2%.

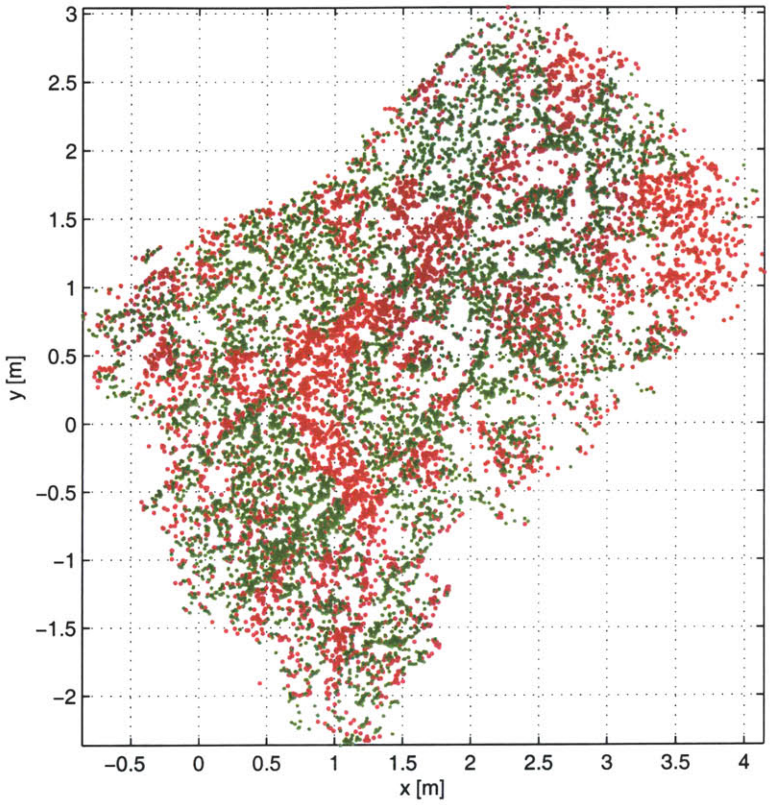


Figure 6-9: Points below the median error (green) and above (red). Registration parameters where calculated using points below the median error. By referring to Figure 6-6 outliers tend to group around the smooth, raised folds of the carpet which clearly do not correspond to the drained carpet surface.

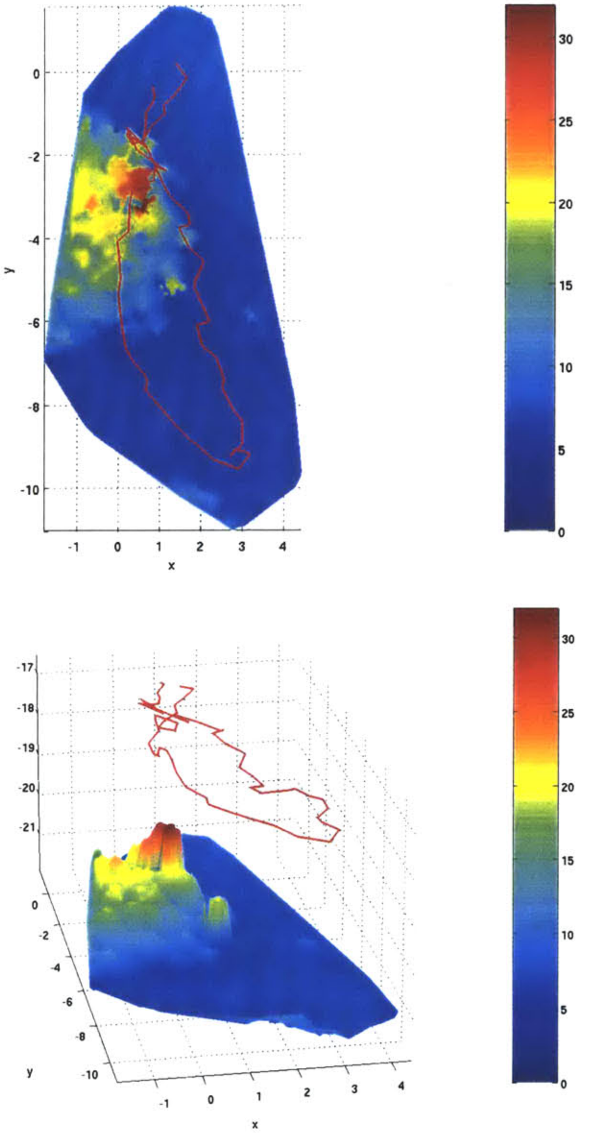


Figure 6-10: Two views of the reconstruction as a surface through the recovered 3D points. The camera trajectory is also presented as a red line. Strong swell significantly perturbed the vehicle trajectory yet the consistency of the reconstruction is apparent in the persistent features such as the sand ripples on the bottom.

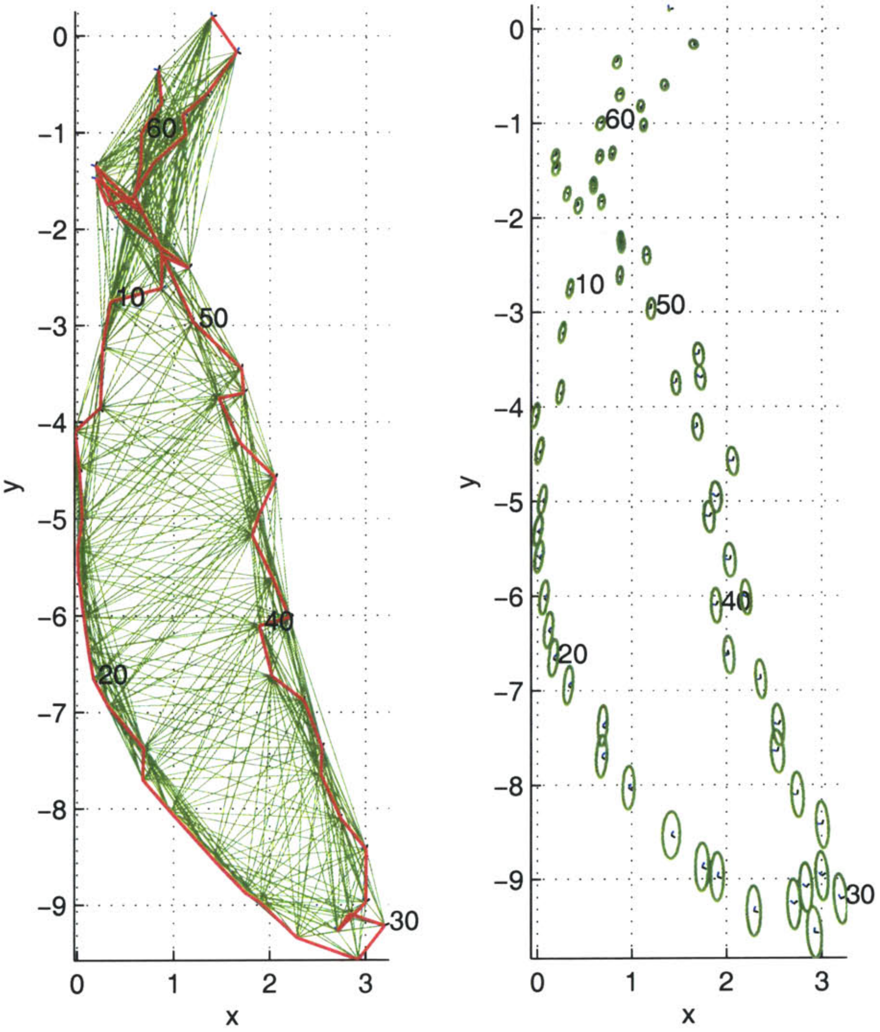


Figure 6-11: (Left) Plan view of the camera trajectory (red) and common features between cameras (green links). (Right) The 99% confidence ellipses for the xy position of the cameras. Every tenth camera is numbered on both figures to suggest the temporal sequence.

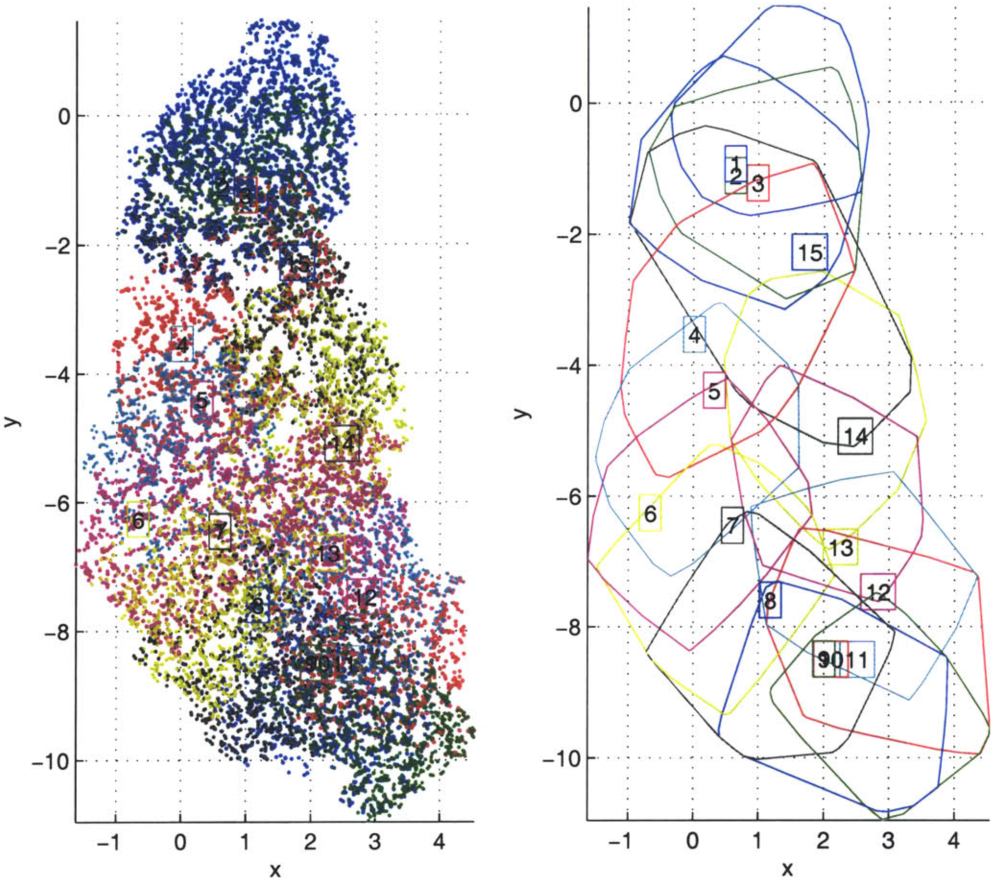


Figure 6-12: (Left) Plan view of the features for each submap. (Right) Convex hull of the 3D features of each submap. The varying degrees of spatial overlap between submaps is apparent in these figures.

between temporally adjacent submaps is consistent while across track overlap is a function of the trajectory followed by the vehicle.

6.3.3 Multiple Passes

A section of 169 images demonstrates matching and reconstruction along the temporal sequence and across track with multiple passes over the same area. Figure 6-16 presents Delaunay triangulated surfaces through the reconstructed points and the camera trajectory. Plan views of the camera trajectory, the links (common 3D features) between views and the uncertainty in the xy position of the cameras are shown in figure 6-17.

Figure 6-18 shows features points and the convex hull of the submaps. Spatial overlap

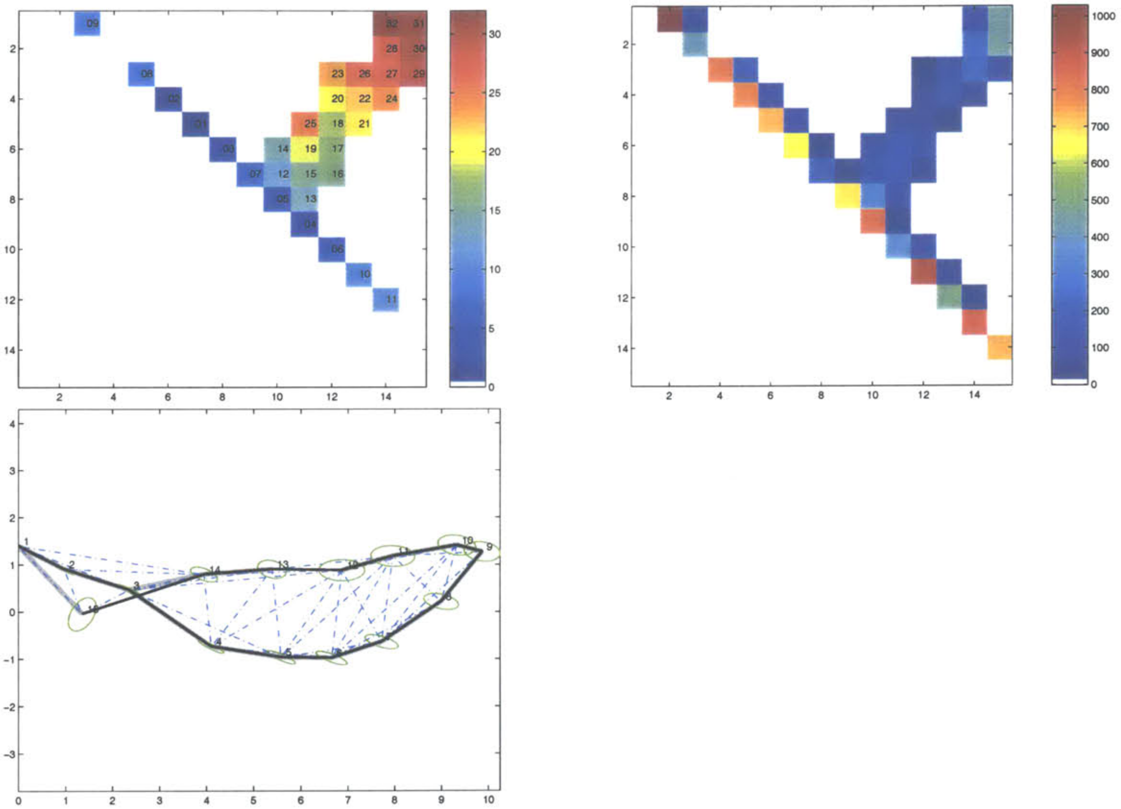


Figure 6-13: (a) Order in which links across track were added to the graph. The 'zipper' effect is apparent as links close in time tend to be established before more distant ones. (b) The number of matching features between submaps. The closing of the loop can be seen in the relatively high number of common features between the first and last submaps. (c) The plane view of the submap origins according to the shortest path algorithm: the temporal sequence (fine black), the additional links (dot-dashed blue) and the shortest uncertainty path from the origin node (wide gray).

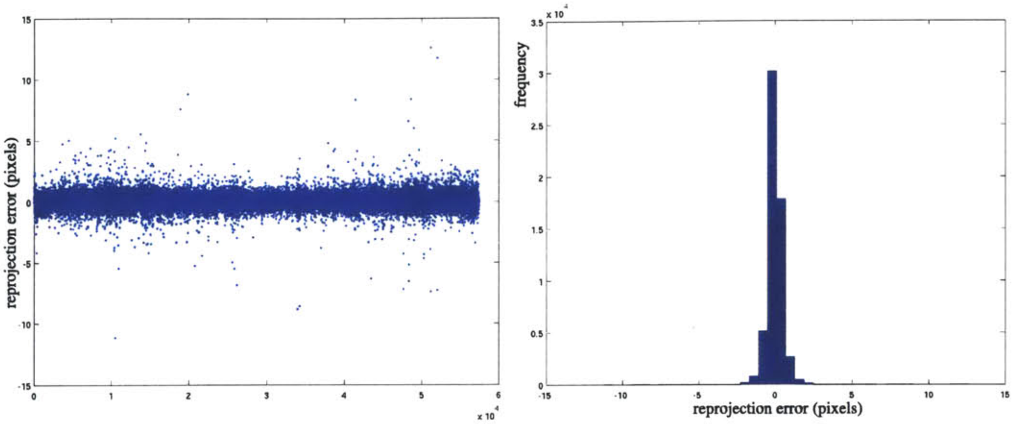


Figure 6-14: (Left) The reprojection errors (both x and y coordinates) for all reconstructed features. Some outliers are present though their effect is reduced by using an m estimator in the bundle adjustment.(Right) A histogram of the same errors. For visualization purposes 95% of the features with lowest associated reprojection errors are displayed in the reconstructions of Figure 6-10.

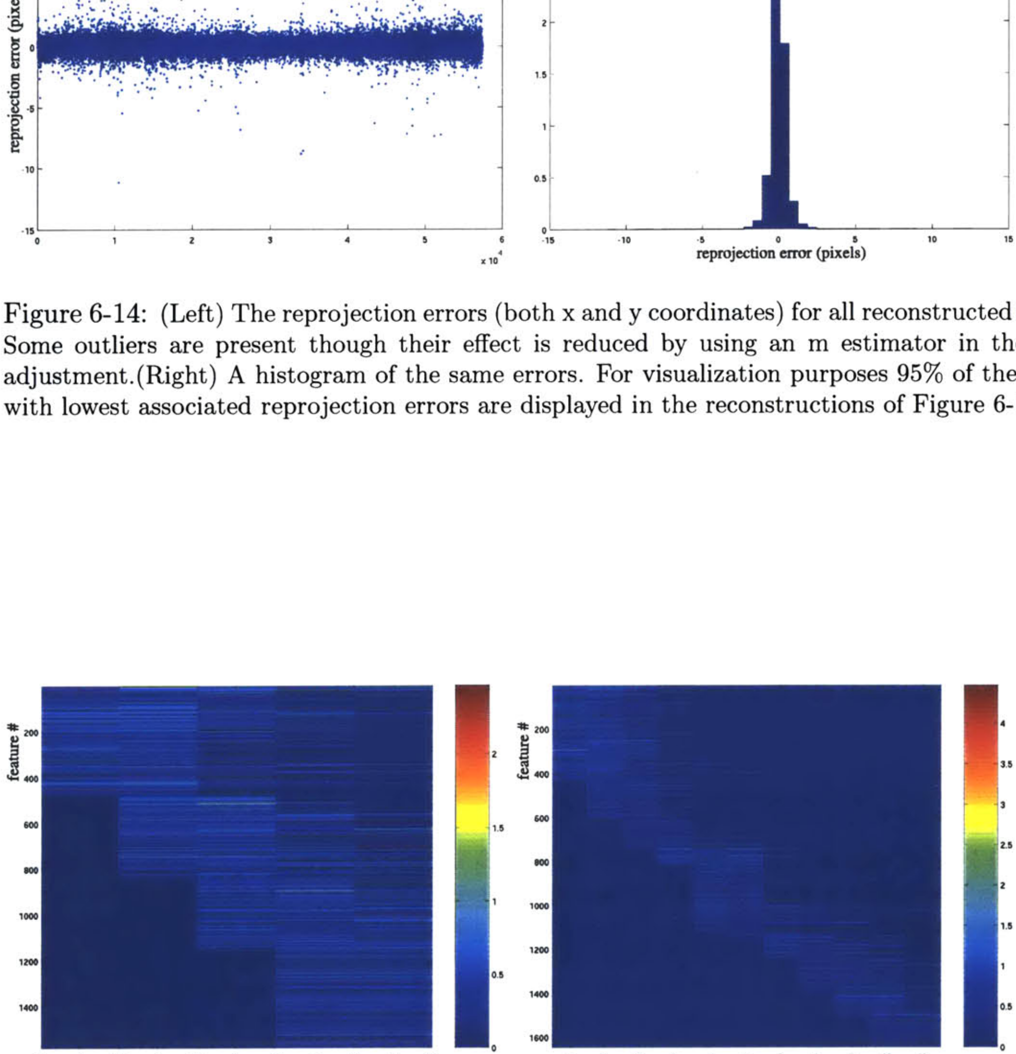


Figure 6-15: The reprojection error for submaps 10 (left) and 14 (right) displayed in a feature versus image number matrix. Most feature tracks persist between two to five images.

between temporally adjacent submaps is consistent while across track overlap is a function of the trajectory followed by the vehicle.

6.4 Self Consistency for calibration and corrections

The self-consistency imposed by refining camera poses and 3D structure provides an opportunity to refine some biases and errors present in vehicle sensors.

6.5 Compass Correction

A compass, such as the TCM2 magneto-inductive electronic compass used on the SeaBED AUV, is a low cost option for a heading reference in underwater navigation. The magnetic field around the compass can be affected by ferrous metals distorting the heading measurements. Though these effects can be minimized by hard and soft iron calibration, errors of a few degrees remain. A 3D reconstruction from imagery has 7 gauge freedoms including orientation and it is not possible to infer absolute heading from it. However, it is possible to calculate relative transformations independent of the gauge. We propose using the relative transformations between cameras of a bundle adjusted reconstruction as measurements to compare to the relative headings according to the compass measurements.

If we compare the image-based heading to the compass heading we observe that the difference $(heading_{im} - heading_{nav})$ has a roughly periodic nature to it (Figure 6-19).

This difference can be thought of as a correction term to be added to the compass heading to make it consistent. This correction does not guarantee that the compass North will correspond to true North, it only attempts to make changes in heading consistent throughout the unit circle. We describe the correction as a truncated Fourier series and solve for it via linear least squares. Given that the data appears quite noisy we only consider up to the fifth harmonic.

$$\Delta h_c(h_{nav}) = a_1 cos(h_{nav}) + b_1 sin(h_{nav}) + a_2 cos(2h_{nav}) + b_2 sin(2h_{nav}) + \dots \qquad (6.1)$$

More compactly

$$\Delta h_c(h_{nav}) = a_0 + \sum_{k=1}^{5} a_k cos(k \cdot h_{nav}) + b_k sin(k \cdot h_{nav}) \tag{6.2}$$

Given multiple measurements between heading according to imagery h_{im} and the compass h_{nav} we form a system of equations

$$\begin{bmatrix} 1 & c(h_{nav1}) & s(h_{nav1}) & c(2h_{nav1}) & s(2h_{nav1}) & \dots & s(5h_{nav1}) \\ 1 & c(h_{nav2}) & s(h_{nav2}) & c(2h_{nav2}) & s(2h_{nav2}) & \dots & s(5h_{nav2}) \\ & & & \vdots & & & \\ 1 & c(h_{navN}) & s(h_{navN}) & c(2h_{navN}) & s(2h_{navN}) & \dots & s(5h_{navN}) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \\ b_5 \end{bmatrix} = \begin{bmatrix} h_{im1} - h_{nav1} \\ h_{im2} - h_{nav2} \\ \vdots \\ h_{imN} - h_{navN} \end{bmatrix} \tag{6.3}$$

and solve for the vector of coefficients $[a_0 a_1 b_1 \cdots b_5]^\top$. Once the compass correction is available it is possible to reduce the assumed heading sensor noise in the Kalman filter and improve inter-image matching. Figure 6-20 shows the effect of compass correction on the data it was derived from. As expected, the corrected navigation-based trajectory is closer to the image-based trajectory. Figure 6-21 presents results of applying the correction on a completely independent data set. The navigation-based trajectory is consistent with the image-based trajectory once the compass correction is considered.

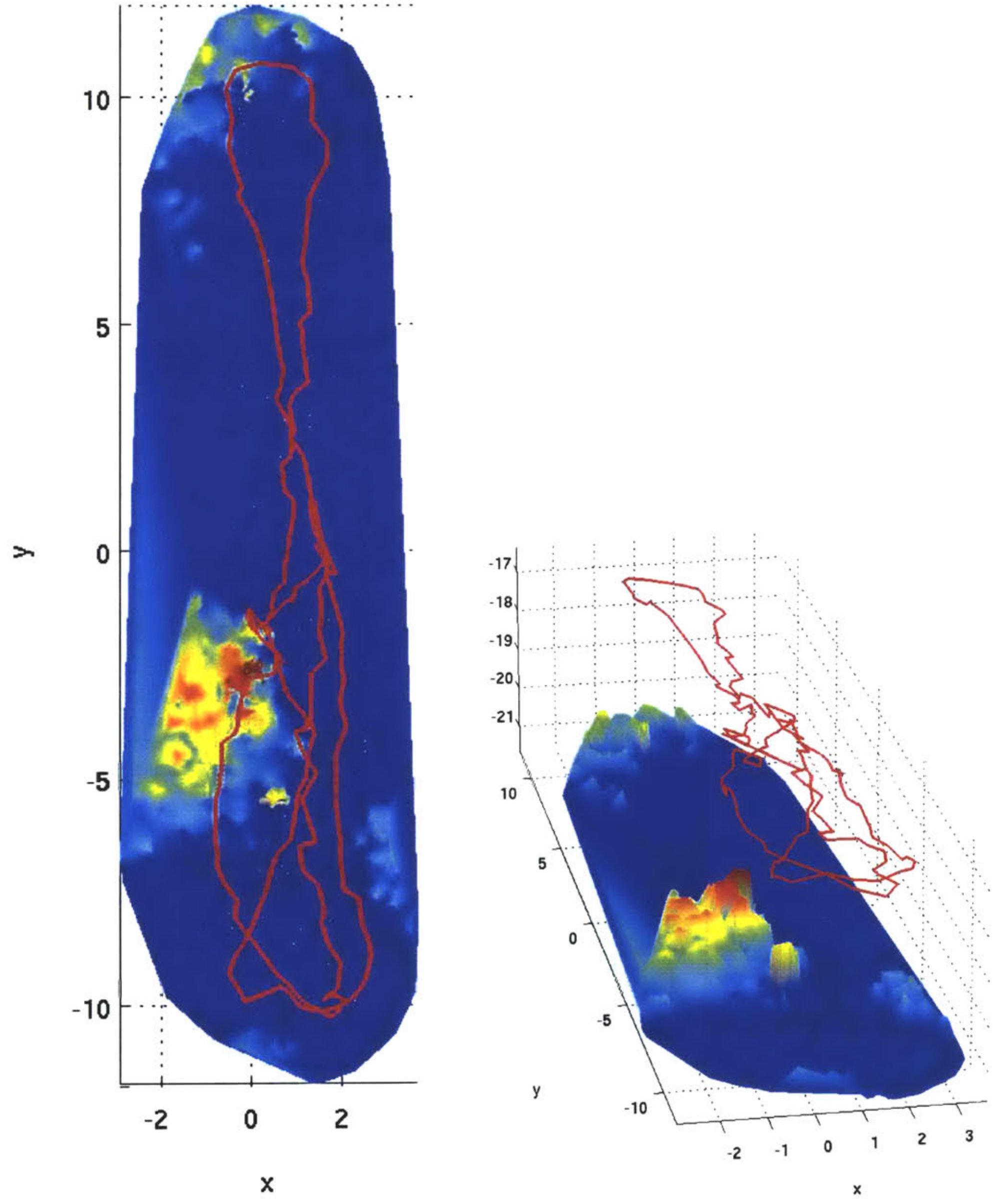


Figure 6-16: Two views of the reconstruction as a surface through the recovered 3D points. The camera trajectory is also presented as a red line. Strong swell significantly perturbed the vehicle trajectory.

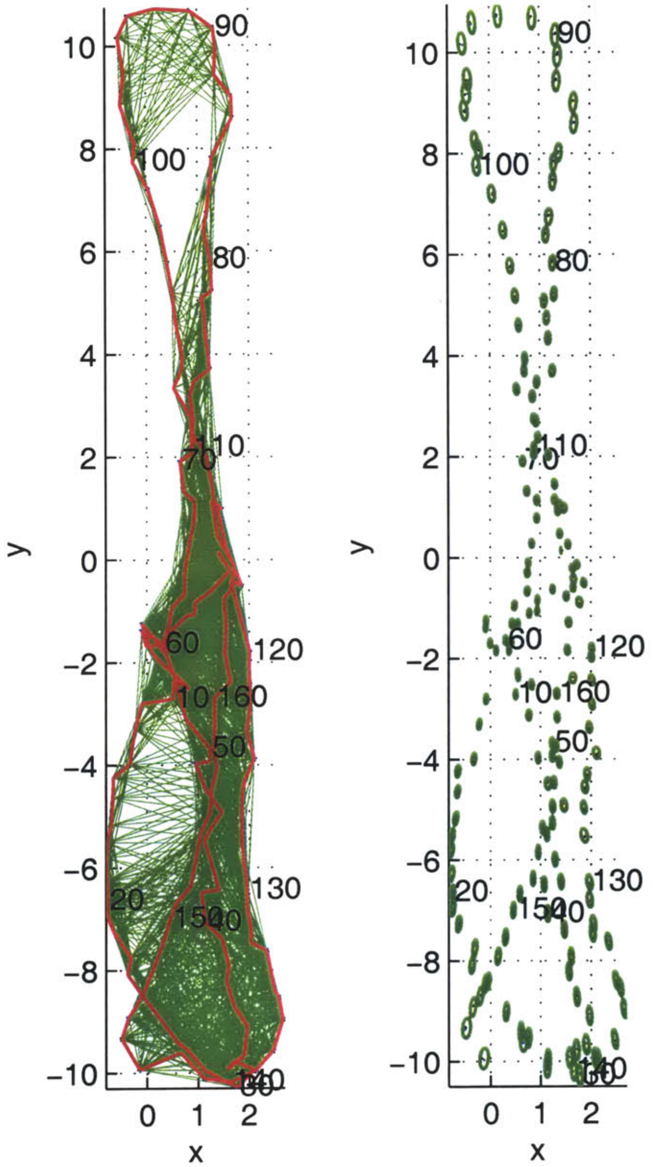


Figure 6-17: (Left) Plan view of the camera trajectory (red) and common features between cameras (green links). (Right) The 99% confidence ellipses for the xy position of the cameras. Every tenth camera is numbered on both figures to suggest the temporal sequence

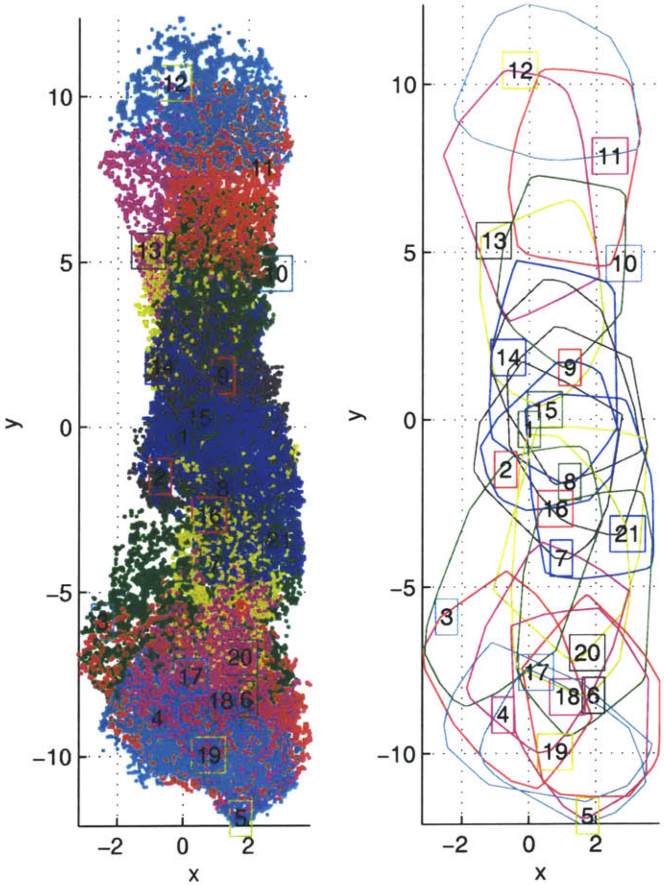


Figure 6-18: (Left) Plan view of the features for each submap. (Right) Convex hull of the 3D features of each submap. The varying degrees of spatial overlap between submaps is apparent in these figures.

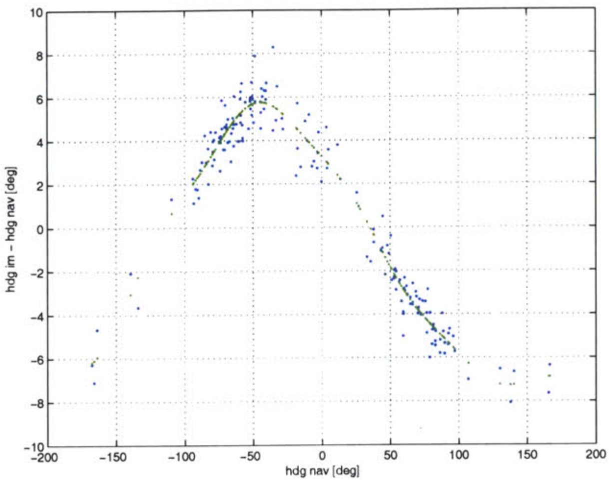


Figure 6-19: The differences in heading between the image-based poses and the compass. The approximating curve is fit using a truncated Fourier series.

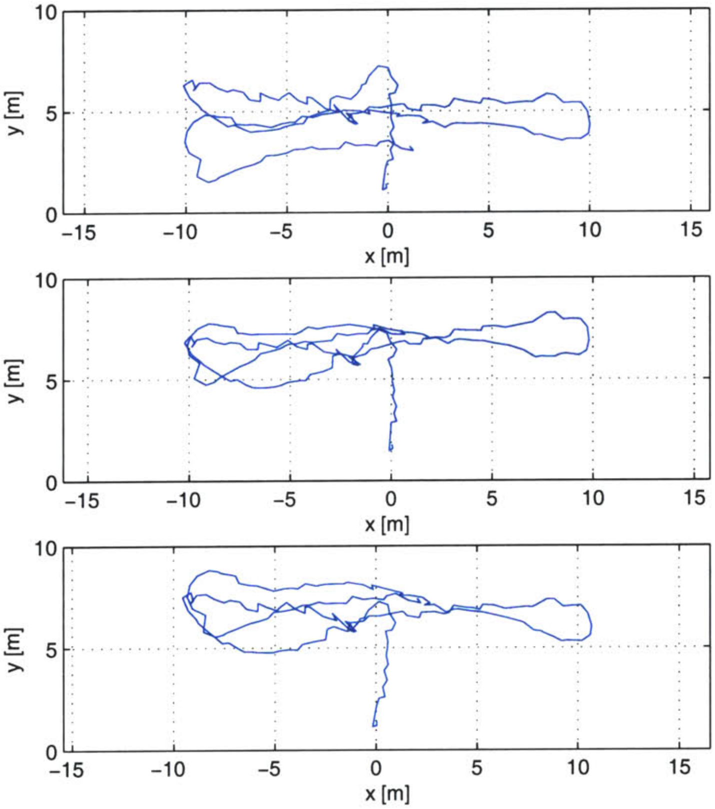


Figure 6-20: (Top) Plane view of the original navigation-based trajectory. (Middle) Image-based trajectory after bundle-adjustment. (Bottom) Corrected navigation-based trajectory.

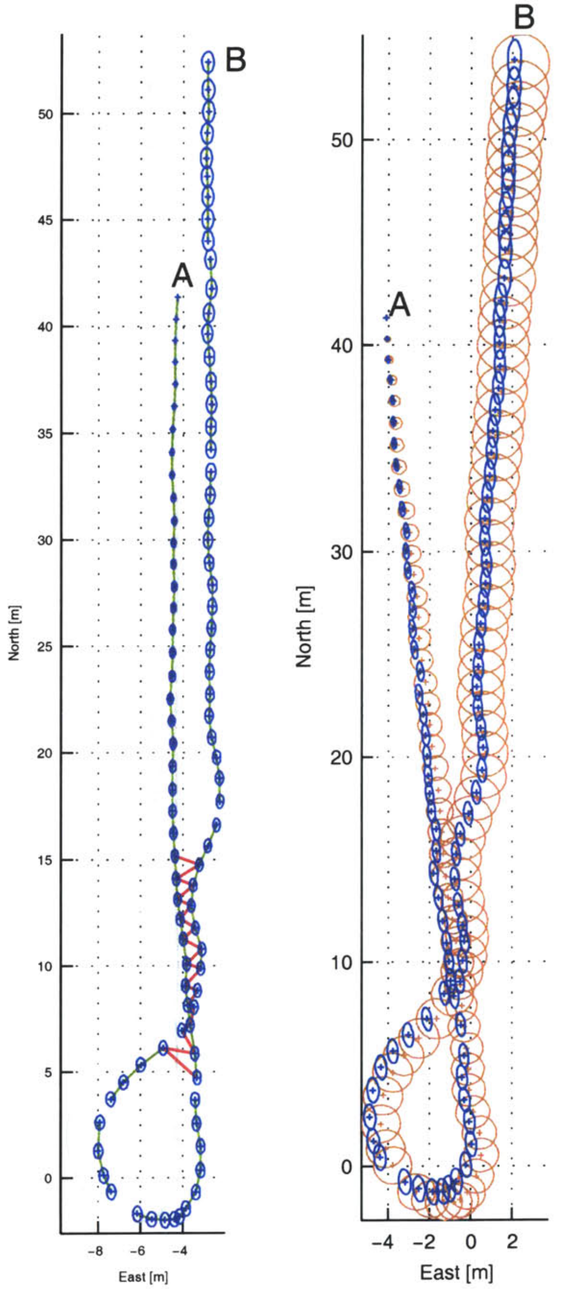


Figure 6-21: (Left) Original uncorrected navigation-based trajectory (light gray) and image-based poses (blue uncertainty ellipses). (Right) Corrected trajectory (brown uncertainty ellipses) and image-based trajectory (blue uncertainty ellipses). For the corrected case the image-based solution is within the uncertainty of the navigation. Figures courtesy of Ryan Eustice [21].

Chapter 7

Conclusions

This thesis has presented a framework for large scale structure from motion from autonomous underwater vehicles. By recognizing the constraints and challenges in underwater imaging, as well as taking advantage of the additional information provided by navigation sensors this framework is able to produce corrected paths and 3D ocean floor reconstructions from real survey data.

7.1 Limitations

Vision-based mapping relies on being able to relate images. Using interest points as features to match assumes that the scene will provide a sufficient density of such features. While navigation allows us to relate images that do not overlap, the uncertainty is higher and the map may contain 'holes'. The data used in this thesis was rich in textures such that the dreaded 'featureless bottom' did not occur.

Map matching is conservative, meaning that the system tends to miss overlapping submaps instead of proposing incorrect matches. This is mostly due to matching based both on appearance and geometry. A missed match leads to repeated features and fewer constraints on the reconstruction. These errors might not be apparent at all if there is enough redundancy in matches, or might lead to obvious shifts when the missed link was onc of the few that should have been established. This conservative approach is in evidence in the missed links in the multi-pass Bermuda data set reconstruction (Figure 6-17).

An improved approach to submap matching would limit the correspondence search between submaps when the uncertainty of the prior transformation is deemed small enough. Submap matching could also be reexamined once additional links are established that suggest that overlap should exist.

Minimization of runtimes was not a priority of this thesis. In fact, our current implementation runs in Matlab. Runtimes for processing the larger datasets were of the order of 10 hours.

7.2 Future work

7.2.1 Large Scale Autonomous Mapping

This framework explicitly focused on producing an initial guess for bundle adjustment. The temporal image sequence is considered an ordering device rather than a causal constraint. For autonomous mapping and localization a real time implementation is needed. Eustice [21] and others are already working on image based navigation but there are still many open questions on how to bring together SLAM and underwater imaging, specifically to deliver data of interest to oceanographers.

7.2.2 Imaging Underwater

This thesis demonstrated 3D reconstructions underwater assuming a simple imaging configuration of a single camera and a single strobe light. Insight gathered in the process suggests that significant improvements could be realized by designing an imaging configuration specifically for underwater structure from motion. For example, the approach used in this thesis benefits from having a motion prior from navigation. This could be improved upon by using two or more cameras with fixed baseline to complement the scale estimates from the Doppler Velocity Log. In addition, more images would be acquired for the same amount of energy expended in lighting.

Improvements in imaging sensors and lighting offer the potential of high dynamic range imagery at video rates under battery power. By narrowing the baseline, matching along the temporal sequence will be simplified. How to best match images or submaps across track remains an open question. Recent results [62] and [98] suggest that more sophisticated lighting and camera arrangements could play an important part in improving matching by engineering the lighting and shadows in a scene.

7.2.3 Applications

The ability to use images to measure and come up with estimates of uncertainty will bring some of the fruits of photogrammetry to underwater archeology such as being able to measure objects and generate euclidean 'sketches' of an underwater site.

The sparse structure produced so far can lead to dense surface reconstructions through dense stereo and dense multi-view algorithms. With dense range estimates it should be possible to correct for range and wavelength-dependent attenuation of light underwater, improving the quality of the imagery delivered by underwater vehicles.

The self-consistency of the SFM reconstruction could be exploited to fully calibrate the sensor frames and biases (up to gauge). This would be of great value in establishing self-consistency in other sensing modalities such as sonar where effective self-similarity measures are not possible.

Appendix A

Camera model and Multiple view geometry

The estimation of structure and camera poses from overlapping imagery can be understood in the context of image formation and the relationship between views of a scene. We consider images as the 2D projection of a 3D scene and present a camera model for the process.

A.1 Camera

The pinhole camera model captures the essence of the image formation process. The camera can be abstracted to a center of projection (the pinhole aperture) and an imaging plane. A 3D object point will be mapped onto the imaging plane along the ray that joins the object point and the center of projection. The basic principle at work is the collinearity of the three points that define the ray: the object (3D) point, the center of projection (pinhole aperture) and the image point.

Consider an euclidean frame with its origin at the center of projection of the camera. X and Y directions are along the imaging plane and the Z direction is out into the scene. The imaging plane is at a distance $Z = 1$ from the center of projection (and parallel to X, Y). A scene point $[X, Y, Z]^\top$ will be mapped onto $x = X/Z$ and $y = Y/Z$ by similar triangles.

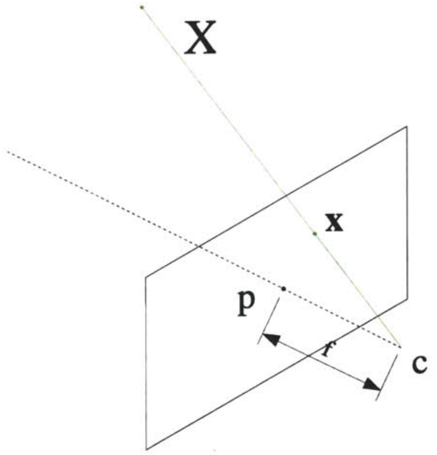


Figure A-1: The pinhole camera. The ray from the scene point $\mathbf{X}$ to the camera center $\mathbf{c}$ intersects the imaging plane at $\mathbf{x}$. The imaging plane is at a focal length f from the camera center. The projection of the camera center onto the imaging plane is the principal point $\mathbf{p}$. The ray from $\mathbf{c}$ to $\mathbf{p}$ is the optical axis (dashed).

In homogeneous coordinates this can be expressed as

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{A.1}$$

where the equality holds only up to scale for homogeneous quantities. For compactness we define the vector representation $\mathbf{x} = [x, y]^\top$, as well as its normalized homogeneous representation $\underline{\mathbf{x}} = [\mathbf{x}^\top, 1]^\top$. Likewise we define a vector of the imaged scene point in camera frame coordinates as ${}^{\mathcal{C}}\mathbf{X} = [X, Y, Z]^\top$ and its normalized homogeneous representation ${}^{\mathcal{C}}\underline{\mathbf{X}} = [{}^{\mathcal{C}}\mathbf{X}^\top, 1]^\top$. We can now express the projection as $\underline{\mathbf{x}} = \mathrm{M}\,{}^{\mathcal{C}}\underline{\mathbf{X}}$.

Intrinsic parameters

At this point the projection has only scaled Euclidean rays to scene points, such that the rays extend from the projection center to the imaging plane. In practice, the image of a scene point is reported in a reference frame that does not correspond to the physical direction of the ray. We assume that image coordinate $[u, v]^\top$ for a ray $[x, y]^\top$ is available from a CCD or CMOS sensor and that the additional coordinate transformation accounts for scaling (from a focal length f different than $Z = 1$ and from the pixel size) and translation (the origin of the image is usually the top left corner rather than the projection of the projection center onto the imaging plane), as well as skew in the pixel shape or array. This affine coordinate frame transformation can be expressed as an upper triangular matrix of intrinsic parameters, known as the calibration matrix K such that $\underline{\mathbf{u}} = \mathrm{K}\underline{\mathbf{x}}$. In more detail

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & s & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

where α_x is the focal length in pixel widths, α_y is the focal length in pixel heights, (u_0, v_0) is the coordinate of the principle point in pixels, and s is the skew in pixel shape. If K is known, the camera is considered calibrated. This allows recovering ray directions from pixel coordinates such that $\underline{\mathbf{x}} = \mathrm{K}^{-1}\underline{\mathbf{u}}$.

Exterior orientation

Scene point coordinates are normally expressed in a different reference frame than the one used by the camera, in particular if a scene is viewed from multiple cameras. The projection of a world point ${}^{w}\mathbf{X}$ onto the camera imaging plane must consider the coordinate frame transformation from world to camera reference frame.

$${}^{c}\mathbf{X} = [{}^{c}_{w}\mathrm{R}\ |{}^{c}\mathbf{t}_{cw}]{}^{w}\underline{\mathbf{X}} \qquad \text{(A.2)}$$

where ${}^{c}_{w}\mathrm{R}$ is the $[3 \times 3]$ orthonormal rotation matrix from the world frame to the camera frame, and ${}^{c}\mathbf{t}_{cw}$ is the translation from the origin of the camera frame to the world frame as seen in the camera frame.

The image point of a scene point is world coordinate frames is then

$$\underline{\mathbf{u}} = \mathrm{K}[{}^{c}_{w}\mathrm{R}\ |{}^{c}\mathbf{t}_{cw}]{}^{w}\underline{\mathbf{X}} \qquad \text{(A.3)}$$

and we define the camera projection matrix as $\mathrm{P} = \mathrm{K}[{}^{c}_{w}\mathrm{R}\ |{}^{c}\mathbf{t}_{cw}]$.

Deviations from the ideal pinhole

In practice, real cameras use lenses that capture more light than a pinhole camera, at the expense of not exactly satisfying the collinearity constraint. Radial distortion, in which projected points differ by a radial displacement from the ideal (linearly projected) points is usually the most significant deviation for short focal lengths. Decentering distortion introduces both radial and tangential components and is usually associated iwth lenses not being perfectly aligned. The distorted ray $[x_d, y_d]^{\mathrm{T}}$ can be expressed in terms of the ideal ray and distortion terms [41]:

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = \begin{bmatrix} x + \delta x_r + \delta x_t \\ y + \delta y_r + \delta y_t \end{bmatrix} \tag{A.4}$$

where the radial distortion δx_r and δy_r terms are approximated by a series expansion

$$\begin{bmatrix} \delta x_r \\ \delta y_r \end{bmatrix} = \begin{bmatrix} x(k_1 r^2 + k_2 r^4 + \ldots) \\ y(k_1 r^2 + k_2 r^4 + \ldots) \end{bmatrix} \tag{A.5}$$

with $r = \sqrt{(x - x_c)^2 + (y - yc)^2}$ and (x_c, y_c) the center of radial distortion. Usually two or three terms are sufficient to account for most of the distortion. Likewise, the tangential terms δx_t and δy_t are approximated by

$$\begin{bmatrix} \delta x_t \\ \delta y_t \end{bmatrix} = \begin{bmatrix} 2p_1 xy + p_2(r^2 + 2x^2) \\ p_1(r^2 + 2y^2) + 2p_2 xy \end{bmatrix} \tag{A.6}$$

These expressions represent the forward distortion model, which is convenient when attempting to determine the distortion parameters and ideal rays via optimization. To correct an image it is necessary to apply the inverse distortion model, which can be inverted locally from the direct model.

Underwater imaging through flat glass plates introduces significant radial distortion due to refraction from the air-glass interface and the glass-water interface. For practical fields of view (greater than 20°) this effect should not be ignored. This thesis assumes a camera calibrated in water. We use a variant of the procedure recommended by [41].

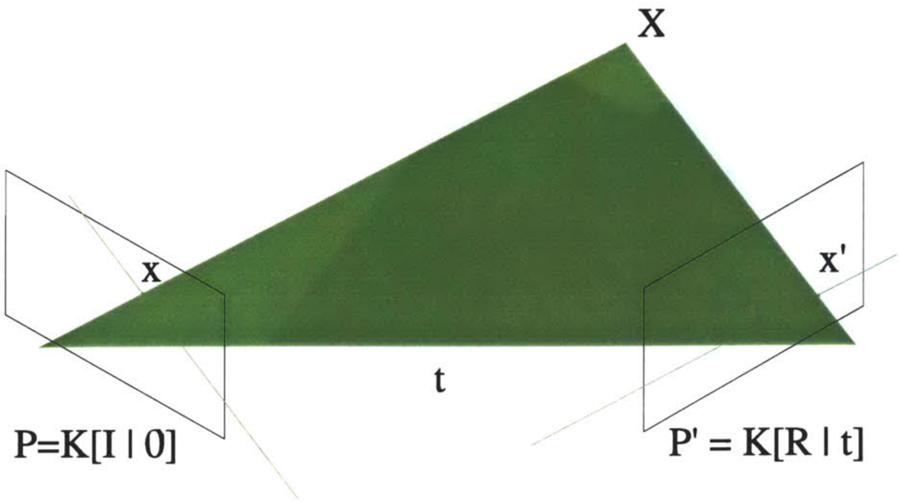


Figure A-2: Two view geometry. The rays $\underline{\mathbf{x}}$,$\underline{\mathbf{x}}'$ and $\mathbf{t}$ form the epipolar plane. This plane intersects the imaging planes in the epipolar lines.

A.2 Multi-view Geometry

The recovery of structure and motion from multiple images relies on tightly coupled parameters: the scene structure parameters expressed relative to some reference frame, the camera parameters (internal and external orientation) and the correspondences which associate scene points to their projections. It is important to understand how views of a common scene are related.

A.2.1 Two View Geometry

Given an image of a scene point, the collinearity condition implies that the 3D point must lie on the ray back-projected from the projection center of the camera and through the image point of the point on the imaging plane. On a second camera viewing the same scene, this ray will be imaged as a line and the image of the 3D point in the second camera will lie on that line, known as the *epipolar line.* The *epipolar plane* contains the scene point and the camera centers such that the rays from the camera centers to the scene points and the ray between camera centers (the baseline vector) all lie in the plane, satisfying the coplanarity constraint for image points in correspondence (Figure A-2 and A-3).

The epipolar lines are the intersection of the epipolar planes with the imaging planes.

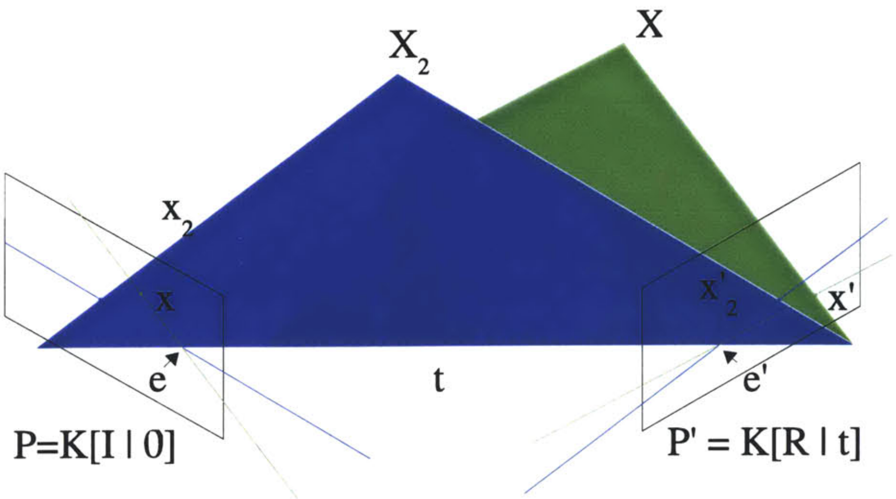


Figure A-3: Multiple epipolar planes intersect on the baseline $\mathbf{t}$ and define the epipoles $\underline{\mathbf{e}}$ and $\underline{\mathbf{e}}'$ on the imaging planes.

They correspond to the image the ray going from the object point to one camera makes on the other camera. Multiple epipolar planes (for different scene points) will all contain the camera centers and the baseline vector joining them. The *epipoles* are the intersection of the baseline with the imaging planes and correspond to the image of the one camera's center on the other view.

In order for the image rays and the baseline vector to be on a plane their triple product must be null. Assume projective camera matrices $\mathrm{P} = \mathrm{K}[\mathrm{I} \mid \mathbf{0}]$ and $\mathrm{P}' = \mathrm{K}[\mathrm{R} \mid \mathbf{t}]$. The triple product for a ray $\underline{\mathbf{x}}$ in the first camera, its corresponding ray $\underline{\mathbf{x}}'$ and the baseline $\mathbf{t}$ in the reference frame of the second camera is

$$\underline{\mathbf{x}}^\top \cdot (\mathbf{t} \times \mathrm{R}\underline{\mathbf{x}}) = 0 \tag{A.7}$$

Where $R\underline{\mathbf{x}}$ is the rotation of $\underline{\mathbf{x}}$ into the frame of the second camera. We define the *essential matrix* E as

$$\mathrm{E} = [\mathbf{t}]_\times \mathrm{R} \tag{A.8}$$

where $[\mathbf{t}]_\times$ is the skew symmetric matrix based on $\mathbf{t}$ such that $[\mathbf{t}]_\times \mathbf{a} = \mathbf{t} \times \mathbf{a}$. The essential matrix encodes the relative pose between two cameras up to scale of the baseline. The epipolar constraint is expressed as

$$\underline{\mathbf{x}}'^{\mathsf{T}}\mathrm{E}\underline{\mathbf{x}} = 0 \tag{A.9}$$

The elements of the essential matrix can be recovered up to a scale factor from point correspondences.

In the case of uncalibrated cameras, the epipolar constraint is still valid between pixel coordinates given that $\underline{\mathbf{x}} = \mathrm{K}^{-1}\underline{\mathbf{u}}$:

$$\underline{\mathbf{x}}'^{\mathsf{T}}\mathrm{E}\underline{\mathbf{x}} = \underline{\mathbf{u}}'^{\mathsf{T}}\mathrm{K}^{-\mathsf{T}}\mathrm{E}\mathrm{K}\underline{\mathbf{u}} \tag{A.10}$$

The *fundamental matrix* $\mathrm{F} = \mathrm{K}^{-\mathsf{T}}\mathrm{E}\mathrm{K}$ then satisfies $\underline{\mathbf{u}}'^{\mathsf{T}}\mathrm{F}\underline{\mathbf{u}} = 0$.

Most recent multi view computer vision applications rely on this relationship to calculate an fundamental matrix from correspondences and given an fundamental matrix, restrict the search for correspondences.

A.2.2 Triangulation

If camera poses and image points of a scene point are known, it is possible to determine the location of the scene point by intersecting the rays back-projected from each camera. Linear triangulation methods are simple but do not minimize a physically meaningful quantity in the case of noisy measurements. Starting from the collinearity condition $\underline{\mathbf{x}}_i = \mathrm{P}_i\underline{\mathbf{X}} = \mathrm{R}_i\mathbf{X} + \mathbf{t}_i$ and taking the cross product $\underline{\mathbf{x}}_i \times \mathrm{P}_i\underline{\mathbf{X}} = 0$ or $[\underline{\mathbf{x}}_i]_\times\mathrm{R}_i\mathbf{X} = [\underline{\mathbf{x}}_i]_\times\mathbf{t}$; each view of a 3D point provides two independent equations, so that with two views it is possible to solve for the three unknowns of $\mathbf{X}$.

If the projections are noisy it is possible to solve for the ideal projections that satisfy the epipolar constraint and intersect [40].

Another possibility is to use the noisy measurements and solve for a 3D point that minimizes the distance to all rays that should intersect but don't because of noise [15].

Appendix B

Sensors and Navigation

B.1 Overview

The imagery collected in optical surveys of the ocean floor presents a challenging application of traditional structure from motion techniques. Since underwater vehicles are the platform of choice for deep or extensive surveys the additional instruments on the vehicle can be used to increase the reliability of the reconstruction process. This chapter describes the basic set of navigation instruments on an ocean-going AUV, and a recursive filter implementation to estimate the vehicle trajectory.

B.2 Sensors on robotic underwater vehicles

We used a pose instrumented underwater vehicle with a downward-looking calibrated camera. The vehicle has the basic set of instruments used for scientific surveys. It uses an acoustic Doppler velocity log (DVL) [99] to measure velocity and altitude relative to the bottom. Typical speeds are in the order of 0 to $1m/s$ with accuracies in the order of mm/s. Absolute orientation (in the world frame) is measured to within a few degrees over the survey area by a magnetic compass and inclinometers. A pressure sensor provides depth measurements with depth-dependent accuracies on the order of 0.01% which can be considered as a bounded accuracy. A rate sensor provides angular velocities with accuracies on the order of $1^\circ/s$. Table B.1 summarizes the sensors and their characteristics. The vehicle

Variable	**Instrument**	**Precision**	**Type**	**Range**	**Update Rate**
Body Velocities (u, v, w)	Bottom-Lock DVL	1 mm/s	Proprio	30-200 m	Fast: 1-5 Hz
Heading ψ	electronic compass	2°	Extero	360°	Medium: 1-2 Hz
Roll/Pitch (ϕ, θ)	2-axis tilt sensors	0.5°	Extero	±20°	Medium: 1-2 Hz
Depth (z)	Pressure sensor	0.01 m	Extero	full ocean	Medium: 1 Hz
Altitude	Altimeter / DVL	0.1 m	Extero	Varies	Varies: 0.1-10 Hz
Angular Rates (p, q, r)	3-axis gyro	1°/s	Proprio	$50°/s$	Fast: 5-10 Hz

Table B.1: Summary of sensors typically used on oceanographic AUVs.

xy position is estimated from integrating velocities which leads to an unbounded growth in the uncertainty. Though external references for xy based on triangulation with beacons are certainly used in oceanography, and the framework we propose can take advantage of it, we focus on the case where no additional beacons are deployed as this mode of operation seems particularly suited for fast, low cost exploration with AUVs.

AUVs have limited power budgets that do not allow for continuous lighting with conventional sources (incandescent bulbs). Instead, strobed lighting is used to acquire still images. For extended surveys, the lowest overlap admissible for the scientific objectives is used since energy consumption of the strobe is proportional to the number of images acquired.

B.3 Vehicle and sensor frames

Each sensor provides a measurement of pose in a specific frame of reference. We assume this measurement corresponds to a random variable. For engineering purposes, the measurement is described by the first two moments (mean and covariance). We wish to estimate the trajectory of the vehicle given multiple measurements. The trajectory can be thought of as a time-varying pose. The vehicle *state* is formed by the pose and additional variables useful for state propagation.

The local-level frame is a convenient reference frame to describe the vehicle pose. It corresponds to a right-hand frame positioned at the surface of the ocean (zero depth) with axes oriented as $+X \rightarrow$ North, $+Y \rightarrow$East, $+Z \rightarrow$Down. The 6DOF vehicle pose vector

$\mathbf{x}_{\ell v}$ describes the vehicle frame relative to the local-level frame:

$$\mathbf{x}_{\ell v} = \begin{bmatrix} {}^{\ell}\mathbf{t}_{\ell v} \\ \boldsymbol{\Theta}_{\ell v} \end{bmatrix} \tag{B.1}$$

Where ${}^{\ell}\mathbf{t}_{\ell v} = [x, y, z]^{\top}$ is the position (the vector from local-level origin to vehicle frame origin as described in the local-level frame) and $\boldsymbol{\Theta}_{\ell v} = [\phi, \theta, \psi]^{\top}$ is the orientation represented by roll, pitch, heading Euler angles [34].

Vehicle motion is represented in a body-fixed frame with a generalized velocity vector $\boldsymbol{\nu}$

$$\boldsymbol{\nu} = \begin{bmatrix} {}^{v}\mathbf{u} \\ {}^{v}\boldsymbol{\omega} \end{bmatrix} \tag{B.2}$$

Where ${}^{v}\mathbf{u} = [u, v, w]^{\top}$ is the vector of body-frame linear velocities and ${}^{v}\boldsymbol{\omega} = [p, q, r]^{\top}$ is the body-fixed angular velocity [34].

The linear velocities transformation from body-frame to local-level is given by ${}^{\ell}\dot{\mathbf{t}}_{\ell v} = {}^{\ell}_{v}\mathrm{R}(\boldsymbol{\Theta}_{\ell v}){}^{v}\mathbf{u}$ where the orthonormal rotation matrix ${}^{\ell}_{v}\mathrm{R}(\boldsymbol{\Theta}_{\ell v}) = \mathrm{R}_{z,\psi}\mathrm{R}_{y,\theta}\mathrm{R}_{x,\phi}$ follows the zyx-convention for Euler angles [34].

The angular velocity transformation expressing body frame angular rates as time derivatives of the Euler angles $\dot{\boldsymbol{\Theta}}_{\ell v} = \mathrm{T}_{\Theta}(\boldsymbol{\Theta}_{\ell v}){}^{v}\boldsymbol{\omega}$ where $\mathrm{T}_{\Theta}(\boldsymbol{\Theta}_{\ell v})$ is more easily described by the inverse transformation:

$${}^{v}\boldsymbol{\omega} = \begin{bmatrix} \dot{\phi} \\ 0 \\ 0 \end{bmatrix} + \mathrm{R}_{x,\phi}^{\top} \begin{bmatrix} 0 \\ \dot{\theta} \\ 0 \end{bmatrix} + \mathrm{R}_{x,\phi}^{\top}\mathrm{R}_{y,\theta}^{\top} \begin{bmatrix} 0 \\ 0 \\ \dot{\psi} \end{bmatrix} = \mathrm{T}_{\Theta}^{-1}(\boldsymbol{\Theta}_{\ell v})\dot{\boldsymbol{\Theta}}_{\ell v} \tag{B.3}$$

The 6DOF kinematic equations transform the velocities in body-frame to the local-level:

$$\begin{bmatrix} {}^{\ell}\dot{\mathbf{t}}_{\ell v} \\ \dot{\boldsymbol{\Theta}}_{\ell v} \end{bmatrix} = \begin{bmatrix} {}^{\ell}_{v}\mathrm{R}(\boldsymbol{\Theta}_{\ell v}) & 0_{3\times 3} \\ 0_{3\times 3} & \mathrm{T}_{\Theta}(\boldsymbol{\Theta}_{\ell v}) \end{bmatrix} \begin{bmatrix} {}^{v}\mathbf{u} \\ {}^{v}\boldsymbol{\omega} \end{bmatrix} \tag{B.4}$$

which can be summarized as:

$$\dot{\mathbf{x}}_{\ell v} = \mathrm{M}(\mathbf{x}_{\ell v})\boldsymbol{\nu} \tag{B.5}$$

B.4 Vehicle trajectory estimation

We propose using an Extended Kalman filter (EKF) [6] to generate estimates of the vehicle pose based on navigation sensors. Poses derived from instruments were used in the process of generating a 3D reconstruction from imagery to constrain searches and regularize optimizations. The standard Kalman filter formulation requires defining a state vector, a process model that describes the evolution of the state and an observation model that relates the state to sensor measurements.

The vehicle state vector as the vehicle pose and generalized velocities

$$\mathbf{x}_v = \begin{bmatrix} \mathbf{x}_{\ell v} \\ \boldsymbol{\nu} \end{bmatrix}. \tag{B.6}$$

Though our principal goal is estimating the vehicle pose at instants when images were acquired, the velocities aid in propagating the vehicle pose through time, as suggested by the kinematic equations of the previous section.

We assume the vehicle state evolves according to a process model $\mathbf{f}_v(t)$ driven by white noise $w(t) \sim N(\mathbf{0}, \mathrm{Q}(t))$. The sensor measurements are incorporated through a discrete time observation model $\mathbf{h}_v(t_k)$ in the presence of time-independent additive Gaussian noise $v[t_k] \sim N(\mathbf{0}, \mathrm{R}_k)$ which is uncorrelated with the process noise, $E\left[\mathbf{w}\mathbf{v}^\top\right]$.

$$\dot{\mathbf{x}}_v(t) = \mathbf{f}_v(\mathbf{x}_v(t), t) + \mathbf{w}(t) \tag{B.7}$$

$$z[t_k] = \mathbf{h}_v(\mathbf{x}_v[t_k], t_k) + \mathbf{v}[t_k] \tag{B.8}$$

The vehicle state $\mathbf{x}_v$ and its covariance P_v are estimated using extended Kalman filter (EKF) equations for the system B.7 with Jacobian of the process model $\mathrm{F}_v = \left.\frac{\partial \mathbf{f}_v(\mathbf{x}_v(t),t)}{\partial \mathbf{x}_v(t)}\right|_{\bar{\mathbf{x}}_v(t)}$ and of the observation model $\mathrm{H}_v = \left.\frac{\partial \mathbf{h}_v(\mathbf{x}_v[t_k],t_k)}{\partial \mathbf{x}_v[t_k]}\right|_{\bar{\mathbf{x}}_v[t]}$

- The prediction step is given by

$$\dot{\bar{\mathbf{x}}}_v = \mathbf{f}_v\left(\bar{\mathbf{x}}_v(t), t\right) \tag{B.9}$$

$$\dot{\mathrm{P}}_v(t) = \mathrm{F}_v \mathrm{P}_v(t) + \mathrm{P}_v(t)\mathrm{F}_v^\top + \mathrm{Q}(t) \tag{B.10}$$

- while the update step is given by

$$\mathrm{K} = \mathrm{P}_v^- \mathrm{H}_v^\top \left[\mathrm{H}_v \mathrm{P}_v^- \mathrm{H}_v^\top + \mathrm{R}_k\right]^{-1} \tag{B.11}$$

$$\bar{\mathbf{x}}_v^+ = \bar{\mathbf{x}}_v^- + \mathrm{K}\left[\mathbf{z}\left[t_k\right] - \mathbf{h}_v\left(\bar{\mathbf{x}}_v^-, t_k\right)\right] \tag{B.12}$$

$$\mathrm{P}_v^+ = \left[\mathrm{I} - \mathrm{K}\mathrm{H}_v\right] \mathrm{P}_v^- \left[\mathrm{I} - \mathrm{K}\mathrm{H}_v\right]^\top + \mathrm{K}\mathrm{R}_k\mathrm{K}^\top \tag{B.13}$$

B.5 Vehicle process model

An underwater imaging platform has relatively slow dynamics. We choose to approximate the vehicle dynamics with a constant velocity process model.

$$\dot{\mathbf{x}}_v = \begin{bmatrix} \dot{\mathbf{x}}_{\ell v} \\ \dot{\boldsymbol{\nu}} \end{bmatrix} = \begin{bmatrix} 0_{6\times 6} & \mathrm{M}(\mathbf{x}_{\ell v}) \\ 0_{6\times 6} & 0_{6\times 6} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\ell v} \\ \boldsymbol{\nu} \end{bmatrix} + \begin{bmatrix} 0_{6\times 1} \\ \mathbf{w}_{\boldsymbol{\nu}} \end{bmatrix} \tag{B.14}$$

where $\mathrm{M}(\mathbf{x}_{\ell v})$ encodes the generalized velocity transformation from vehicle to local-level (B.4), and $\mathbf{w}_{\boldsymbol{\nu}} = [\mathbf{w}_{\mathbf{u}}^\top, \mathbf{w}_{\boldsymbol{\omega}}^\top]^\top$ is the process noise that accounts for unmodeled dynamics. The process model allows us to propagate the state between navigation sensor measurements, essentially rotating body frame velocities into the local-level. Since the process model is time varying and nonlinear, the prediction is performed by using a 4th order Runge-Kutta approximation to integrate the state derivatives.

B.6 Vehicle observation model

We seek to estimate the pose of the vehicle in the local-level frame using multiple sensors. Navigation sensors can be abstracted into *proprioceptive* sensors that measure motion of the vehicle in the sensor frame and *exteroceptive* sensors that provide information about the vehicle pose relative to its environment. Proprioceptive sensors include the Doppler Velocity Log (DVL), angular rate sensors, accelerometers, wheel encoders, etc. A robot's change in pose can be estimated by integrating these measurements in the appropriate reference frame. Exteroceptive sensors include the depth sensor (which measures range to the surface), the compass (for heading relative to the local magnetic field), and tilt sensors that provide orientation relative to gravity. Receivers that triangulate ranges from acoustic beacons and GPS receiver are also exteroceptive sensors.

B.6.1 Proprioceptive Sensors

Proprioceptive sensors provide motion measurements in the sensor frame, which can be transformed into the vehicle frame by knowledge of the static sensor to vehicle transformation. Proprioceptive measurements can then be placed in the local-level frame by using the rigid body transformation implied by the vehicle pose. However, the estimation process uses the discrepancy between the predicted measurement according to the current state estimate and the actual measurement to correct the state estimate. For proprioceptive sensors we choose to use the measurement in the sensor frame as the observation which requires formulating an observation model that transforms the vehicle state into a predicted sensor reading in the sensor frame.

The coordinate transformation from vehicle to sensor frame can be represented by a homogeneous transformation:

$$ {}^{s}_{v}\mathrm{T} = \begin{bmatrix} {}^{s}_{v}\mathrm{R} & {}^{s}\mathbf{t}_{sv} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \tag{B.15} $$

Since these sensors measure motion the velocities and angular rates in the vehicle frame are expressed in the sensor frame as

$$ {}^{s}\mathbf{u} = {}^{s}_{v}\mathrm{R}\,({}^{v}\mathbf{u} + {}^{v}\boldsymbol{\omega} \times {}^{v}\mathbf{t}_{vs}) \tag{B.16} $$

$$ {}^{s}\boldsymbol{\omega} = {}^{s}_{v}\mathrm{R} \cdot {}^{v}\boldsymbol{\omega} \tag{B.17} $$

where ${}^{v}_{s}\mathrm{T} = {}^{s}_{v}\mathrm{T}^{-1}$ is a fixed transformation that describes how the sensor is mounted relative to the vehicle reference frame.

B.6.2 Exteroceptive Sensors

Exteroceptive sensors express the sensor pose in the sensor local-level frame s_o. In its general form the observation model expresses the vehicle to local-level transformation as a sensor to sensor local-level transformation by use of the static vehicle to sensor transformation and sensor local-level to (vehicle) local-level transformation. In homogeneous transformation notation:

$$ {}^{s_o}_{s}\mathrm{T} = {}^{s_o}_{\ell}\mathrm{T} \cdot {}^{\ell}_{v}\mathrm{T} \cdot {}^{v}_{s}\mathrm{T}. \tag{B.18}$$

This composition of coordinate frame transformations can also be expressed compactly using Smith, Self and Cheeseman's notation [114] as

$$\mathbf{x}_{s_o s} = \mathbf{x}_{s_o \ell} \oplus \mathbf{x}_{\ell v} \oplus \mathbf{x}_{vs}. \tag{B.19}$$

It is common that the sensor local-level s_o corresponds to the vehicle local-level ℓ in which case ${}^{s_o}_{\ell}\mathrm{T}$ is the identity transformation. This is the case for the depth sensor and the compass and tilt sensors.

B.6.3 DVL observation model

The observation model for the DVL, which returns velocities ${}^{d}\mathbf{u}$ is

$$ {}^{d}\mathbf{u} = {}^{d}_{v}\mathrm{R}\,({}^{v}\mathbf{u} + {}^{v}\boldsymbol{\omega} \times {}^{v}\mathbf{t}_{vd}) \tag{B.20}$$

B.6.4 Rate sensor observation model

The observation equation for the orientation rates from the gyroscope are

$$ {}^{g}\boldsymbol{\omega} = {}^{g}_{v}\mathrm{R} \cdot {}^{v}\boldsymbol{\omega} \tag{B.21}$$

B.6.5 Depth sensor observation model

In most cases, the sensor provides only partial pose information. For example, the Paroscientific depth sensor provides pose information of the z coordinate of the sensor in the local-level, ${}^{\ell}z_{\ell p}$. This can be extracted from B.18 to yield

$$ {}^{\ell}z_{\ell p} = {}^{\ell}_{v}\mathbf{R}_3^{\top} \cdot {}^{v}\mathbf{t}_{vp} + {}^{\ell}z_{\ell v} \tag{B.22}$$

where ${}^{\ell}_{v}\mathbf{R}_3^{\top}$ represents the third row of ${}^{\ell}_{v}\mathrm{R}$. This could be expressed in composition notation as

$$ z_{\ell p} = z\left(\mathbf{x}_{\ell v} \oplus \mathbf{x}_{vp}\right) \tag{B.23}$$

as a shorthand for

$$\begin{bmatrix} \cdot \\ \cdot \\ z \\ \cdot \end{bmatrix}_{\ell p} = \begin{bmatrix} \cdot \\ \cdot \\ z \\ \boldsymbol{\Theta} \end{bmatrix}_{\ell v} \oplus \begin{bmatrix} x \\ y \\ z \\ \cdot \end{bmatrix}_{vp} \tag{B.24}$$

B.6.6 Attitude sensor observation model

We consider the heading and tilt sensors as an attitude module. The observation model for attitude from compass and tilt sensors, ${}^{\ell}_{a}\boldsymbol{\Theta}$ is also a case of partial pose information. From B.18 we have

$$ {}^{\ell}_{a}\mathrm{R}({}^{\ell}_{a}\boldsymbol{\Theta}) = {}^{\ell}_{v}\mathrm{R}({}^{\ell}_{v}\boldsymbol{\Theta}) \cdot {}^{v}_{a}\mathrm{R}({}^{v}_{a}\boldsymbol{\Theta}) \tag{B.25}$$

which can be written using compositions of orientations as

$$ \boldsymbol{\Theta}_{\ell a} = \boldsymbol{\Theta}\left(\mathbf{x}_{\ell v} \oplus \mathbf{x}_{va}\right) = \boldsymbol{\Theta}_{\ell v} \oplus \boldsymbol{\Theta}_{va} \tag{B.26}$$

which is shorthand for

$$\begin{bmatrix} \cdot \\ \boldsymbol{\Theta} \end{bmatrix}_{\ell a} = \begin{bmatrix} \cdot \\ \boldsymbol{\Theta} \end{bmatrix}_{\ell v} \oplus \begin{bmatrix} \cdot \\ \boldsymbol{\Theta} \end{bmatrix}_{va} \tag{B.27}$$

B.7 Augmented state for relative pose estimation

The state vector representation of B.6 is convenient for trajectory estimation, and the poses from the trajectory can be used to form a relative pose prior for matching temporally adjacent images. The vehicle state is correlated in time scales of a few seconds since dynamics tend to be slow and the xy position is derived from integrating velocities from a previous position. It is appropriate to consider the cross-covariances between two poses when calculating the uncertainty of the relative pose between them. These cross-covariances are lost in the simple trajectory estimation case, so we propose using an augmented state vector that keeps the vehicle pose estimate for the previous camera [21]. This allows calculating relative pose between temporally adjacent images with full covariance.

Consider that at time t_i image I_i is acquired and that the next image I_j is acquired at time t_j. In order to estimate the cross-covariance between states at times t_i and t_j we augment the state vector with a delayed state corresponding to the pose of the vehicle when the image was taken $\mathbf{x}_{\ell v}(t_i) = \mathbf{x}_{\ell v_i}$. For $t_i \leq t \leq t_j$ the augmented state and covariance matrix are:

$$\mathbf{x}_{aug}(t) = \begin{bmatrix} \mathbf{x}_{\ell v}(t) \\ \boldsymbol{\nu}(t) \\ \mathbf{x}_{\ell v_i} \end{bmatrix}, \mathrm{P}_{aug}(t) = \begin{bmatrix} \mathrm{P}_{\mathbf{x}_{\ell v}}(t) & \mathrm{P}_{\mathbf{x}_{\ell v}\boldsymbol{\nu}}(t) & \mathrm{P}_{\mathbf{x}_{\ell v}\mathbf{x}_{\ell v_i}}(t) \\ \mathrm{P}^{\mathrm{T}}_{\mathbf{x}_{\ell v}\boldsymbol{\nu}}(t) & \mathrm{P}_{\boldsymbol{\nu}}(t) & \mathrm{P}_{\boldsymbol{\nu}\mathbf{x}_{\ell v_i}}(t) \\ \mathrm{P}^{\mathrm{T}}_{\mathbf{x}_{\ell v}\mathbf{x}_{\ell v_i}}(t) & \mathrm{P}^{\mathrm{T}}_{\boldsymbol{\nu}\mathbf{x}_{\ell v_i}}(t) & \mathrm{P}_{\mathbf{x}_{\ell v_i}} \end{bmatrix} \tag{B.28}$$

At time t_j the the next image is acquired and at that point we will have propagated the covariance of the vehicle state and its covariance with the pose for the previous image. We have all the pieces needed to calculate the relative pose $\mathbf{x}_{v_i v_j}$ and its uncertainty $\mathrm{P}_{\mathbf{x}_{v_i v_j}}$ using the tail-to-tail transformation $\mathbf{x}_{v_i v_j} = \ominus \mathbf{x}_{\ell v_i} \oplus \mathbf{x}_{\ell v_j}$. The covariance is given by:

$$\mathrm{P}_{\mathbf{x}_{v_i v_j}} = {}_{\ominus}\mathrm{J}_{\oplus} \begin{bmatrix} \mathrm{P}_{\mathbf{x}_{\ell v_i}} & \mathrm{P}_{\mathbf{x}_{\ell v_i}\mathbf{x}_{\ell v_j}} \\ \mathrm{P}^{\mathrm{T}}_{\mathbf{x}_{\ell v_i}\mathbf{x}_{\ell v_j}} & \mathrm{P}_{\mathbf{x}_{\ell v_j}} \end{bmatrix} {}_{\ominus}\mathrm{J}^{\mathrm{T}}_{\oplus} \tag{B.29}$$

with the ${}_{\ominus}\mathrm{J}_{\oplus}$ the Jacobian of the tail-to-tail relationship

$${}_{\ominus}\mathrm{J}_{\oplus} = \frac{\partial \mathbf{x}_{v_i v_j}}{\partial(\mathbf{x}_{\ell v_i}, \mathbf{x}_{\ell v_j})} = \frac{\partial \mathbf{x}_{v_i v_j}}{\partial(\mathbf{x}_{v_i \ell}, \mathbf{x}_{\ell v_j})} \frac{\partial(\mathbf{x}_{v_i \ell}, \mathbf{x}_{\ell v_j})}{\partial(\mathbf{x}_{\ell v_i}, \mathbf{x}_{\ell v_j})} \tag{B.30}$$

Before the filter continues to process new measurements, the relative pose and its uncertainty are stored and the delayed state $\mathbf{x}_{\ell v_i}$ is replaced with $\mathbf{x}_{\ell v_j}$. The covariance and cross-covariances are also initialized to the new delayed state. This basically allows for the estimation to have enough 'memory' to establish relative poses between temporally adjacent images while estimating a trajectory in a global frame.

The augmentation of the process model is such that the vehicle state continues to evolve according to the dynamic model $\mathbf{f}_v$ and the delayed state does not evolve

$$\dot{\mathbf{x}}_{aug} = \begin{bmatrix} \mathbf{f}_v\left(\mathbf{x}_v(t), t\right) + \mathbf{w}(t) \\ \mathbf{0}_{[6\times1]} \end{bmatrix} \tag{B.31}$$

The sensor observation model only depends on the current vehicle state and not the delayed state.

Bibliography

[1] M. Antone and S. Teller. Scalable extrinsic calibration of omni-directional image networks. *International Journal of Computer Vision*, 49(2-3):143–174, 2002.

[2] F. Badra, A. Qumsieh, and G. Dudek. Rotation and zooming in image mosaicing. In *Fourth IEEE Workshop on Applications of Computer Vision*, pages 50–55, Princeton, NJ, 1998.

[3] T. Bailey. *Mobile Robot Localisation and Mapping in Extensive Outdoor Environments.* Phd, Australian Center for Field Robotics, University of Sydney, August 2002.

[4] R.D. Ballard, A.M. McCann, D.R. Yoerger, L.L. Whitcomb, D.A. Mindell, J. Oleson, H. Singh, B. Foley, and J. Adams. The discovery of ancient history in the deep sea using advanced deep submergence technology. *Deep-Sea Research I*, 47(9):1591–1620, September 2000.

[5] R.D. Ballard, L.E. Stager, D. Master, D.R. Yoerger, D.A. Mindell, L.L. Whitcomb, H. Singh, and D. Piechota. Iron age shipwrecks in deep water off Ashkelon, Israel. *American Journal of Archaeology*, 106(2):151–168, April 2002.

[6] Y. Bar-Shalom, X.R. Li, and T. Kirubarajan. *Estimation with Applicaions To Tracking and Navigation.* John Wiley & Sons, Inc., 2001.

[7] A. Baumberg. Reliable feature matching across widely separated views. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[8] P.A. Beardsley, A. Zisserman, and D. Murray. Sequential Updating of Projective and Affine Structure from Motion. *International Journal of Computer Vision*, 23(3):235–259, June 1997.

[9] S. Beaulieu and K.L. Smith. Phytodetritus entering the benthic boundary layer and aggregated on the sea floor at an abyssal station in the NE Pacific: macro- and microscopic composition. *Deep-Sea Research II*, 45(4-5):781–815, June 1998.

[10] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, Santa Margherita Ligure, Italy, May 1992.

[11] P.J. Besl and N.D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-14(2):239–256, February 1992.

[12] M. Bosse. *Atlas Framework for Scalable Mapping.* Phd, Massachusetts Institute of Technology, February 2004.

[13] M. Bosse, P. Newman, J. Leonard, M. Soika, W. Feiten, and S. Teller. Atlas Framework for Scalable Mapping. In *IEEE International Conference on Robotics and Automation*, pages 1899–1906, Taiwan, September 2003.

[14] M. Bosse, P. Newman, J. Leonard, and S. Teller. An atlas framework for scalable mapping. Technical Report Technical memorandum 2002-04, MIT Marine Robotics Laboratory, 2002.

[15] S. Coorg and S. Teller. Matching and pose refinement with camera pose estimates. In *DARPA97*, pages 857–862, 1997.

[16] S. Coorg and S. Teller. Spherical mosaics with quaternions and dense correlation. *International Journal of Computer Vision*, 37(3):259–273, 2000.

[17] A.J. Davison. *Mobile Robot Navigation Using Active Vision.* Ph.d., University of Oxford, 1999.

[18] A.J. Davison and N. Kita. Sequential localisation and map-building in computer vision and robotics. In *SMILE Workshop at European Conference on Computer Vision*, 2000.

[19] M.C. Deans. *Bearing-Only Localization and Mapping.* Ph.d., Carnagie Mellon University, July 2002.

[20] S.Q. Duntley. Light in the sea. *Journal of the Optical Society of America*, 53(2):214–233, 1963.

[21] R. Eustice, O. Pizarro, and H. Singh. Visually augmented navigation in an unstructured environment using a delayed state history. In *Proceedings of the IEEE International Conference on Robotics and Automation ICRA2004*, volume 1, pages 25–32, April 2004.

[22] R. Eustice, O. Pizarro, H. Singh, and J. Howland. UWIT: Underwater Image Toolbox for Optical Image Processing and Mosaicking in Matlab. In *Proceedings of the 2002 International Symposium on Underwater Technology*, pages 141–145, Tokyo, Japan, April 2002.

[23] N. Farr. LED underwater lighting system. Private communication, March 2004.

[24] O. Faugeras. *Three-Dimensional Computer Vision.* MIT Press, 1993.

[25] O. Faugeras and O.-T. Luong. *The Geometry of Multiple Images : The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications.* MIT Press, 2001.

[26] O. Faugeras and S. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4(3):225–246, 1990.

[27] H.J.S. Feder, J.J. Leonard, and C.M. Smith. Adaptive concurrent mapping and localization using sonar. In *Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 892–898, Victoria, B.C., Canada, October 1998.

[28] M. A. Fischler and R Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[29] A.W. Fitzgibbon and A. Zisserman. Automatic Camera Recovery for Closed or Open Image Sequences. In *Proceedings of the 5th European Conference on Computer Vision*, pages 311–326, Freiburg, Germany, June 1998. Springer-Verlag.

[30] S.D. Fleischer, R.L. Marks, S.M. Rock, and M.J. Lee. Improved real-time video mosaicking of the ocean floor. In *MTS/IEEE Conference Proceedings OCEANS' 95*, volume 3, pages 1935–1944, San Diego, CA, October 1995.

[31] S.D. Fleischer, H.H. Wang, S.M. Rock, and M.J. Lee. Video mosaicking along arbitrary vehicle paths. In *Proceedings of the 1996 Symposium on Autonomous Underwater Vehicle Technology, 1996*, pages 293–299, Monterey, CA, June 1996.

[32] K.G. Foote. Censusing marine living resources in the gulf of Maine: A proposal. In *MTS/IEEE Oceans 2001*, volume 3, pages 1611–1614, Hononulu, 2001.

[33] D.A. Forsyth and Ponce J. *Computer Vison. A Modern Approach.* John Wiley & Sons, Inc., 2003.

[34] T.I. Fossen. *Guidance and Control of Ocean Vehicles.* John Wiley and Sons Ltd., New York, Date 1994.

[35] G. Godin, D. Laurendeau, and R. Bergevin. A Method for the Registration of Attributed Range Images. In *Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling (3DIM)*, pages 179–186, Quebec City, Que. , Canada, 2001.

[36] N. Gracias and J. Santos-Victor. Underwater mosaicing and trajectory reconstruction using global alignment. In *Oceans*, pages 2557–2563, Honolulu, Hawaii, 2001.

[37] S. Granger, X. Pennec, and A. Roche. Rigid Point-Surface Registration using Oriented Points and an EM variant of ICP for Computer Guided Oral Implantology. Research

report RR-4169, INRIA, 2001. Published in MICCAI'01, Utrecht, Netherlands, LNCS 2208, p.752-761.

[38] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, Manchester, U.K., 1988.

[39] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[40] R.I. Hartley and P. Sturm. Triangulation. *CVIU*, 68(2):146–157, November 1997.

[41] J. Heikkila and O. Silven. A Four-Step Camera Calibration Procedure with Implicit Image Correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1112, Puerto Rico, 1997.

[42] B.K.P. Horn. *Robot Vision.* McGraw–Hill, Cambridge, Massachusetts, 1986.

[43] B.K.P Horn. Closed form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987.

[44] B.K.P. Horn. Closed form solutions of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7):1127–1135, 1987.

[45] B.K.P Horn. Relative orientation. *International Journal of Computer Vision*, 4(1):58–78, 1990.

[46] J. Howland. Digital Data Logging and Processing, Derbyshire Survey, 1997. Technical report, Woods Hole Oceanographic Institution, December 1999.

[47] S.C. Hsu and H.S. Sawhney. Influence of global constraints and lens distortion on pose and appearance recovery from a purely rotating camera. In *Fourth IEEE Workshop on Applications of Computer Vision, 1998*, pages 154–159, October 1998.

[48] M. Hu. Pattern recognition by moment invariants. *Proceedings of the Institute of Radio Engineers*, 49:1428, 1961.

[49] M. Irani, P. Anandan, J.R. Bergen, R. Kumar, and S.C. Hsu. Mosaic representations of video sequences and their applications. *Signal Processing: Image Communication, special issue on Image and Video Semantics: Processing, Analysis, and Application*, 8(4), May 1995.

[50] J.S. Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111, April 1990.

[51] J.S. Jaffe, J. McLean, M.P. Strand, and K.D. Moore. Underwater optical imaging:status and prospects. *Oceanography*, 14(3):66–76, 2001.

[52] T. Jebara, A. Azarbayejani, and A.P. Pentland. 3d structure from 2d motion. In *Vismod*, 1999.

[53] A. Johnson and M. Hebert. Surface Registration by Matching Oriented Points. In *International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, pages 121–128, Ottawa, Ontario, Canada, May 1997.

[54] S.J. Julier and J.K. Uhlmann. A Non-divergent Estimation Algorithm in the Presence of Unknown Correlations. In *Proceedings of the American Control Conference*, volume 4, pages 2369–2373, Albuquerque, NM USA, 1997.

[55] F. Kahl and R.I. Hartley. Critical curves and surfaces for euclidean reconstruction. In *ECCV02*, volume 2, pages 447–462, 2002.

[56] K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice.* Elsevier, Amsterdam, The Netherlands, 1996.

[57] K. Kanatani and N. Ohta. Comparing Optimal 3-D Reconstruction for Finite Motion and Optical Flow. *Memoirs of the Faculty of Engineering, Okayama University*, 36(1):91–106, December 2001.

[58] A. Khotanzad and Y.H. Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[59] W.Y. Kim and Y.S. Kim. Robust rotation angle estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):768–773, 1999.

[60] W.Y. Kim and Y.S. Kim. A region-based shape descriptor using Zernike moments. *Signal Processing:Image Communication*, 16(1-2):95–102, September 2000.

[61] J.J. Leonard and H.J.S. Feder. Decoupled Stochastic Mapping. *IEEE Journal of Ocean Engineering*, 26(4):561–571, October 2001.

[62] M. Levoy, B. Chen, V. Vaish, M. Horowitz, and M. McDowall, I.and Bolas. Synthetic aperture confocal imaging. *ACM Transactions on Graphics. Special Issue: Proceedings of the 2004 SIGGRAPH Conference*, 23(3):825–834, 2004.

[63] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

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

[65] H.C. Longuet-Higgins. The Reconstruction of a Plane Surface from Two Perspective Projections. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 227(1249):399–410, May 1986.

[66] H.C. Longuet-Higgins. Multiple interpretations of a pair of images of a surface. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 418(1854):1–15, July 1988.

[67] F. Lu and E. Milios. Globally Consistent Range Scan Alignment for Environment Mapping. *Autonomous Robots*, 4(4):333–349, 1997.

[68] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, September 2002.

[69] S. Maybank and O. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–152, 1992.

[70] P.F. McLauchlan. Gauge Independence in Optimization Algorithms for 3D Vision. In *Workshop on Vision Algorithms*, pages 183–199, 1999.

[71] P.F. McLauchlan and A.H. Jaenicke. Image mosaicing using sequential bundle adjustment. In *British Machine Vision Conference*, pages 616–625, Bristol, UK, 2000.

[72] E. Mikhail, J. Bethel, and J.C. McGlone. *Introduction to Modern Photogrammetry.* John Wiley & Sons, Inc., 2001.

[73] K. Mikolajczyk and C. Schmid. An Affine Invariant Interest Point Detector. In *Proceedings of the 7th European Conference on Computer Vision*, pages 128–142, Copenhagen, Denmark, May 2002.

[74] P.H. Milne. *Underwater Acoustic Positioning Systems.* Gulf Publishing Company, Houston, 1983.

[75] F. Mindru, T. Moons, and L. Van Gool. Recognizing color patterns irrespective of viewpoint and illumination. In *Proceedings of CVPR99*, pages 368–373, 1999.

[76] C. Mobley. *Light and Water: Radiative Transfer in Natural Waters.* Academic Press, 1994.

[77] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002. AAAI.

[78] D. Morris. *Gauge Freedoms and Uncertainty Modeling for 3D Computer Vision.* Phd, Carnegie Mellon University, March 2001.

[79] National Transportation Safety Board, Washington, DC. *Aircraft Accident Brief: EgyptAir Flight 990, Boeing 767-366ER, SU-GAP, 60 Miles South of Nantucket, Massachuesetts, October 31, 1999*, 2002. Aircraft Accident Brief NTSB/AAB-02/01.

[80] S. Negahdaripour and H. Madjidi. Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures. *IEEE Journal of Oceanic Engineering*, 28(4):625–650, October 2003.

[81] S. Negahdaripour, X. Xu, and L. Jin. Direct estimation of motion from sea floor images for automatic station-keeping of submersible platforms. *IEEE Journal of Oceanic Engineering*, 24(3):370–382, July 1999.

[82] S. Negahdaripour and X. Xun. Mosaic-Based Positioning and Improved Motion-Estimation Methods for Automatic Navigation of Submersible Vehicles. *IEEE Journal of Oceanic Engineering*, 27(1):79–99, January 2002.

[83] P. Newman and J. Leonard. Pure range-only sub-sea SLAM. In *Proceedings of the IEEE International Conference on Robotics & Automation*, pages 1921–1926, September 2003.

[84] D. Nister. Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. In *ECCV00*, pages I: 649–663, 2000.

[85] D. Nister. An efficient solution to the five-point relative pose problem. In *CVPR03*, pages II: 195–202, 2003.

[86] X. Pennec. Computing the mean of geometric features - application to the mean rotation. Research Report RR-3371, INRIA, March 1998.

[87] X. Pennec and J-P. Thirion. A framework for uncertainty and validation of 3d registration methods based on points and frames. *International Journal of Computer Vision*, 25(3):203–229, 1997.

[88] J. Philip. A non-iterative algorithm for determining all essential matrices corresponding to five point pairs. *Photogrammetric Record*, 15(88):589–599, October 1996.

[89] J. Philip. Critical point configurations of the 5-,6-,7-, and 8-point algorithms for relative orientation. Technical Report TRITA-MAT-1998-MA-13, KTH Royal Institute of Technology, 1998.

[90] O. Pizarro, R. Eustice, and H. Singh. Relative pose estimation for instrumented, calibrated imaging platforms. In *Proceedings of the 2003 Conference on Digital Image Computing Techniques and Applications*, Sydney, Australia, 2003.

[91] O. Pizarro and H. Singh. Toward Large-Area Underwater Mosaicing for Scientific Applications. *IEEE Journal of Oceanic Engineering*, 28(4):651–672, 2003.

[92] C.J. Poelman and T. Kanade. A Paraperspective Factorization Method for Shape and Motion Recovery. *PAMI*, 19(3):206–218, March 1997.

[93] M. Pollefeys. 3D Modelling from Images (Tutorial). In *ECCV2000*, 2000.

[94] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Metric 3D Surface Reconstruction from Uncalibrated Image Sequences. In *3D Structure from Multiple Images of Large Scale Environments, LNCS Series*, volume 1506, pages 139–154. Springer-Verlag, 1997.

[95] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Hand-held acquisition of 3d models with a video camera. In *Second International Conference on 3-D Digital Imaging and Modeling*, pages 14–23, Los Alamitos, CA, 1999. IEEE Computer Society Press.

[96] M. Pollefeys, F. Verbiest, and L. Van Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *ECCV (2)*, pages 837–851, 2002.

[97] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760, 1998.

[98] R. Raskar, K. Tan, R. Feris, J. Yu, and M. Turk. Non-photorealistic Camera: Depth Edge Detection and Stylized Rendering using Multi-Flash Imaging. *ACM Transactions on Graphics. Special Issue: Proceedings of the 2004 SIGGRAPH Conference*, 23(3):679–688, 2004.

[99] RD Instruments, San Diego, CA. *Acoustic Doppler Current Profiler. Principles of Operation. A Practical Primer*, 1996.

[100] J.R. Reynolds, R.C. Highsmith, B. Konar, C.G. Wheat, and D. Doudna. Fisheries and fisheries habitat investigations using undersea technology. In *MTS/IEEE Oceans 2001*, pages 812–819, Honolulu, Hawaii, 2001.

[101] S. Rusinkiewicz and M. Levoy. Efficient Variants of the ICP Algorithm. In *Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling (3DIM)*, pages 145–152, Quebec City, Que. , Canada, 2001.

[102] H.S. Sawhney, S.C. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *European Conference on Computer Vision*, pages 103–119, Freiburg, Germany, 1998.

[103] H.S. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):235–243, March 1999.

[104] C. Schlegel and T. Kämpke. Filter Design for Simultaneous Localization and Map Building (SLAM). In *IEEE International Conference on Robotics & Automation*, pages 2737–2742, Washington, DC, May 2002.

[105] C. Schmid and R. Mohr. Matching by Local Invariants. Technical Report 2644, INRIA, August 1995.

[106] C. Schmid and R. Mohr. Local Greyvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

[107] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[108] R. Sedgewick. *Algorithms in C, Part 5: Graph Algorithms.* Addison Wesley Professional, 1998.

[109] G. Sharp, S. Lee, and D. Wehe. Multiview Registration of 3D Scenes by Minimizing Error Between Coordinate Frames. In *Proceedings of the 7th European Conference on Computer Vision*, pages 587–597, Copenhagen, Denmark, May 2002.

[110] H.-Y. Shum and R. Szeliski. Panoramic image mosaics. Technical Report MSR-TR-97-23, Microsoft Research, Redmond, WA, 1997.

[111] H. Singh, R. Eustice, C. Roman, O. Pizarro, R. Armstrong, F. Gilbes, and J. Torres. Imaging coral I: Imaging coral habitats with the SeaBED AUV. *Subsurface Sensing Technologies and Applications*, 5(1):25–42, January 2004.

[112] C.C. Slama, editor. *Manual of Photogrammetry.* American Society of Photogrammetry, Bethesda, MD, fourth edition, 1980.

[113] C.R. Smith. Whale falls: Chemosynthesis at the deep-sea floor. *Oceanus*, 35(3):74–78, 1992.

[114] R. Smith, M. Self, and P. Cheeseman. *Estimating Uncertain Spatial Relationships in Robotics.* Autonomous Robot Vehicles. Springer-Verlag, 1990.

[115] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV96*, pages II:709–720, 1996.

[116] R. Szeliski. Image mosaicing for tele-reality applications. Technical report CRL 94/2, Cambridge Research Laboratory, Cambridge, MA, May 1994.

[117] C.-H. Teh and R.T. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1998.

[118] D. Tell and S. Carlsson. Combining Appearance and Topology for Wide Baseline Matching. In *ECCV*, volume 1, pages 68–81, Copenhagen, Denmark, 2002.

[119] S. Thrun, D. Koller, Z. Ghahramani, H.F. Durrant-Whyte, and A.Y. Ng. Simultaneous Mapping and Localization with Sparse Extended Information Filters: Theory and Initial Results. In J.D. Boissonnat, J. Burdick, K. Goldberg, and S. Hutchinson, editors, *Proceedings of the Fifth International Workshop on Algorithmic Foundations of Robotics*, Nice, France, to appear 2002.

[120] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9(2):137–154, November 1992.

[121] P. Torr, A. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated images. *International Journal of Computer Vision*, 32(1):27–44, August 1999.

[122] Philip H. S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *Workshop on Vision Algorithms, ICCV*, pages 278–294, 1999.

[123] P.H.S. Torr and A. Zisserman. Performance characterization of fundamental matrix estimation under image degradation. *MVA*, 9(5-6):321–333, 1997.

[124] B. Triggs. Autocalibration and the absolute quadric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 609–614, San Juan, Puerto Rico, 1997.

[125] B. Triggs. Routines for relative pose of two calibrated cameras from 5 points. Technical report, INRIA, 2000.

[126] B. Triggs, P.F. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. In Bill Triggs, A. Zisserman, and Robert Szeliski, editors, *Vision Algorithms: Theory & Practice*. Springer-Verlag, 2000.

[127] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proceedings of the British Machine Vision Conference 2000*, pages 736–739, Bristol, UK, 2000.

[128] T. Tuytelaars and L. Van Gool. Matching Widely Separated Views baed on Affine Invariant Regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[129] S. Umeyama. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[130] D.R. Yoerger, A.M. Bradley, M.-H. Cormier, W.B.F. Ryan, and B.B. Walden. Fine-scale seafloor survey in rugged deep-ocean terrain with an autonomous robot. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1767–1774, San Francisco, USA, 2000.

[131] D.R. Yoerger, A.M. Bradley, H. Singh, B.B. Walden, M.H. Cormier, and W.B.F. Ryan. Multisensor Mapping of the Deep Seafloor with the Autonomous Benthic Explorer. In *Proceedings of the 2000 International Symposium on Underwater Technology*, pages 248–253, Tokyo, Japan, May 2000.

[132] Z. Zhang. Iterative Closest Point Matching for Registration of Free-Form Curves and Surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.

[133] Z. Zhang. Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. *Image and Vision Computing Journal*, 15(1):59–76, 1997.

[134] Z. Zhang and Y. Shan. Incremental motion estimation through local bundle adjustment. Technical Report MSR-TR-01-54, Microsoft Research, May 2001.

[135] K. Zuiderveld. Contrast limited adaptive histogram equalization. In Paul Heckbert, editor, *Graphics Gems (IV)*, pages 474–485. Academic Press, Boston, Date 1994.